Human Genetics
https://doi.org/10.1007/s00439-023-02553-1

ORIGINAL INVESTIGATION

High diagnostic potential of short and long read genome sequencing with transcriptome analysis in exome-negative developmental disorders

François Lecoquierre[1] · Olivier Quenez[1] · Steeve Fourneaux[1] · Sophie Coutant[1] · Myriam Vezain[1] · Marion Rolain[1] · Nathalie Drouot[1] · Anne Boland[2] · Robert Olaso[2] · Vincent Meyer[2] · Jean-François Deleuze[2] · Dana Dabbagh[3] · Isabelle Gilles[4] · Claire Gayet[5] · Pascale Saugier-Veber[1] · Alice Goldenberg[1] · Anne-Marie Guerrot[1] · Gaël Nicolas[1]

Received: 21 January 2023 / Accepted: 5 April 2023

Abstract

Exome sequencing (ES) has become the method of choice for diagnosing rare diseases, while the availability of short-read genome sequencing (SR-GS) in a medical setting is increasing. In addition, new sequencing technologies, such as long-read genome sequencing (LR-GS) and transcriptome sequencing, are being increasingly used. However, the contribution of these techniques compared to widely used ES is not well established, particularly in regards to the analysis of non-coding regions. In a pilot study of five probands affected by an undiagnosed neurodevelopmental disorder, we performed trio-based short-read GS and long-read GS as well as case-only peripheral blood transcriptome sequencing. We identified three new genetic diagnoses, none of which affected the coding regions. More specifically, LR-GS identified a balanced inversion in *NSD1*, highlighting a rare mechanism of Sotos syndrome. SR-GS identified a homozygous deep intronic variant of *KLHL7* resulting in a neoexon inclusion, and a de novo mosaic intronic 22-bp deletion in *KMT2D*, leading to the diagnosis of Perching and Kabuki syndromes, respectively. All three variants had a significant effect on the transcriptome, which showed decreased gene expression, mono-allelic expression and splicing defects, respectively, further validating the effect of these variants. Overall, in undiagnosed patients, the combination of short and long read GS allowed the detection of cryptic variations not or barely detectable by ES, making it a highly sensitive method at the cost of more complex bioinformatics approaches. Transcriptome sequencing is a valuable complement for the functional validation of variations, particularly in the non-coding genome.

Introduction

High-throughput sequencing has significantly advanced our understanding of the molecular basis of genetic diseases. In particular, trio exome sequencing (ES) has provided insight into the impact of de novo mutations on the coding sequence of the genome (Deciphering Developmental Disorders Study 2017; Kaplanis et al. 2020). The usefulness of ES for gene discovery and clinical applications is now well established.

Genome sequencing (GS) has also been used for investigating rare diseases for around a decade (Gilissen et al. 2014) but its use in routine diagnosis is more recent. The cost of GS, the more complex informatics procedures compared to ES, and the unclear added value in terms of diagnostic yield have slowed the adoption of GS in diagnostic routine compared to ES. However, sequencing procedures and data analysis have improved, making GS more viable in healthcare systems. As a result, several countries have developed public health strategies around GS, with

✉ François Lecoquierre
francois.lecoquierre@chu-rouen.fr

✉ Gaël Nicolas
gaelnicolas@hotmail.com

[1] Univ Rouen Normandie, Inserm U12045 and CHU Rouen, Department of Genetics and Reference Center for Developmental Disorders, FHU-G4 Génomique, F-76000 Rouen, France

[2] Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057 Evry, France

[3] Department of Pediatrics, Elbeuf Hospital, Elbeuf, France

[4] Department of Pediatrics, Evreux Hospital, Evreux, France

[5] Department of Pediatrics, CHU Rouen, F-76000 Rouen, France

Published online: 19 April 2023

significant investments (Lévy 2016; Kovanda et al. 2021; 100,000 Genomes Project Pilot Investigators et al. 2021). Recommendations for the diagnostic use of GS have been established, both long-standing (van El et al. 2013) and recent (Souche et al. 2022), including guidance on the interpretation of non-coding variations (Ellingford et al. 2022). However, the contribution of non-coding regions to rare diseases and the benefits of genome sequencing compared to ES as a diagnostic tool are not well understood. While certain short-read protocols, such as the large-insert GS method, have been developed to mitigate the limitations of small molecule sequencing with respect to the detection sensitivity of structural variations (Dong et al. 2019), long-read GS approaches are increasingly used in patients with rare diseases.

Long-read genome sequencing (LR-GS) provides access to certain types of variation that are difficult to capture in short-read genome sequencing (SR-GS), such as variations in complex regions, short tandem repeats (STRs) and other repeats, balanced structural variations, and mobile elements of the genome (Mantere et al. 2019). LR-GS has been used at the scale of small populations (Beyter et al. 2021; Wu et al. 2021) and is expected to be a key part of the future of genomics (De Coster et al. 2021), although its use for the diagnosis of rare diseases has been limited to a small number of patients (Pauper et al. 2021; Hiatt et al. 2021). The two main long molecule sequencing techniques currently available (Eid et al. 2009; Clarke et al. 2009) historically had a high error rate per base, including many insertions and deletions. However, they still allowed accurate detection of structural variations. Recently, substantial optimizations of these protocols have enabled obtaining high-quality data on short variants as well (Wenger et al. 2019). Since these optimizations were not available for this study, we used SR-GS and LR-GS in a complementary way to detect a wide range of variations of all types.

The use of RNA sequencing (RNA-seq) in the form of bulk transcriptome analysis on whole blood or other accessible biological samples is emerging as a versatile tool for the diagnosis of rare diseases, through its global analysis of transcripts. Transcriptome sequencing allows the detection of quantitative and qualitative abnormalities in gene expression resulting from genetic variations. Specifically, the search for decreased expression, monoallelic expression, or splicing defects in the form of abnormal retention or exclusion of portions of transcripts can serve as biomarkers of the biological effect of candidate variations. Although it is possible to detect variations in RNA-seq data, the sensitivity of detection is low and this technique is often used in combination with more robust genomic DNA sequencing (such as ES or GS). These combined approaches have been shown to have higher diagnostic yield than genomic DNA sequencing alone (Lee et al. 2020; Colin et al. 2022; Coursimault et al. 2022), particularly through the functional validation of candidate variants.

The objective of this pilot study was to evaluate the contribution of these innovative techniques in identifying pathogenic variants. We present the results of trio short-read and long-read genome sequencing, along with case-only transcriptome sequencing, in a series of five patients with unexplained developmental abnormalities after exome analysis.

Materials and methods

Patients and samples

Inclusion criteria were: (i) patient with a developmental disorder of suspected genetic origin, (ii) no known genetic cause following genetic investigation including at least exome sequencing, and (iii) agreement of the parents to participate. Patients were selected among those followed in the clinical genetics consultation of the Rouen University hospital, until $n = 5$ trios were included.

Each patient received genetic analysis according to clinical orientation and diagnostic strategies, and had undergone ES that was considered negative prior to this study. Informed written consent was obtained from both parents of each proband for genetic analysis in a medical setting and participation to the study. The study was approved by the CPP Ouest V (20/043-2) ethics committee. Biological material for SR-GS, LR-GS, and RNA-seq analyses was obtained from EDTA and Paxgene blood samples. Figure 1 summarizes the sampling, preparation methods, sequencing, and data analysis.

SNV/indel detection by trio short-read genome sequencing

DNA was extracted from whole blood using standard procedures. SR-GS was performed in the *Centre National de Recherche en Génomique Humaine* (CNRGH, Institut de Biologie François Jacob, CEA, Evry, France). After quality control, genomic DNA (1 µg) was used to prepare libraries for whole genome sequencing using the Illumina TruSeq DNA PCR-Free Library Preparation Kit (Illumina Inc., CA, USA) according to the manufacturer's instructions. After quality control and normalization, qualified libraries were sequenced on a NovaSeq6000 platform from Illumina (Illumina Inc., CA, USA) as paired-end 150 bp reads. Samples were pooled on a NovaSeq6000 S4 flowcell to reach an average sequencing depth of > 30×. Sequence quality parameters were assessed throughout the sequencing run and standard bioinformatics analysis of the sequencing data was based on the Illumina pipeline to generate fastq files for each sample. Read alignment

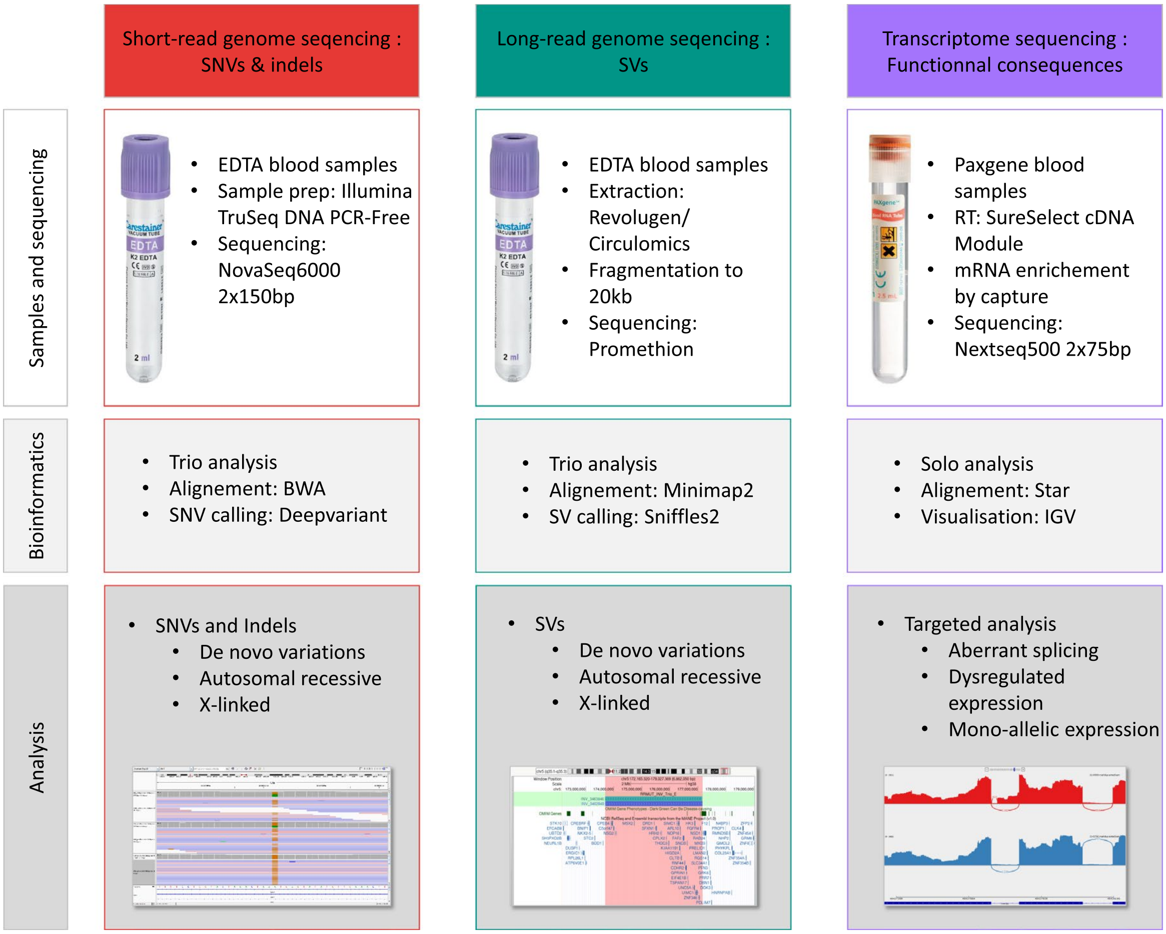


Fig. 1 Sequencing methods and strategies. The main methods are presented, but other complementary bioinformatics approaches were performed, including for SR-GS the detection of uniparental disomies, the genotyping of STRs involved in pathology, and the detection of regions of homozygosity (see "Materials and methods")

to the reference genome GRCh38 was performed using BWA, and SNV/indel calling was performed on each alignment file using Deepvariant (Poplin et al. 2018). Individual gVCFs were merged using GLnexus. Standard gene and variant-based annotations were added using custom scripts. De novo variants were called using an in-house pipeline as previously described (Coursimault et al. 2022), consisting of (i) the detection of de novo candidates through a series of filters applied by bcftools on the multi-vcf, and (ii) manual review of the de novo candidates on IGV using a dedicated python script (see web resources). Complementary variant types were also called on SR genomes, including the analysis of disease-associated STRs using ExpansionHunter, the detection of uniparental disomies using a custom script (UPD_plotter, see web resources), and the detection of regions of homozygosity using AutoMap (Quinodoz et al. 2021).

SV detection by long-read genome sequencing

DNA of high molecular weight were extracted from PBMCs using Revolugen kit for trios A–D, and from frozen blood using Circulomics kit for trio E. DNA was quantified and fragmented to a target size of 20 kb. Samples were prepared using SQK-LSK109 or SQK-LSK110 ligation kits and sequenced on a Promethion instrument (Oxford Nanopore Technologies, ONT). QC was performed on raw data by nanoplot and reads were aligned on GRCh38 by minimap2. Calling of structural variants (DEL, INS, DUP, INV, BND) was performed on individual alignment files and merged into a single multi-vcf by Sniffles2 with standard parameters. Variants were annotated by annotSV and filtered via BCFtools. Variants were ranked according to patients phenotype and variant effect using SvAnna (Danis et al. 2022).

Transcriptome sequencing

RNA was extracted from Paxgene blood samples following the manufacturer's protocol. Reverse transcription was performed using the random primers based SureSelect cDNA Module (Agilent Technologies). Agilent Magnis was used to capture coding exons with the SureSelect all exon V7 probes. Sequencing was performed on Illumina Nextseq 500 in the Rouen sequencing facility. Eight samples were pooled and sequenced using High Output Kit v2.5 (150 Cycles, 400 M reads). Data was processed using nf-core rnaseq pipeline (see web resources), including read alignment on GRCh38 by STAR, quantification by Salmon, and aberrant junction analysis by RSeQC. Analysis of aberrant splicing and mono-allelic expression at specific loci of interest were performed using IGV. For quantitative analysis, we compared the adjusted gene expression using the transcripts per million (TPM) metrics between the five probands plus three additional samples from the same sequencing batch belonging to patients with developmental disorders and a variant of unknown significance in unrelated genes.

Variant filtration

Variant filtration and analysis was performed following the major inheritance modes in developmental disorders, i.e. (i) de novo, (ii) recessive, either homozygous or compound heterozygous, and (iii) X-linked. Standardized variant interpretation guidelines were applied (Richards et al. 2015). Global methodology for variant detection and analysis are depicted on Fig. 1.

Results

We included five probands (three males, two females, age at inclusion ranged from 2.7 to 13.4 years) and their unaffected parents. All probands presented with a neurodevelopmental disorder of unknown cause following clinical and molecular evaluations and follow-up by clinicians specializing in rare developmental disorders. A sporadic presentation was noted in four families, while in family B, the proband had a younger brother with similar symptoms in a context of consanguinity, highly suggestive of an autosomal recessive disorder.

Sequencing quality and variant calling

Sequencing metrics of the 15 samples are summarized in Fig. 2. SR-GS produced about 500 million reads per individual, leading to a median depth of 43 × after read alignment and deduplication. SNV/indels variant counts were highly homogeneous between individuals, with a median of 4.7 million SNV/indels per sample after calling from SR data using Deepvariant, including 1.200 rare (frequency < 1% in gnomAD v2.1) variants with a high, moderate or low effect as annotated by VEP (Supplementary Fig. 1). Probands harbored from 49 to 111 de novo SNVs and from 7 to 11 de novo indels with high confidence (Supplementary Fig. 2), with a paternal age correlation consistent with current knowledge.

LR-GS achieved similar average depth of coverage of 40×. Read length N50 was about 14.9 kb, meaning that 50% of bases sequenced were within reads of this size or more. Structural variants were detected from LR alignments by the Sniffles2 algorithm. Similar to SNV/indels, SV count was highly homogeneous between the 15 individuals with an average of 26,191 (range 25,770–27,479) SV calls (deletion, duplication, insertion, inversion and break-ends). Supplementary Fig. 3 summarizes the counts and sizes of SV calls among the 15 individuals. Mendelian transmission, evaluated by the mendelian bcftools plugin was concordant for 92% of variants detected in probands, ranging from 81% for break-ends to 95% for duplications. Deletions and insertions represented 53% and 47% of all calls respectively, while the other SV types jointly accounted for less than 1% of events. SV count was inversely correlated with their size for all variant types. Two peaks at ~ 300 and ~ 7.000 bp were observed in the insertion and deletion events, recapitulating previous observations of Alu and Line mediated SVs, respectively (De Coster et al. 2019; Pauper et al. 2021; Beyter et al. 2021; Wu et al. 2021).

Identification of a likely pathogenic or pathogenic variant in 3/5 patients

Prioritization of rare variants as detected by SR (SNV/indels) and LR (SV) GS following Mendelian hypotheses (de novo mutation, autosomal recessive, X-linked inheritance) allowed us to identify a (likely) pathogenic variant in 2/5 patients (Patient C and Patient E) and a candidate homozygous deep intronic variant in one patient (Patient B). Transcriptome sequencing data provided extra evidence for pathogenicity in all three patients thus leading to a final diagnosis in 3/5 patients.

Identification of a homozygous deep intronic variant in KLHL7

An autosomal recessive disorder was highly suspected in proband B based on familial information. Her unaffected parents were first cousins and her younger brother presented with the same specific phenotype including intrauterine growth retardation (IUGR), microcephaly, facial dysmorphic features, and intellectual disability (Fig. 3A, supplementary information). Inspection of homozygous

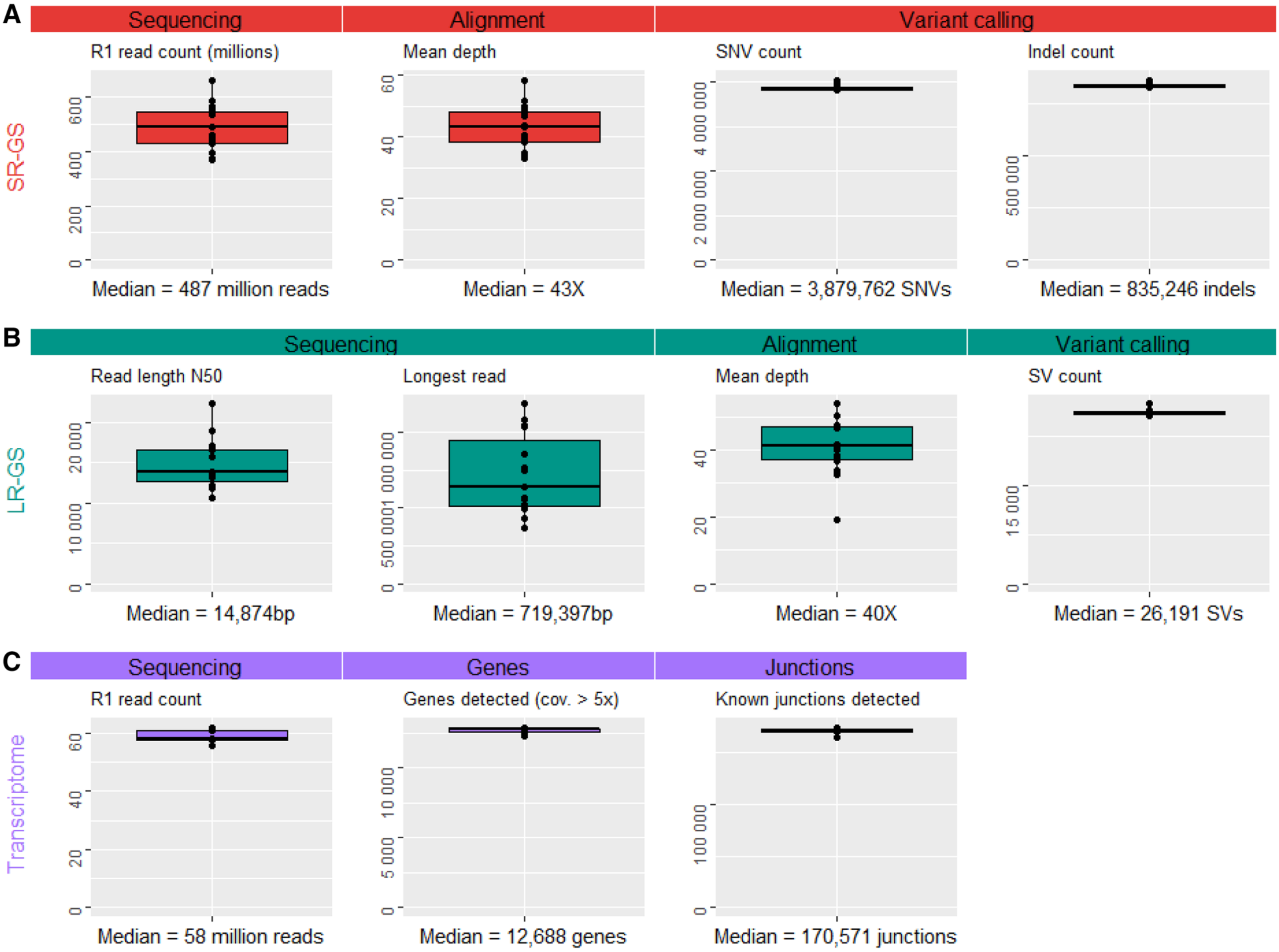
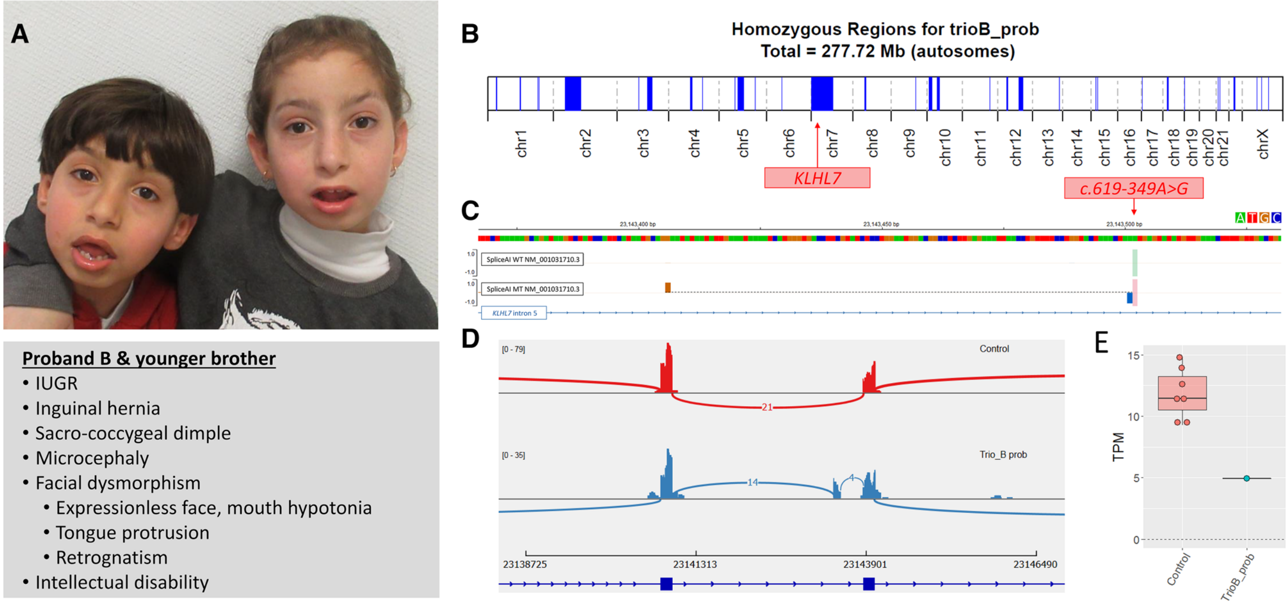


Fig. 2 Quality metrics and variant counts. Each point represents one sample (*n* = 15 for SR and LR-GS, *n* = 5 for transcriptomes). The individual values of each metrics is available in Supplementary Material. **A** Short-read Illumina genomes: number of reads sequenced (bash script from R1 FASTQs), median depth on GRCh38 (Mosdepth), number of SNVs and Indels (DeepVariant calling report). **B** Long-read ONT genomes: read size assessed by N50 read size (50% of sequenced bases are in reads longer than this measure), and by the length of the longest read in the library (source: Nanoplot). Median depth on GRCh38 evaluated by Mosdepth. Variant calling by Sniffles2. **C** Transcriptomes on whole blood samples: read count (FASTQC), genes detected, as evaluated by a depth of coverage of > 5X (StingTie) and count of known junctions detected (RSeQC)

regions using the Automap software was consistent with the level of consanguinity. Analysis of homozygous variants revealed a variant initially considered as a variant of uncertain significance in the *KLHL7* gene, located within a large homozygous region on chr7 (Fig. 3B): NM_001031710.3:c.619-349A > G, p.?. This variant retained our attention because bi-allelic loss-of-function variants in *KLHL7* are involved in Perching Syndrome (OMIM #617055), an autosomal recessive syndromic developmental disorder (Angius et al. 2016; Bruel et al. 2017; Jeffries et al. 2019) with a strong clinical overlap with the trio B siblings. In addition, this candidate intronic variant was predicted by SpliceAI to result in a gain of a donor site (delta score = 0.78), while a combined use of a predicted acceptor site (delta score = 0.74) would result in the inclusion of a hypothetical 95-bp neo-exon (Fig. 3C). Variant segregation analysis in the affected younger brother showed that he was also homozygous for this variant. Transcriptome analysis from a blood sample of the proband confirmed the presence of the suspected out-of-frame neo-exon (Fig. 3D) and quantitative analysis of TPM using the Salmon software showed reduced *KLHL7* expression, likely due to nonsense-mediated decay on both alleles (Fig. 3E). Altogether, we identified a deep intronic variant with a significant deleterious effect on splicing in the *KLHL7* gene, with concordant clinical and segregation arguments. These findings allowed us to finally classify this variant as likely pathogenic and to solve this familial case.



Proband B & younger brother

- IUGR
- Inguinal hernia
- Sacro-coccygeal dimple
- Microcephaly
- Facial dysmorphism
  - Expressionless face, mouth hypotonia
  - Tongue protrusion
  - Retrognatism
- Intellectual disability

Fig. 3 Short read genome sequencing identifies a homozygous deep intronic variant in KLHL7 leading to a neo-exon. The variant NM_001031710.3:c.619-349A > G, p.? was identified at a homozygous state in proband B and her younger brother. **A** Clinical data. Photograph represent proband A at 10 years of age, and her younger brother at 7 years of age, showing hypotonia of the mouth and expressionless face in both siblings. **B** Regions of homozygosity in proband B's genome, as detected by the Automap software (Quinodoz et al. 2021). Note the presence of the gene KLHL7 in the largest homozygous region on chr7. **C** SpliceAI predictions of splicing in wild-type (upper) and mutant (lower) panels. The reference base is highlighted in green and the alternative base is in red. Note the prediction of both a donor splice site in blue and an acceptor site in brown, framing a 95-bp pseudo exon at the genomic position NC000007.14:g.23,143,406–23,143,500, depicted as a dashed line. Plot generated using SpliceAI Visual (de Sainte Agathe et al. 2023) on the Mobidetails platform (Baux et al. 2021). **D** Transcriptome data on Paxgene blood sample confirming the intronic retention of a 95-bp neoexon in proband B and not in batch controls. IGV sashimi plot, GRCh38. **E** Quantitative TPM analysis shows a lowed *KLHL7* expression compared to batch controls. *IUGR* intra-uterine growth retardation

Identification of a complex balanced structural variant in NSD1

Proband E was a 7-year-old patient of Caucasian descent. She presented with a sporadic developmental disorder including statural advance with macrocephaly, scoliosis, pectus excavatum, arachnodactyly, facial dysmorphic features, and delayed motor and language acquisitions (Fig. 4A, supplementary information). A disorder in the spectrum of Sotos syndrome was the main hypothesis, but neither panel analysis nor ES analysis of the proband led to a molecular diagnosis. Sniffles2, the structural variant caller used on LR-GS data, identified two de novo inversion events on chr5 (Fig. 4B) that were further inspected on short-read and long-read alignments (Fig. 4C). These two events recapitulated a complex and mostly balanced structural variant consisting of a 3.5 Mb inversion, flanked by two deletions of ~ 5 kb at each side: NC_000005.10:g.[173693667_173699649del;173699650_177157517inv;177157518_177162599del]. A UCSC session showing the event is available in web resources. While the proximal breakpoint was intergenic, the distal breakpoint was located within intron 2 of *NSD1* (NM_022455.5), the gene responsible for Sotos syndrome. LR-GS allowed us to phase this variant to the paternal haplotype (data not shown). The disruption of *NSD1* coding sequence and the separation of most of the transcript from its promoter led us to hypothesize the presence of monoallelic expression of the maternal allele. Consistent with this hypothesis, the exonic rs1363405 SNP, heterozygous in proband E and her father, was not detected in the transcriptome, indicating a total absence of expression of the paternal allele in proband E (Fig. 4D). Furthermore, relative expression compared to controls showed a decreased expression of *NSD1* of ~ 50% (Fig. 4E). The clinical hypothesis, the de novo nature of this variant, and the clear consequence on transcription led us to classify this variant as pathogenic.

Identification of a splicing pathogenic variant in KMT2D

Proband C was a non-consanguineous 15-year-old patient affected by a sporadic neurodevelopmental disorder. He presented with intellectual disability associated with scoliosis, congenital heart defect and facial dysmorphic features (Fig. 5A, supplementary information). This association was evocative of Kabuki syndrome (KS), but neither a gene panel nor an exome approach identified a molecular cause.

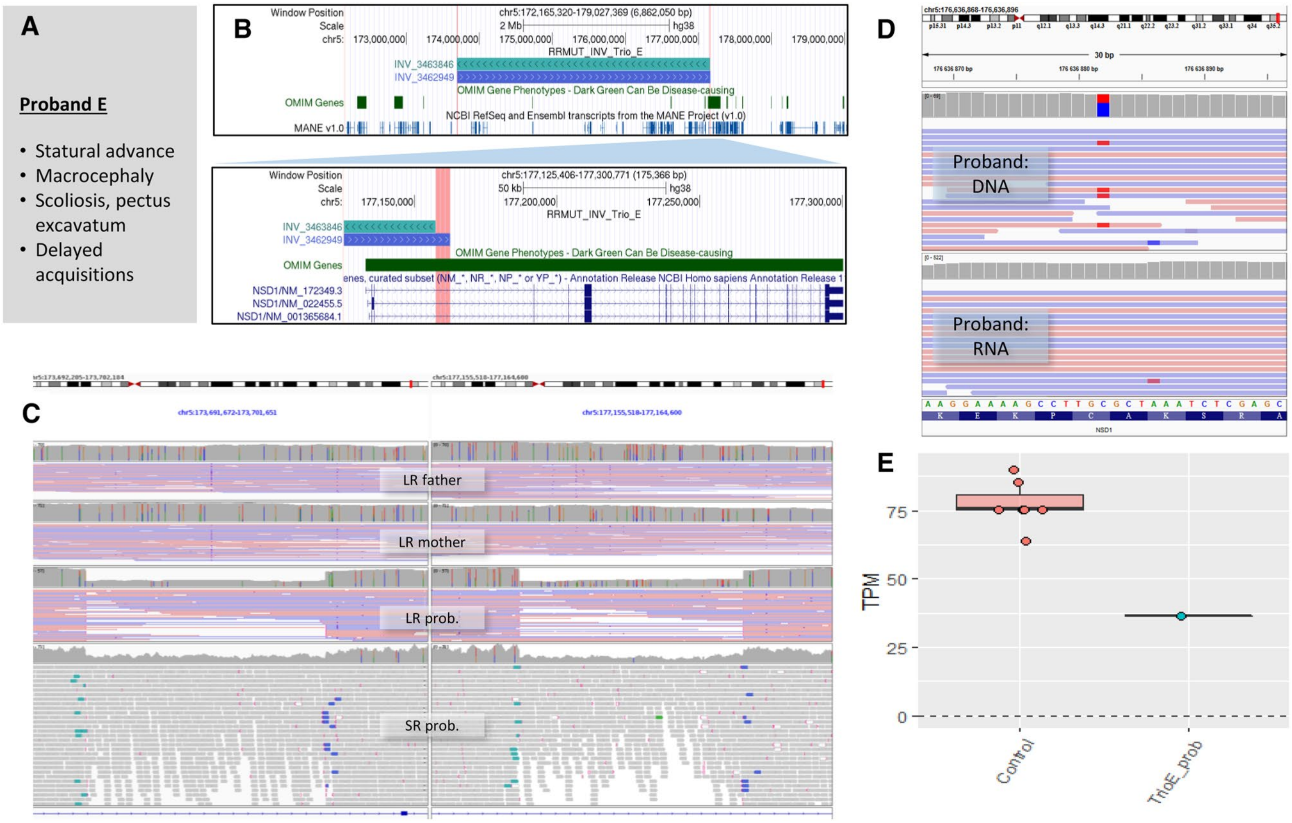


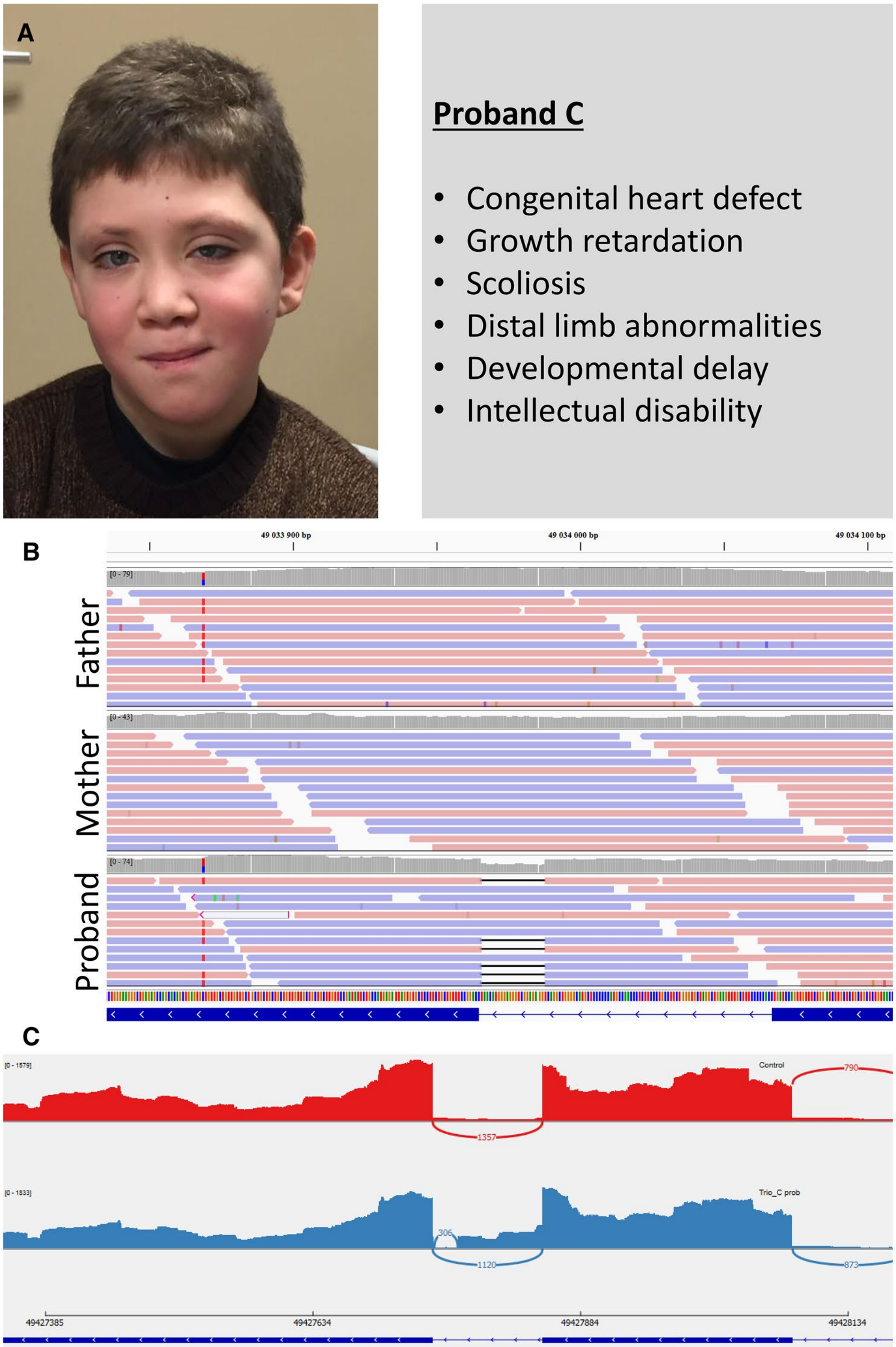
Fig. 4 Long read genome sequencing identifies a de novo inversion on chr5 leading to *NSD1* haploinsufficiency with mono-allelic expression. **A** Clinical summary of proband E. **B** UCSC visualization the de novo event detected by Sniffles 2. The SV was detected as two distinct but close 3.5 Mb inversions. Sniffles2 called two events because of 5979 bp and 5082 bp deletions at breakpoints (depicted in red). Note the presence of an OMIM morbid gene on the second breakpoint, namely *NSD1*. Upper panel: global view of the event. Lower panel: zoomed view on the breakpoint within one intron of *NSD1*. **C** IGV visualization of long read and short read alignments. Two genomic windows focusing on breakpoints are presented. Note the de novo nature of this complex SV. The two ~ 5 Kb heterozygous deletions are visible in short and long read alignments. Short read representation using « color by pair orientation» confirms that the 3.5 Mb fragment between the two deletions is inverted. **D** Transcriptomic consequences: monoallelic expression. The paternal exonic rs1363405 SNP, located in the coding sequence of *NSD1* (exon 5, NM_022455.5), and present at a heterozygous state in proband B's genome was used as a marker to distinguish both alleles. Transcriptome did not show any read supporting the alternate allele of rs1363405, indicating a strict monoallelic expression of proband B's maternal allele. **E** Transcriptomic consequences: accordingly, quantitative TPM analysis shows a lower *NSD1* expression compared to batch controls

Analysis of de novo SNV and indels in trio-based SR-GS showed a 22pb deletion in intron 39 of *KMT2D*, one of the two genes associated with KS: NM_003482.4:c.10741-23_10741-2del, NC_000012.12:g.49033967_49033988 del (Fig. 5B). Variant allelic fraction was 20%, suggesting a post-zygotic event. Proband C was heterozygous for an exonic SNP, rs3782357, which was located at close proximity of the deletion. This variant, inherited from the father, could be used for variant phasing directly in SR-GS data. Only about half of the reads spanning the heterozygous SNP harbored the de novo splicing variant, which validated a post-zygotic mosaicism on the paternal allele (data not shown). This variant was located directly adjacent to the canonical 3′ splice site of intron 39 of *KMT2D* and was predicted to result in a loss of this splice site by in-silico algorithms including SpliceAI. This effect was assessed using transcriptome sequencing, which showed partial intron retention, which was consistent both with the mosaic distribution of the variant and putative partial degradation by NMD (Fig. 5C). Based on its de novo transmission, loss-of-function consequence, and high clinical concordance with KS in this patient, this variant was classified as pathogenic. While this variant could have been seen on previous exome and panel analysis, it is likely that the mosaicism is involved in this missed diagnosis.

Discussion

We performed short-read and long-read genome sequencing, as well as peripheral blood transcriptome analysis, in a series of five patients without genetic diagnosis after extensive

Fig. 5 De novo variant analysis identifies a mosaic splicing variant in *KMT2D*. **A** Clinical data. Photograph represent proband C at 10 years of age. Note the flat face, abnormal eyebrows, ptosis, long and everted palpebral fissures, long philtrum and dysplastic ear lobes. **B** Short read alignments show a de novo 22 bp deletion: NM_003482.4:c.10741-23_10741-2del, with an allelic ratio evocative of mosaicism. The paternally inherited SNP rs3782357 appears on the same reads as the de novo indel indicating that the de novo event occurred on the paternal haplotype. Blue bars refer to C, green refer to A, red refer to T and orange refer to G. **C** Transcriptomic consequences: presence of intronic reads compared to controls indicating a partial retention on intron 39



investigations including exome sequencing. These combined approaches allowed us to identify the cause of the disease in three families. Of note, it is not possible to infer from this study the performance of these approaches on larger cohorts of developmental diseases with negative exome analysis. Indeed, this study, based on only five families, is underpowered to accurately assess the precise contribution of these techniques. Furthermore, the particularly selected clinical context of these five patients may have introduced a bias, resulting in a higher yield than analyses performed on a larger scale. Nevertheless, this study provides insights into potential diagnostic opportunities beyond the current standard approaches.

The main input from long-read genome sequencing (LR-GS) was the identification of a complex, mostly balanced de novo structural variant disrupting the *NSD1* gene. Inspection of the short-read alignments and retrospective analysis of structural variants on short-read data using SR-specific callers (Manta and Canvas) confirmed the presence of this complex variant. While it could have been detected using short-read data alone, calling and interpreting structural variants on SR-GS can be challenging due to numerous artifacts. In contrast, data obtained from LR-GS had fewer artifacts and was more readily accessible. The SvAnna software, which ranks variants detected in LR-GS according to the patient's phenotype and the effect of variants on genes, was released during the writing of this manuscript, after the identification of the complex inversion in the proband (E), and was retrospectively tested. With minimal phenotypic information (2 HPO terms) and no information on parental transmission (i.e., de novo), SvAnna ranked the two detected inversions as the top events from the approximately 26,000 events detected in proband E, indicating good performance without the need for fine adjustment. However, using SvAnna in both negative patients did not help prioritizing candidate variants that would have been missed by our analysis strategy. Additionally, the recent version 2 of Sniffles features the ability to merge individual variant calls into a single multi-VCF file, simplifying the analysis of cohorts or families. Overall, new and powerful tools are emerging to facilitate the analysis of structural variants detected in LR-GS. While the chr5 inversion could have been detected using SR-GS, other variant types, such as mobile element insertions, can be particularly challenging to detect in SR data. These insertions appear to be a recurrent source of novel diagnoses in rare disease cases (Hiatt et al. 2021; Walsh et al. 2021). It is likely that future long-read studies on larger cohorts will provide a better understanding of the contribution of LR-GS. Additionally, bioinformatics methods are likely to improve and approach the accessibility and robustness of short-read methods.

Genome sequencing offers the access to virtually all non-coding regions. However, we are still facing significant limitations in non-coding variant interpretation. One increasingly accessible source of pathogenic non-coding regions is the creation of neo-exons by deep intronic variants. A recent study showed that out of five unsolved Cornelia de Lange cases, two were caused by an *NIPBL* de novo variant leading to a frameshift neo-exon (Coursimault et al. 2022). However, the global contribution of deep intronic variants leading to deleterious neo-exons in developmental disorders is unknown and could be the subject of future investigations. In this study, we identified a homozygous *KLHL7* variant leading to an out-of-frame neo-exon, which decreased gene expression in two siblings with features of Perching syndrome.

Finally, the third variant detected was a post-zygotic *KMT2D* variant affecting the paternal allele. The identification of a mosaic variant using GS after negative results from exome and panel sequencing may seem paradoxical, but the high homogeneity of coverage in SR-GS makes it a useful tool for detecting mosaics with relatively high allelic ratios (which have the potential to cause a clinical phenotype). Additionally, bioinformatics pipelines have improved in recent years, allowing for very robust detection of SNV-indels, particularly using the DeepVariant algorithm (Poplin et al. 2018), putatively contributing to explain this novel diagnosis missed by exome and gene panel sequencing although theoretically detectable.

Interestingly, among the three new diagnoses made, all had visible consequences on the blood-derived transcriptome, which allowed us to validate these diagnoses. This was particularly important in the diagnosis of *KLHL7*. The peripheral blood transcriptome appears to provide independent evidence of pathogenicity, making it a good biomarker in certain situations, particularly for the assessment of variants in non-coding regions. It is a universal functional test with the obvious limitations of the expression of the gene of interest and the tissue evaluated. Of note, RNAseq data can also be used primarily to identify candidate variants. Here, we interpreted DNA sequencing data first, mainly because the strategies available to interpret variants from RNAseq data require more comprehensive bioinformatics pipelines including comparison with a large number of samples, which was not available to us (Yépez et al. 2022).

For two patients with strongly suspected developmental disease, our multimodal analysis of SNV/indels, SVs, and other types of disease-associated variations (i.e., UPD and STR expansions) failed in identifying a diagnosis. It should be noted that one of these two patients had a Cornelia de Lange-like syndrome. This syndrome is associated with a strikingly high rate of tissue-specific mosaicism, often with no detectability of the variation in blood and low mosaicism in other tissues, such as saliva (Latorre-Pellicer et al. 2021). A mosaic in saliva was previously excluded in this patient by a deep coverage gene panel analysis, but our understanding of this type of mosaicism is still incomplete and it is possible that an even more confined mosaicism is responsible for the symptoms. In complement to our analysis, we used the Phenomizer tool (Köhler et al. 2009) to help us to rank known Mendelian disorders by confronting Human Phenotype Ontology (HPO) codes associated to each patient's phenotype. This phenotype-first approach indeed identified Cornelia de Lange syndrome as a candidate disorder for Proband A, along with others. We then got back to the sequencing data and did not identify any candidate variant among the list of disorders with lowest p-values (< 0.05) following prioritization based on HPO information (Supplementary Table 1). In addition to variants not present in our data, it is also possible that our data does contain the diagnosis, but our analysis failed to identify it. For these cases, it is likely that new diagnoses will emerge in the coming years with the evolution of bioinformatics procedures, as well as scientific knowledge (loci and genes not implicated, and new types of variants). Finally, technological advances are also emerging, such as the outstanding new sequencing protocol called Circular Consensus Sequencing (Wenger et al. 2019) with long molecules and a very low error rate per base, resulting in a "near perfect genome" (Olson et al. 2022). This technique has the ability to detect virtually all types of variants in a single experiment (Hiatt et al. 2021) and will hopefully both simplify genomic testing and increase diagnosis rates.

In summary, a thorough, multi-technique analysis of patients with rare diseases that remained unsolved after exome sequencing resulted in a diagnostic yield of 3/5. It is likely that in the future, with the democratization and automation of these techniques, the vast majority of patients with a suspected rare genetic disease will have access to a molecular diagnosis.

Web resources

- Custom IGV-based filtration interface:
- https://github.com/francois-lecoquierre/genomic_shortcuts/
- Custom tool to detect Uniparental Disomies: UPD_plotter
- https://github.com/francois-lecoquierre/UPD_plotter/
- Nf-core rnaseq pipeline:
- https://nf-co.re/rnaseq
- UCSC session showing NSD1 complex inversion:
- http://genome-euro.ucsc.edu/s/francois.leco/RRMUT_NSD1_trioE

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-023-02553-1.

Acknowledgements We thank all the patients and their families as well as their referring physicians for their participation to this study. Europe gets involved in Normandie with the European Regional Development Fund (ERDF). This work was generated within the European Reference Network for Developmental Anomalies and Intellectual Disability. This work was performed in the framework of FHU-G4 Génomique.

Funding This study was supported by a grant from GIRCI Nord-Ouest (AAP-AE_19-36). Collaboration CEA-DRF-Jacob-CNRGH—CHU de Rouen. This study was co-supported by the European Union and Région Normandie in the context of Recherche Innovation Normandie (RIN2018). This work did benefit from support of the France Génomique National infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This study was performed in line with the principles of the Declaration of Helsinki. The study was approved by the CPP Ouest V (20/043–2) ethics committee.

Consent to participate Written informed consent was obtained from the parents of each included individual.

Consent to publish The authors affirm that human research participants provided informed consent for publication of the images in Figs. 3, 4 and 5.

References

100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, et al (2021) 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. N Engl J Med 385:1868–1880. https://doi.org/10.1056/NEJMoa2035790

Angius A, Uva P, Buers I et al (2016) Bi-allelic mutations in KLHL7 cause a Crisponi/CISS1-like phenotype associated with early-onset retinitis pigmentosa. Am J Hum Genet 99:236–245. https://doi.org/10.1016/j.ajhg.2016.05.026

Baux D, Van Goethem C, Ardouin O et al (2021) MobiDetails: online DNA variants interpretation. Eur J Hum Genet 29:356–360. https://doi.org/10.1038/s41431-020-00755-z

Beyter D, Ingimundardottir H, Oddsson A et al (2021) Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet 53:779–786. https://doi.org/10.1038/s41588-021-00865-4

Bruel A-L, Bigoni S, Kennedy J et al (2017) Expanding the clinical spectrum of recessive truncating mutations of KLHL7 to a Bohring-Opitz-like phenotype. J Med Genet 54:830–835. https://doi.org/10.1136/jmedgenet-2017-104748

Clarke J, Wu H-C, Jayasinghe L et al (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270. https://doi.org/10.1038/nnano.2009.12

Colin E, Duffourd Y, Tisserant E et al (2022) OMIXCARE: OMICS technologies solved about 33% of the patients with heterogeneous rare neuro-developmental disorders and negative exome sequencing results and identified 13% additional candidate variants. Front Cell Dev Biol 10:1021785. https://doi.org/10.3389/fcell.2022.1021785

Coursimault J, Cassinari K, Lecoquierre F et al (2022) Deep intronic NIPBL de novo mutations and differential diagnoses revealed by whole genome and RNA sequencing in Cornelia de Lange syndrome patients. Hum Mutat. https://doi.org/10.1002/humu.24438

Danis D, Jacobsen JOB, Balachandran P et al (2022) SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. Genome Med 14:44. https://doi.org/10.1186/s13073-022-01046-6

De Coster W, De Rijk P, De Roeck A et al (2019) Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res 29:1178–1187. https://doi.org/10.1101/gr.244939.118

De Coster W, Weissensteiner MH, Sedlazeck FJ (2021) Towards population-scale long-read sequencing. Nat Rev Genet 22:572–587. https://doi.org/10.1038/s41576-021-00367-3 de Sainte Agathe J-M, Filser M, Isidor B et al (2023) SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Hum Genom 17:7. https://doi.org/10.1186/s40246-023-00451-1

Deciphering Developmental Disorders Study (2017) Prevalence and architecture of de novo mutations in developmental disorders. Nature 542:433–438. https://doi.org/10.1038/nature21062

Dong Z, Yan J, Xu F et al (2019) Genome sequencing explores complexity of chromosomal abnormalities in recurrent miscarriage. Am J Hum Genet 105:1102–1111. https://doi.org/10.1016/j.ajhg.2019.10.003

Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. https://doi.org/10.1126/science.1162986

Ellingford JM, Ahn JW, Bagnall RD et al (2022) Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med 14:73. https://doi.org/10.1186/s13073-022-01073-3

Gilissen C, Hehir-Kwa JY, Thung DT et al (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344–347. https://doi.org/10.1038/nature13394

Hiatt SM, Lawlor JMJ, Handley LH et al (2021) Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. HGG Adv 2:100023. https://doi.org/10.1016/j.xhgg.2021.100023

Jeffries L, Olivieri JE, Ji W et al (2019) Two siblings with a novel nonsense variant provide further delineation of the spectrum of recessive KLHL7 diseases. Eur J Med Genet 62:103551. https://doi.org/10.1016/j.ejmg.2018.10.003

Kaplanis J, Samocha KE, Wiel L et al (2020) Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586:757–762. https://doi.org/10.1038/s41586-020-2832-5

Köhler S, Schulz MH, Krawitz P et al (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet 85:457–464. https://doi.org/10.1016/j.ajhg.2009.09.003

Kovanda A, Zimani AN, Peterlin B (2021) How to design a national genomic project-a systematic review of active projects. Hum Genom 15:20. https://doi.org/10.1186/s40246-021-00315-6

Latorre-Pellicer A, Gil-Salvador M, Parenti I et al (2021) Clinical relevance of postzygotic mosaicism in Cornelia de Lange syndrome and purifying selection of NIPBL variants in blood. Sci Rep 11:15459. https://doi.org/10.1038/s41598-021-94958-z

Lee H, Huang AY, Wang L-K et al (2020) Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genet Med 22:490–499. https://doi.org/10.1038/s41436-019-0672-1

Lévy Y (2016) Genomic medicine 2025: France in the race for precision medicine. Lancet 388:2872. https://doi.org/10.1016/S0140-6736(16)32467-9

Mantere T, Kersten S, Hoischen A (2019) Long-read sequencing emerging in medical genetics. Front Genet 10:426. https://doi.org/10.3389/fgene.2019.00426

Olson ND, Wagner J, McDaniel J et al (2022) PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genom 2:100129. https://doi.org/10.1016/j.xgen.2022.100129

Pauper M, Kucuk E, Wenger AM et al (2021) Long-read trio sequencing of individuals with unsolved intellectual disability. Eur J Hum Genet 29:637–648. https://doi.org/10.1038/s41431-020-00770-0

Poplin R, Chang P-C, Alexander D et al (2018) A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36:983–987. https://doi.org/10.1038/nbt.4235

Quinodoz M, Peter VG, Bedoni N et al (2021) AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. Nat Commun 12:518. https://doi.org/10.1038/s41467-020-20584-4

Richards S, Aziz N, Bale S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–424. https://doi.org/10.1038/gim.2015.30

Souche E, Beltran S, Brosens E et al (2022) Recommendations for whole genome sequencing in diagnostics for rare diseases. Eur J Hum Genet 30:1017–1021. https://doi.org/10.1038/s41431-022-01113-x

van El CG, Cornel MC, Borry P et al (2013) Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. Eur J Hum Genet 21:580–584. https://doi.org/10.1038/ejhg.2013.46

Walsh T, Casadei S, Munson KM et al (2021) CRISPR-Cas9/long-read sequencing approach to identify cryptic mutations in BRCA1 and other tumour suppressor genes. J Med Genet 58:850–852. https://doi.org/10.1136/jmedgenet-2020-107320

Wenger AM, Peluso P, Rowell WJ et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wu Z, Jiang Z, Li T et al (2021) Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. Nat Commun 12:6501. https://doi.org/10.1038/s41467-021-26856-x

Yépez VA, Gusic M, Kopajtich R et al (2022) Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med 14:38. https://doi.org/10.1186/s13073-022-01019-9

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