Genetics in Medicine (2026) 28, 101661

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine

ARTICLE

Polygenic variants in DNA repair genes are associated with neurodevelopmental disorders, regression and increased burdens of somatic variants and short tandem repeat expansions

Maria Cerminara[1,2], Giovanni Spirito[3], Luca Pandolfini[4], Silvia Boeri[5], Giulia Rosti[2,6], Margherita Mancardi[5], Livia Pisciotta[7,8], Marco Fontana[1,2], Alessandra Bianchi[1,2], Iris Chen[2], Loretta Ferrera[1], Francesco Caroli[1], Marco Di Duca[1], Andrea Cavalli[9,10,11], Maria Teresa Divizia[6], Elisa De Grandis[2,5], Silvia Casabona[5], Sara Trova[3], Diego Vozzi[4], Antonio Amoroso[3], Laure Obino[12], Remo Sanges[4,13], Manuela Vecchi[3], Lino Nobili[2,5], Federico Zara[1,2], Stefano Gustincich[3,4,*], Aldamaria Puliti[1,2,*]

ARTICLE INFO

Article history:
Received 16 December 2024
Received in revised form
14 November 2025
Accepted 24 November 2025
Available online 8 December 2025

Keywords:
FAN1
Genome instability
Long-read sequencing
Neurodevelopmental disorders
Short tandem repeat

ABSTRACT

Purpose: Developmental regression, characterized by the loss of acquired milestones, occurs in some individuals with neurodevelopmental disorders (NDDs); yet, its molecular basis remains unclear. Studies suggest that DNA damage repair (DDR) genes, such as *FAN1*, may protect against neurological dysfunction by modulating the somatic stability of short tandem repeats (STRs). This study explores the contribution of DDR gene variants in NDD cases presenting with regression.

Methods: We analyzed 1087 NDD patients, focusing on those carrying variants in DDR genes and presenting regression. We assessed the sensitivity to DNA damage using mitomycin C on lymphoblastoid cells. Somatic variants and STR expansions were evaluated through high-depth short-read genome sequencing. To further investigate the pathogenetic role of STR expansions, we performed long-read genome sequencing on the most severely affected proband.

Results: Probands with regression carried multiple DDR gene variants, several within the Fanconi anemia pathway. Their lymphoblastoid cells showed increased sensitivity to mitomycin C-induced cytotoxicity compared with parental and control samples. Probands with severe phenotypes and regression exhibited an accumulation of somatic variants and STR instability, enriched in neurodevelopmental genes.

Conclusion: Our findings suggest that polygenic DDR gene variants may contribute to developmental regression in NDDs by promoting the accumulation of somatic variants and STR expansions.

The Article Publishing Charge (APC) for this article was paid by the IRCCS Istituto Giannina Gaslini.

Maria Cerminara and Giovanni Spirito contributed equally to this work.

*Correspondence and requests for materials should be addressed to Aldamaria Puliti. *Email address:* aldamaria.puliti@unige.it OR Stefano Gustincich. *Email address:* stefano.gustincich@iit.it

Affiliations are at the end of the document.

doi: https://doi.org/10.1016/j.gim.2025.101661

Introduction

Neurodevelopmental disorders (NDDs) represent a broad spectrum of conditions characterized by impaired brain function and development, leading to difficulties in cognitive, social, motor, and emotional functioning.[1,2] Although the etiology of NDDs remains multifaceted, recent research suggests a strong genetic component,[2-6] highlighting a complex interplay between inherited and de novo genetic variants with moderate to high penetrance.[2,7,8]

Currently, thousands of genes have been implicated in NDD susceptibility.[9,10] Although most NDD-associated genes are involved in neuronal function, intriguingly, a subset participates in DNA repair.[11] This pathway is essential for maintaining genomic integrity both during embryonic development and throughout the lifespan of an organism, functioning in proliferating neuronal progenitors, as well as in postmitotic neurons.[12] Dysregulation of DNA damage repair (DDR) functions has been previously reported in postmortem analysis of individuals with autism spectrum disorder (ASD) at older ages[13] and has been proposed as a contributing factor to somatic mosaicism and focal cortical dysplasia.[14] Recently, alterations in DDR pathways have been associated with the most severe ASD subtypes, as identified through integrated clinical and molecular stratification (Zahiri J, Mirzaie M, Duan K, et al. Beyond the spectrum: subtype-specific molecular insights into autism spectrum disorder via multimodal data integration. *medRxiv*. 2024. https://doi.org/10.1101/2024.09.17.24313857). Notably, some DDR genes also act as modifiers of short tandem repeat (STR) expansions, an emerging mutagenic mechanism linked to ASD etiology. Trost et al[15] analyzed STR expansions ranging from 2 to 20 base pairs in the genomes of families with ASD and control individuals. They found that rare STR expansions were more prevalent in individuals with ASD, especially in regions critical for nervous system development.[15] Similarly, Mitra et al[16] reported that de novo STR expansions were significantly larger in ASD individuals compared with their parents and were enriched in fetal brain regulatory regions. Although the specific STR expansions identified in these studies did not overlap and the presence of somatic STR variation was left unexplored, these findings reveal a previously underappreciated source of genomic variations contributing to ASD.

Among genes involved in DDR and STR expansion control, *FAN1* (HGNC:29170) has emerged as an important modifier of ASD, epilepsy, and schizophrenia, influencing disease onset and susceptibility.[9,10] *FAN1* is a Fanconi-anemia-associated DNA nuclease that promotes the repair of DNA damaged by interstrand cross-links in a manner dependent on its ability to accumulate close to DNA damage sites and on monoubiquitylation of the fanci-fancd2 complex, a central event in the activation of the Fanconi anemia (FA) pathway.[17] Importantly, *FAN1* has been established as a strong modifier of Huntington disease (HD)(OMIM 143100), with a role in slowing CAG expansions and delaying disease onset.[18-20] Evidence shows that *FAN1* contributes to genomic stability maintenance by engaging in DNA repair mechanisms, thereby mitigating the aberrant expansion of STRs.[21] Indeed, findings suggest that *FAN1* safeguards against somatic CAG expansions in HD cell lines[19] and CGG expansions in *FMR1* transgenic mice,[22] highlighting its protective role against STR expansions in diverse disease contexts. Recently, 15 highly polymorphic STRs were identified through computational analysis of DNA sequences from over 700 000 participants in UK Biobank and the All of Us Research Program and some of them associated with human diseases.[23] Interestingly, *FAN1* was identified as one of the 7 loci at which inherited variants modulate the high somatic STR instability within the *TCF4* (HGNC:11634), confirming *FAN1* as one of the most relevant modifiers.

The mechanisms and the extent of *FAN1* involvement in causing or modifying NDD phenotypes remain poorly understood, as well as whether these involve regulation of STR expansions at both germinal and somatic levels. Moreover, it is plausible to speculate that additional DDR NDD genes and STR modifiers may act through similar mechanisms.[11] A subset of individuals with NDDs exhibits a phenomenon known as developmental regression, in which previously acquired skills and abilities deteriorate over time.[24] Although the molecular basis of regression remains largely unknown, ultrarare variants in DDR genes have recently been identified in seventeen individuals with NDDs and/or regression. In 7 cases, these variants affect genes involved in the FA complex.[25] However, the mechanisms by which dysfunction of these genes might contribute to regressive NDD phenotypes remain largely unexplored. To date, no studies have investigated the presence of germinal or somatic STR instability in these individuals, despite the FA complex being extensively associated with instability in cancer.[26]

In this study, we analyzed a cohort of 1087 individuals with NDD and identified 6 probands who showed a complex neurodevelopmental phenotype characterized by ASD and psychomotor regression in the presence of polygenic pathogenic variants in DDR genes, including *FAN1*. To assess the functional impact of these variants, we performed mitomycin C (MMC)-induced cytotoxicity assays in 4 probands, which revealed increased DNA damage in all patient-derived cell lines, indicating impaired DNA repair. We then conducted

high-coverage short-read genome sequencing (srGS) and observed a notable increase in STR amplifications in 1 of the probands. In addition, 3 of the 4 trios showed a higher burden of somatic STRs in the proband compared with their parents. To further investigate STR instability, we carried out high-coverage long-read GS in 1 trio, selected for the severity of the proband's phenotype. This analysis uncovered an elevated number of putative somatic STR expansions in the proband compared with both parents.

Together, these findings expand our understanding of NDD pathophysiology, showing that polygenic variants in DDR genes may give rise to the regression phenotype. Intriguingly, almost all cases carry variants in genes involved in the FA pathway, including *FAN1*.

Materials and Methods

Clinical assessment

Patients were initially evaluated through a medical history, a neurological-behavioral examination, a developmental/cognitive assessment, and a basic metabolic screening. They were also evaluated using genetic tests for common genetic causes of NDD, tests for Fragile X syndrome (OMIM 300624), and chromosomal microarray analysis (CMA) with a whole-genome 180 K Agilent array (Human Genome CGH Microarray, Agilent Technologies). The assessment of developmental regression was conducted using a multifaceted approach, incorporating caregiver reports, behavioral observations, and motor and neurological examinations in comparison with prior evaluations. According to recent findings, the manifestation of regression was indicated by the loss of social and/or communication skills, as well as by impairment in motor abilities.[24]

This study was reviewed and approved by the Ethics Committee of the Italian Regione Liguria (P.R. 399REG2017). Written informed consent to participate in this study was provided by all participants or their legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s) legal guardian/next of kin, for the publication of data included in this article.

Exome/genome sequencing and variants prioritization

DNA from blood lymphocytes was obtained from the patients and their parents. ES was carried out by Italian Institute of Technologies in Genoa, Italy, using methods already described.[27] Sequencing provided a 60x medium coverage. Fastq reads were aligned to the reference genome (GRCh37). The resulting file was used for manual variant assessment. We selected variants on the basis of specific criteria, as also reported by Cerminara et al[27]: (1) variants with a frequency $\leq$ 3% in ExAC, GnomAD,[28] and 1000G databases; (2) only exonic, splicing, nonsynonymous, and stopgain variants with a good coverage; (3) variants with a deleterious effects predicted with CADD (score $\geq$ 20), SIFT, PolyPhen-2 and AlphaMissense. We discarded those variants found in more than 1% of in-house controls (ie, not diagnosed with ASD or any neuropsychiatric disease, aged from 18 to 65 years old). Finally, we further filtered the variants by considering annotation in public databases such as OMIM,[29] ClinVar,[30] and Mouse Genome Database.[31] Selected variants were classified using the Franklin Genoox platform (https://franklin.genoox.com/), which integrates advanced annotations, population-specific allelic frequencies, in silico prediction tools to obtain pathogenicity scores and the classification using the American College of Medical Genetics and Genomics (ACMG)/Association for Molecular Pathology (AMP) guidelines.[32,33]

To unveil the contribution of multiple genes to the patient's phenotype, we used a platform for the prediction and exploration of candidate disease-causing oligogenic variant combinations (Oligogenic Resource for Variant Analysis [ORVAL])[34] following authors recommendations and as described in the Supplemental document file (Supplemental Figure 1).

The contribution of DNA repair genes to disease was investigated by searching for variants in the genes associated with the Gene Ontology (GO) term "DNA repair" and focused on genes reported as modifiers for HD disease, according to recently published articles (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, Lee JM, McLean ZL, et al. Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease reveal shared and tissue-specific effects. *bioRxiv*. 2024. https://doi.org/10.1101/2024.06.10.597797),[35] and those shown to modulate *TCF4* repeat instability in blood cells of over 700,000 participants in UK Biobank and the All of Us Research Program according to Hujoel et al.[23] More in detail, we searched for patients with at least 2 variants, de novo or inherited, located in 2 genes that were predicted to interact by STRING and/or ORVAL and that could additively contribute to the patients' phenotype.

Genome sequencing

Genomic DNA (gDNA) was extracted from peripheral blood of all sequenced individuals. srGS of gDNA was conducted on an Illumina NovaSeq 6000 platform (Illumina Inc) to produce paired 151 base-pair reads. The library preparation and sequencing of gDNA utilized barcode adapters. Specifically, 350 ng of gDNA was prepared using the Illumina DNA PCR-Free Prep, Tagmentation library preparation kit (Cat number 20041795), and IDT for Illumina DNA/RNA Unique Dual Indexes Set A and B, Tagmentation (96 Indexes, Cat numbers 20027213 and 20027214). The library preparation was carried out with a T100 thermal cycler instrument (Bio-Rad) and standard laboratory equipment. Library quantification was performed

with KAPA qPCR Library Quantification Kit (Roche, Cat number 0 7960140001). Libraries were pooled by volume, adjusted for the index correction factor according to Illumina's instructions to ensure balanced sample coverage across each run.

DNA extraction and preparation

High-molecular-weight DNA was extracted from blood samples using the Monarch HMW gDNA Extraction kit for Cells and Blood (New England Biolabs) following the manufacturer's protocol. The gDNA was fragmented to a modal size of ~20k using mechanical shearing on a g-TUBE (Covaris), then further purified by AMPureXP beads cleanup (Beckman Coulter).

Long-read sequencing

For long-read sequencing, we used a PromethION 24 instrument (PRO-PRC024) according to the guidelines provided by Oxford Nanopore Technologies. We used R10.4.1 chemistry flowcells (FLO-PRO114M) and the Ligation Sequencing Kit (SQK-LSK114) starting from 1ug of gDNA. Basecalling was performed using Dorado version 0.3.4 in Super-High Accuracy mode (dna_r10.4.1_e8.2_400bps_sup@v4.2.0). Only reads with a Q- score above 10 were included in the subsequent analysis.

Cell culture

Peripheral blood mononuclear cells were obtained from all affected individuals (F36A, F74A, F82A, and F19A), the parents of 3 of them (F36A, F74A, and F82A), and an unaffected individual. From isolated peripheral blood mononuclear cells, Epstein-Barr-virus-transformed lymphoblastoid lines were acquired and grown in RPMI1640 (Euroclone) with 10% fetal bovine serum (FBS; Sigma Aldrich, Merck Life Science S.r.l.), 50 units/mL penicillin-streptomycin (Euroclone), and 2 mmol/L glutamine. The HEK293 cell line was previously purchased by ATCC and routinely cultured in advanced Dulbecco's modified Eagle's medium (Euroclone) supplemented with 10% FBS, 1% L glutamine, 100 units/mL penicillin, and 100 µg/mL streptomycin. Cells were maintained at 37 °C in a 5% CO_2 humidified atmosphere.

MMC sensitivity assay

Lymphoblastoid cells were seeded in T-75 cm(2) flasks (5×10^5 cells) with 5 mL complete medium (RPMI1640, 10% fetal bovine serum, 50 units/mL penicillin-streptomycin, 2 mmol/L glutamine) and treated for 5 days with mitomycin C (M4287, Merck Life Science) at increasing concentration (0, 0.075, 0.15, 0.30, 0.60, 1, 1.5, and 2 µmol/L). Cell viability was then assessed by using an automatic cell counter (LUNA-II Automated Cell Counter, Twin Helix) after Trypan Blue solution (0.4%) cell staining (Twin Helix). For each MMC concentration, 7 independent experiments run in triplicate were performed, and the growth percentage statistically analyzed by the Kruscal-Wallis nonparametric test followed by the Dunn's Multiple Comparison Test. The Mann-Whitney test was used to compare 2 mean populations. Statistical analyses were performed by using GraphPad Prism 9 software. Data are expressed as mean ± SEM and P value < .05 was considered significant.

Downregulation of *FAN1* in HEK293 cells by transfection of siRNA

To downregulate *FAN1*, HEK293 cells were seeded in 60-mm diameter dishes and cultured in antibiotic-free medium (Opti-MEM; Gibco, Life Technologies) for 24 hours and then, at 80% confluency, were transfected with 5-nmol/L small-interfering RNA (siRNA) duplexes, using Lipofectamine 2000 (Invitrogen, Life Technologies) as the transfection agent. We used a pool of 2 commercially available siRNAs complementary to *FAN1* (s22650 and s22649; Invitrogen, Thermofisher Scientific, Life Technologies), and to *GAPDH* (HGNC:4141) (Thermofisher Scientific, Life Technologies) as a positive control. After 6 h of cell incubation at 37 °C, the medium was removed and replaced with fresh Dulbecco's modified Eagle's medium plus serum and antibiotics. After additional 48/72 hours, we lysed transfected cells in Trizol (Invitrogen, Life Technologies) and tested gene silencing by qRT-PCR. Briefly, cDNA was prepared from 1.5 µg RNA using the iScript Reverse Trascription Supermix with combined oligo(dT) and random primers (Bio-Rad, Segrate). Gene expression levels were quantified using the following gene-specific primers: *FAN1* (5′-ATTGACGAATGCAGGCTTTCTAC-3′/5′-AAGTGGAAGGTCTTGGCTAGGGA-3′); *GAPDH* (5′-GTCTCCTCTGACTTCAACAGCG-3′/5′-ACCACCC TGTTGCTGTAGCCAA-3′). The real-time RT-PCR was run on a CFX96 Touch Real-Time PCR Detection System instrument (Bio-Rad). PCR amplifications were performed in triplicate using 1:2 diluted cDNA in a 20 µL final reaction mixture, using the iQ SYBR Green Supermix (Bio-Rad). Data were analyzed by the CFX Maestro Software (Bio-Rad). The mRNA expression was calculated using the ΔΔCt method.[36] *FAN1* expression levels of HEK293 cells were normalized to GAPDH and represented as relative to the nonsilenced HEK293 cells (normalized fold change).

Immunoblot analysis

We prepared protein extracts by lysing cells (HEK293, silenced HEK293 and lymphoblastoid cells) in a denaturing buffer (SDS 2%; Tris-HCl 62.5 mmol/L; pH 6.8). Supernatant protein concentration was calculated using the BCA assay (Thermofisher) following the manufacturer's

instructions. Total proteins were separated by means of a precast 8% Bis-Tris SDS-PAGE gel (GenScript) using SDS Running buffer (6.06g Tris Base; 10.46g MOPS; 1.0g SDS; and 0.3g EDTA). The concentration of proteins in each sample fell in the linear portion of the curve. A triplicate analysis for each sample was performed. Gels were transferred onto 0.2 µm nitrocellulose membrane by using the Trans-Blot Turbo Midi instrument (Bio-Rad). Electroblotted proteins were monitored using Ponceau Red S staining. Membranes were then incubated with the following antibodies: rabbit polyclonal anti-fan1 (MTMR15, 17600-1-AP) antibody (Proteintech) and mouse monoclonal antibeta actin (Abcam). After incubation with appropriate peroxidase-coupled secondary antibodies, protein bands were detected by using a Western blotting detection system (Supersignal West Atto, Thermofisher Scientific). Bands were detected and analyzed for density using an enhanced chemiluminescence system (Uvitech). Bands of interest were normalized for Gapdh level in the same membrane. Western blot quantifications were performed using densitometry measures and the ImageJ software.

Genome alignments and downsampling alignments (short-read sequencing)

We mapped raw fastq GS reads to the reference genome (UCSC, version GRCh38)[37] with BWA.[38] Quality and coverage data relative to the alignments are reported in Supplemental Table 1.

To quantify the amount of somatic short variants (SNVs and Indels) in our samples, we addressed the issue of variability in the number of somatic events due to differences in sequencing depth. Because our samples were sequenced at different depths, we downsampled the aligned data to ensure consistent coverage across samples within the same family. This was achieved by randomly selecting the appropriate fraction of mapped reads from each sample using Samtools,[39] thereby standardizing the coverage. Downsampling was performed before quantifying somatic variation within genomic repeats from the aligned short-read data. Coverage data relative to the downsampled alignments are reported in Supplemental Table 2 and 3.

Identification of STR expansions from short-read GS data

Upon mapping raw fastq reads to the reference genome (UCSC, version GRCh38)[37] with BWA,[38] we run ExpansionHunter Denovo[40] to compute a genome-wide STR profile for each individual. ExpansionHunter Denovo is a tool able to detect information about reads that originate in STRs longer than the read length (in-repeat reads, IRR). We analyzed GS data of each member of our trios alongside 132 healthy individuals who were sequenced using the same protocol and library preparation. We calculated the total amount of normalized IRR for each sample and computed a z-score on the total amount of IRR for each analyzed sample. Finally, z-scores have been transformed to P values and corrected (FDR) with SciPy.

Identification of somatic variation within genomic repeats from short-read GS data

To estimate the total number of somatic structural variants (SNVs and Indels) we mapped raw fastq reads to the reference genome (UCSC, version GRCh38)[37] with BWA, ran Mutect2 on the downsampled alignment files, and filtered the identified variants to retain only short variants that (1) were labeled as "PASS," (2) were associated with a TLOD score above 6.3, (3) were found in a locus supported by a number of reads equal or above the mean sequencing depth of each trio, (4) were heterozygous, (5) were found in autosomes, (6) were absent in a panel of normal produced with 126 healthy individuals, (7) were associated to an allelic frequency below 0.1; (8) were found within genomic repeats annotated in RepeatMasker. Genomic coordinates of all repeats annotated in the human genome (RepeatMasker, GRCh38) were retrieved through UCSC.[37]

SNV and indel signature enrichments were estimated with the R package mutSigExtractor (v1.28), using signatures from COSMIC.[41]

Identification of somatic structural variants from short-read GS data

To estimate the total number of somatic structural variants (SVs) we mapped raw fastq reads to the reference genome (UCSC, version GRCh38)[37] with BWA,[38] ran svABA[42] on the downsampled alignment files, and filtered the identified variants to retain only break ends (BNDs) that (1) were labeled as "PASS," (2) were associated with a TLOD score above 6.3, (3) were found in a locus supported by a number of reads equal or above the mean sequencing depth of each trio, (4) were heterozygous, (5) were found in autosomes, (6) were absent in the DBSNP database, (7) were associated to an allelic frequency below 0.1, (8) were found within genomic repeats annotated in RepeatMasker.[43] Genomic coordinates of all repeats annotated in the human genome (RepeatMasker, GRCh38) were retrieved through UCSC.[37]

Quantification of STR expansions from long-read GS data

We obtained fastq files from long-read sequencing of the 3 members of the F36 trio and aligned them to the reference genome (UCSC, version GRCh38)[37] with minimap2.[44] Upon mapping, we obtained a mean depth of 87.4X, 87.7X, and 78.2X, respectively, for the proband, the mother, and the father, with an average N50 value of 8746 among the 3 members of the trio. Quality and coverage data relative to the fastq and alignment files are detailed in Supplemental

Table 4. To quantify the total amount of long-reads supporting STR expansions we used the tool Straglr[45] (v1.4.1, –exclude telomeres, centromeres –min_support 1 –min_ins_size 10) upon aligning raw fastq reads to the reference genome (version GRCh38) with Minimap2[44] (v2.26, -Y -t 5 -ax map-ont). We then counted the number of reads supporting STR expansions (read_status=full) and normalized the number for the total number of mapped reads. A recent study on repeat instability in blood cells of over 700,000 participants in UK Biobank and the All of Us Research Program pointed to a group of genes as particularly prone to amplification.[23] As quality control of the analysis, we searched for intersections between locations and sequence contexts of CAG repeat loci expanded to ≥45 repeat units in ≥5 UKB participants (Hujoel et al[23]; Supplemental Table 5) and loci with long-reads supporting repeat expansions found in the members of family F36 were also searched.

Identification of somatic STR expansions from long-read GS data

We aligned long-read sequencing fastq reads of all 3 members of F36 trio to the reference genome (UCSC, version GRCh38)[37] with both Minimap2[44] (long-read aligner, required for the analysis with CuteSV[46]; average mapping depth can be viewed in Supplemental Table 4) and with LASTAL, which estimates the rates of insertion, deletion, and substitutions between fastq reads and the reference genome, and is essential for the analysis with the tool Tandem-genotypes.[47]

Tandem-genotypes can be used to calculate the number of repeat copies (STR copy number) present in each read for each locus of interest (±60 bp upstream and downstream from each input locus) compared with the reference genome. In other words, for each input locus we obtained a list of STR copy numbers, in which each reported value corresponds to a mapped read. We used 3 sets of loci of interest: (1) a comprehensive set of 5,683,690, including tandem repeats with unit length from 1 to 13 made with RepeatMasker and obtained from the tandem-genotypes GitHub web page, (2) a set of 59 known STR pathogenic loci,[15,47-49] (3) a set comprising 57 loci relative to STR regions expanded only in ASD and not in healthy controls.[15] For each set of STR loci, we wanted to detect potentially germinal, mosaic, and somatic STR expansions in the proband.

To detect reads containing "somatic STR expansions" for each set of loci, we selected only loci covered by at least 30 reads in all 3 samples and characterized by a normal distribution of STR copy number (Shapiro test, P value < .05). We then calculated a z-score for each read STR copy number and considered all reads containing an STR copy number associated with a z-score above 10 as reads containing a somatic STR expansion (outlier reads). We then counted the number of normalized outlier reads for each locus and normalized it for the total number of reads mapping within the locus.

Furthermore, we exploited the Tandem-genotypes-join utility to prioritize loci characterized by large changes in STR copy number in the proband compared with his parents. We focused on the top 100 prioritized loci.

GO enrichments have been performed with DAVID[50] (online platform), using as query genes all genes overlapping somatic STR events (supported by at least 1 read) in each of the 3 members of the trio and all genes overlapping STR (motif length 2-13) annotated in RepeatMasker and covered by at least 30 reads in all members of the trio as the background. We considered as significant only GO enrichments associated with an FDR < 0.05.

Identification of somatic structural variants from long-read GS data

We also used CuteSV to call structural variants (insertions, duplications, deletions, break ends, and inversions) from the alignment files obtained with Minimap2.[44] CuteSV reports SVs supported by even just a single read. To estimate the rate of somatic and germinal events in each sample, we calculated the ratio between reads supporting the reference allele (DR) and reads supporting the structural variant (DV) for all loci covered by at least 5 total reads. We considered as "somatic" all SVs with a DR/DV ratio above 9. We used Tandem-Repeat Finder[51] to estimate the percentage of genomic sequences involved in structural variants (Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50 Maxperiod = 13).

Statistical analysis

Dunn's multiple comparison test was used for multiple comparisons after the Kruskal-Wallis test. The Mann-Whitney test was used to compare 2 mean populations. Statistical analyses were performed by using GraphPad Prism 9 software (GraphPad Software). Data are expressed as mean ± SEM. P values of less than .05 were considered statistically significant. P values are indicated in all figures according to convention: *P < .05; **P < .01; ***P < .001.

To estimate the amount of short-reads mapped to tandem repeats we used ExpansionHunter Denovo, we considered an amount of normalized reads associated with a z-score > 3 (FDR < 0.01). Read counts normalization was performed with ExpansionHunter Denovo; z-scores and P values were calculated with Scipy.

Results

Summary of clinical features and DDR gene variants in NDD cohorts

To investigate the role of DDR genes in the development of regressive phenotypes in NDD individuals and their

potential association with germline and somatic STR instability, we carried out a clinical and genomic analysis of a cohort of 105 individuals with NDD.

A regressive phenotype was identified in 11 probands. Because ES data were available, we searched for variants in DDR genes, with special emphasis on those involved in the repeat expansion in HD (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, Lee JM, McLean ZL, et al. Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease reveal shared and tissue-specific effects. *bioRxiv*. 2024. https://doi.org/10.1101/2024.06.10.597797),[35] and those shown to modulate *TCF4* repeat instability.[23] We focused on patients with at least 2 variants (either de novo or inherited) located in 2 genes that were predicted to interact by STRING and ORVAL, and that could additively contribute to the patients' phenotype. We found that 6 out of the 105 analyzed patients carried at least 2 variants in DDR genes: 4 patients with a regressive phenotype (F36A, F19A, F20A, and F66A), 1 case (F84A), whose severe phenotype impaired the evaluation of regression, and one proband (F69A) who was the only patient with variants in DDR genes without evidence of regression (Table 1,[32,52] Supplemental Tables 6 and 7). Interestingly, F36A carried a likely pathogenic variant in *FAN1*.

To further investigate the contribution of DDR genes to the regressive phenotype, we analyzed a second cohort of 982 individuals with NDDs, all of whom carried nonbenign copy-number variants (CNVs) identified by CMA during routine diagnostic screening. In this cohort, we identified 9 cases with recurrent deletions on chromosome 15 encompassing *FAN1*. Of the 5 patients who could be contacted for follow-up, 1 was excluded because of the presence of additional pathogenic CNVs. Among the remaining 4, 2 individuals exhibited a regressive phenotype and were included in this study (F74A and F82A).

We also extended the analysis to a cohort of 10 trios, each consisting of a mother, father, and a proband with a nonregressive NDD. Among them, the proband (NED034_P) presented with mild intellectual disability (ID) and epilepsy and was heterozygous for a maternally inherited deletion on chromosome 15, del(15)(q13.2q13.3), encompassing *FAN1*. However, no additional variants in other DDR genes were identified in this case.

Altogether, we focused on 6 NDD patients presenting with a regressive phenotype and carrying variants in DDR genes (Figure 1A, Table 1,[32,52] Supplemental Table 6 and 7), 3 of whom harbored alterations involving *FAN1*.

The patients included 3 males (F19A, F20A, and F36A) and 3 females (F66A, F74A, and F82A), ranging from 14 to 23 years of age. All were born to nonconsanguineous parents after an uneventful pregnancy, with normal birth parameters. Developmental delay and/or behavioral problems during childhood were reported in the fathers of 2 patients (F74A and F82A), whereas the remaining cases were classified as sporadic. Their main clinical features included psychomotor delay and ID with 2 individuals (F19A and F20A) receiving a diagnosis of ASD.

All identified probands showed neuropsychiatric features since they were 1 year old, which progressively worsened over time. ID was evident in all individuals, with more pronounced language regression observed in 4 cases (F19A, F20A, F66A, and F74A). Motor regression was noted in 4 patients (F19A, F36A, F74A, and F82A), particularly in the 2 males (F19A and F36A). Three patients developed epilepsy: 1 of them during his first years of life (F36A), 1 in his late childhood (F19A), and 1 after febrile seizures that later evolved into epilepsy (F74A). Notably, F19A experienced subacute, severe psychomotor regression after the onset of epilepsy, initially raising the suspicion of a dysimmune etiology. However, this hypothesis was excluded based on negative results of the autoimmune screening (Supplemental Table 6). Behavioral dysregulation represented a significant concern in 3 patients (F19A, F36A, and F82A). None of the probands had facial dysmorphisms, and head circumference was within the normal range for all. One patient was myopic (F74A).

Brain MRI was performed in 3 individuals: 2 showed normal findings (F19A and F36A), whereas 1 exhibited a mild dysmorphism of the telencephalic commissural plate (F82A) (see Table 1,[32,52] Supplemental Table 6). Overall, F36A presents the most severe clinical phenotype among the probands in our cohort.

Detailed analysis of variants in *FAN1* and other DDR genes in 6 ASD patients with regression

We then aimed to analyze the genomic landscape of the 6 individuals with NDD and regression, all of whom carry pathogenic variants in DDR genes: F36A, F74A, F82A, F19A, F20A, and F66A.

In the case of F36A, after a variant prioritization process and an oligogenic analysis using ORVAL, we identified 3 key variants affecting DDR genes (Figure 1, Tables 1 and 2).[32,52,53] As described above, 1 is a maternally inherited nonsense variant in *FAN1* (NM_014967.5:c.2116C>T p.(Arg706Ter)). In addition, F36A inherited 2 variants from his unaffected father, a missense variant in *ERCC2* (HGNC:3434) (NM_000400.4:c.720C>G p.(Asp240Glu)) and a nonsense variant in *SFR1* (HGNC:29574) (NM_001002759.2:c.127C>T p.(Arg43Ter)). *ERCC2* encodes a subunit of the transcription/repair factor tfIIh, which functions as a DNA helicase for nucleotide excision repair. Its role consists in unwinding the double helix via 5′-3′ helicase activity. The identified missense variant affects the helicase adenosine triphosphate (ATP)-binding domain and is predicted to be pathogenic by AlphaMissense, a tool that predicts the pathogenicity of all possible single amino acid changes at a given position (Table 2).[52-54] *SFR1* encodes a component of the swi5-sfr1 complex, which is

Table 1 Clinical features of the probands analyzed

Case	Sex	Age	*FAN1* Locus	0-12 Months	1-3 Years	3-6 Years	6-10 Years	10-15 Years	From 15 Years Up
F36A	M	20	Heterozygous *FAN1* variant: NC_000015.10:g.30922298C>T NM_014967.5:c.2116C>T NP_055782.3:p.(Arg706Ter) (PVS1;PM2;PP5)(Pathogenic)[a]	Psychomotor delay	Epilepsy, psychomotor regression	Drug-resistant epilepsy, severe emotional/ behavioral dysregulation[d]	Motor regression[e] Severe ID, persistent seizures	Behavioral worsening, Persistent seizures	Movement impairment, persistent seizures
F74A	F	19	Heterozygous chromosomal deletion: seq[GRCh38] del(15)(q13.2q13.3) NC_000015.10:g.30722306_32218662del 1A (0.00); 2A (+1.00); 3A (0.00); 4L (+0.15); 4N (-0.15); 4O (0.00)(Pathogenic)[b] (including *FAN1*)	Psychomotor delay	Febrile seizures, psychomotor regression[c]	Epilepsy	Moderate ID, Language regression	Mild movement impairment	Mild movement impairment, rare seizures
F82A	F	14	Heterozygous chromosomal deletion: seq[GRCh38] del(15)(q13.2q13.3) NC_000015.10:g.30588776_32393763del 1A (0.00); 2A (+1.00); 2G (0.00); 3A (0.00); 4L (+0.15); 4N (-0.15); 4O (0.00)(Pathogenic)[b] (including *FAN1*)	Psychomotor delay	Mild psychomotor regression[c]	Poor motor coordination	Moderate ID	Mild emotional/ behavioral dysregulation	/
F19A	M	23	NA	Psychomotor delay	Loss of social and language skills	ASD	Severe ID/ emotional/ behavioral/ dysregulation	Epilepsy, Motor regression[e]	Movement impairment, persistent seizures

[a]Prediction by the Franklin Genoox platform, according to the ACMG/AMP guidelines (Richards et al,[32] 2015).
[b]Prediction by the Franklin Genoox platform, according to the ACMG/ClinGen guidelines (Riggs et al,[52] 2020).
[c]Psychomotor regression: loss of social and language skills (loss of previously acquired individual words and social-relational skills, after onset of epileptic seizures).
[d]Emotional/behavioral dysregulation: episodes of psychomotor agitation with screaming, crying, aggression, tremors, and autonomic symptoms (sweating and tachycardia) appearing "suddenly."
[e]Motor regression: awkward, uncoordinated, broad-based gait with camptocormic features and limb rigidity that worsen with age.

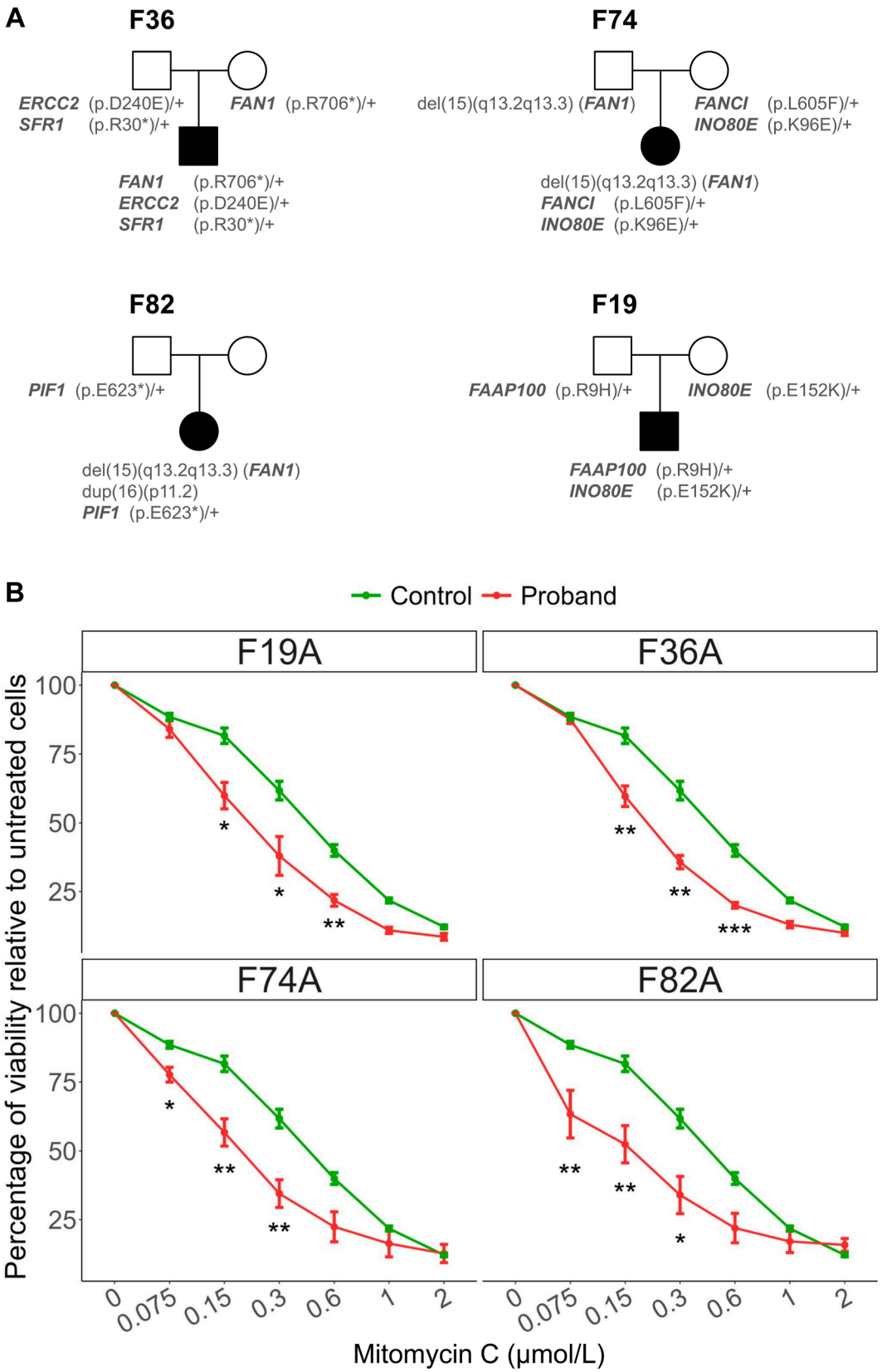

Figure 1 **Family pedigrees and mitomycin C sensitivity of proband-derived lymphoblastoid cells.** A. Pedigree of proband families. Implicated genes and their variants, indicated as amino acid changes, together with chromosomal deletions/duplications are shown; B. Sensitivity to mitomycin C of lymphoblastoid cells obtained from probands F36A, F82A, F19A, and F74A. Lymphoblasts derived from probands and a healthy individual (CTR, control) were treated for 5 days with various concentrations of MMC. For each patient, percentage of cell viability was evaluated relative to untreated cells. Points in the line represent the means ± SEM of 7 independent experiments performed in triplicate. *$P < .05$, **$P < .01$, ***$P < .001$ vs control cells.

Table 2 Potentially deleterious variants found in probands

SNVs						
Case	Gene	Variant	Inh	Variant Effects	Freq 1000g; GnomAD ALL V4.1.0	DNA Repair Pathways
F36A	*FAN1*	NC_000015.10:g.30922298C>T NM_014967.5:c.2116C>T NP_055782.3:p.(Arg706Ter)	Mat	ACMG/AMP: Likely Path (PVS1, PM2)	N/A; <1/10,000	Fanconi anemia pathway/Binds the FANCI- FANCD2 complex
	ERCC2	NC_000019.10:g.45364330G>C NM_000400.4:c.720C>G NP_000391.1:p.(Asp240Glu)	Mat	ACMG/AMP: VUS (PM2) AlphaMissense: Path	N/A; <1/10,000	Helicase ATP-binding, subunit of the transcription/repair factor TFIIH/ nucleotide excision repair pathway
	SFR1	NC_000010.11:g.104123078C>T NM_001002759.2:c.127C>T NP_001002759.1:p.(Arg43Ter)	Pat	ACMG/AMP: VUS (PM2)	N/A; <1/100,000	Component of the SWI5-SFR1 complex/ required for double-strand break repair via homologous recombination
F74A	*FANCI*	NC_000015.10:g.89285210C>T NM_018193.3:c.1813C>T NP_060663.2:p.(Leu605Phe)	Mat	ACMG/AMP: B (BS1, BS2, BP6) AlphaMissense: ambiguous Reduced protein expresssion[54]	0.002; <1/100	Fanconi anemia pathway/Binds FANCD2 to form the ID complex, central for the activation of the Fanconi pathway/directly interacts with FAN1
	INO80E	NC_000016.10:g.30000930A>G NM_173618.3:c.286A>G NP_775889.1:p.(Lys96Glu)	Mat	ACMG/AMP: VUS (PM2, PP3) AlphaMissense: Path	N/A; N/A	INO80 Complex Subunit/DNA repair through Homologous Recombination
F82A	*PIF1*	NC_000015.10:g.64816357C>A NM_001286496.2:c.1867G>T NP_001273425.1:p.(Glu623Ter)	Pat	ACMG/AMP: VUS (PM2)	N/A; <1/100,000	5' to 3' DNA helicase/break-induced replication, a specialized homologous recombination pathway for DNA double-strand break (DSB) repair
F19A	*FAAP100*	NC_000017.11:g.81552305C>T NM_025161.6:c.26G>A NP_079437.5:p.(Arg9His)	Pat	ACMG/AMP: VUS (PM2, BP4) AlphaMissense: B	N/A; <1/10,000	Fanconi anemia pathway/Component of the Fanconi anemia core complex/ required for FANCD2 monoubiquitination
	INO80E	NC_000016.10:g.30001471G>A NM_173618.3:c.454G>A NP_775889.1:p.(Glu152Lys)	Mat	ACMG/AMP: VUS (PM2, BP4) AlphaMissense: B	N/A; <1/100,000	INO80 Complex Subunit/DNA repair through Homologous Recombination
	SHANK3	NC_000022.11:g.50721463dup NM_001372044.2:c.3855dup NP_001358973.1:p.(Val1286ArgfsTer85)	dn	ACMG/AMP: Likely Path (PVS1, PM2)	N/A; N/A	-

CNVs			
Case	Chromosome Anomaly (Inheritance)	Coordinates GRCh38 (size Kbp) Variant Effects	Protein-Coding Genes
F74A	del(15)(q13.2q13.3) (pat)	NC_000015.10:g.30722306_32218662del (1.496) ACMG/ClinGen: 1A (0.00); 2A (+1.00); 3A (0.00); 4L (+0.15); 4N (-0.15); 4O (0.00); (Pathogenic)	*FAN1*[a], *MTMR10*[a], *TRPM1*[a], *KLF13*, *OTUD7A*[a], *CHRNA7*

(continued)

Table 2 Continued

Case	CNVs		
	Chromosome Anomaly (Inheritance)	Coordinates GRCh38 (size Kbp) Variant Effects	Protein-Coding Genes
F82A	del(15)(q13.2q13.3) (dn)	NC_000015.10:g.30588776_32393763del (1.858) ACMG/ClinGen: 1A (0.00); 2A (+1.00); 2G (0.00); 3A (0.00); 4L (+0.15); 4N (-0.15); 4O (0.00); (Pathogenic)	*CHRFAM7A, GOLGA8R, GOLGA8Q, GOLGA8H, ARHGAP11B, FAN1*[a], *MTMR10*[a], *TRPM1*[a], *KLF13, OTUD7A*[a], *CHRNA7*
	dup(16)(p11.2p11.2) (dn)	NC_000016.10:g:29641678_30187279dup (0.545) ACMG/ClinGen: 1A (0.00); 2A (+1.00); 2B (0.00); 2H (0.00); 2K (0.00); 2L (0.00); 3A (0.00); 4L (+0.15); 4O (0.00); (Pathogenic)	*SPN, QPRT, C16orf54, ZG16, KIF22*[a], *MAZ, PRRT2*[a], *PAGR1, MVP, CDIPT, SEZ6L2, ASPHD1, KCTD13, TMEM219, TAOK2, HIRIP3, INO80E, DOC2A, C16orf92, TLCD3B*[a], *LOC112694756, ALDOA*[a], *PPP4C, TBX6*[a], *YPEL3, GDPD3, MAPK3, CORO1A*[a]

Variant effects: *ACMG/AMP*, interpretation of sequence variants according to the American College of Medical Genetics and Genomics recommendations as calculated with the Franklin Genoox tool; *ACMG/ClinGen*, variant effects evaluated according to the ACMG/ClinGen guidelines (Riggs et al,[52] 2020); *AlphaMissense*, molecular effects of missense variants as predicted by the AlphaMissense tool which predicts the pathogenicity of all possible single amino acid changes at a given position; *B*, benign; *DNA repair pathways*, main functional role of involved genes in known DNA repair pathways; *Inh*, inheritance; *Path*, pathogenic; *Protein-coding genes*, protein-coding genes which localize in the genomic region of reported chromosome deletions/duplications.

[a]OMIM genes.

essential for double-strand break (DSB) repair through homologous recombination.

To explore the possibility of synergistic pathogenic mechanisms among these 3 inherited variants, we performed a digenic interaction analysis using ORVAL. The results indicated a likely digenic interaction between *FAN1* and *ERCC2* (score 0.833), whereas *FAN1-SFR1* and *ERCC2-SFR1* pairs classified as major/modifier genes (scores 1.00 and 0.6, respectively) (Supplemental Figure 1). Finally, interrogation of the STRING database revealed that the 3 proteins are functionally interconnected within DDR pathways via rad51, the primary recombinase in mitotic cells and a central player in homologous recombination (Supplemental Figure 2A and B).

We then focused our attention on F74A and F82A probands, the 2 patients with regression and the recurrent deletion of chromosome 15, as identified by CMA. To better characterize their genomic landscape, we performed GS. Both probands were confirmed to carry a heterozygous deletion of chromosome 15, del(15)(q13.2q13.3), encompassing an overlapping region that includes *FAN1* along 5 additional protein encoding genes: *TRPM1* (HGNC:7146), coding for a calcium channel and associated with autosomal recessive congenital stationary night blindness (OMIM 613216); *OTUD7A* (HGNC:20718), coding for a deubiquitinating enzyme and involved in a recessive neurodevelopmental disorder with hypotonia and seizures (OMIM 620790); and 3 genes not currently associated with human disease and coding for a nicotinic acetylcholine receptor (*CHRNA7;* HGNC:1960); a myotubularin related protein (*MTMR10;* HGNC:25999) and a Kruppel-like transcription factor (*KLF13*; HGNC:13672) (Table 2[52,54]). The chromosome 15 deletion was de novo in patient F82A and paternally inherited in F74A (Figure 1A). These hemideletions partially overlap with the recurrent 15q13.3, which is known to be associated with a wide spectrum of clinical manifestations and exhibits incomplete penetrance (OMIM 612001).

Both F74A and F82A probands carry 2 additional pathogenic variants in DDR genes. Patient F74A inherited 2 missense variants from her healthy mother, 1 in *FANCI* (HGNC:25568) and 1 in *INO80E* (HGNC:26905) (Figure 1A, Table 1[32,52]). *FANCI* is part of the FA Complementation Group I (OMIM 609053). This gene plays an essential role in the repair of DNA DSBs by homologous recombination, as well as in the resolution of interstrand DNA crosslinks. The identified missense variant affects a key region of the protein, the helical domain 2 (*HD2*, IPR029312). Although this variant is predicted as benign by the Franklin Genoox platform, according to the ACMG/AMP guidelines,[32,33] and is predicted as ambiguous by the AlphaMissense predictor, functional studies demonstrated that it significantly impairs DNA repair activity. Specifically, cells expressing *FANCI* with the p.Leu605Phe variant[53] exhibit severely reduced localization of *FANCD2* (HGNC:3585), a binding partner of the fanci-fancd2 (ID2) complex, to sites of DNA damage. *INO80E*

encodes subunit E of the *INO80* chromatin remodeling complex, which is involved in DNA recombination, DNA repair, and chromatin remodeling (Table 2[52,54]).

Patient F82A carries 2 additional variants of interest: a de novo duplication of chromosome 16, dup(16)(p11.2p11.2) and a paternally inherited nonsense variant in *PIF1* (HGNC:26220) (NM_001286496.2:xc.1867G>T p.(Glu623Ter)) (Figure 1). The 16p11.2 duplication is a common CNV of approximately 600 Kbp, encompassing 27 protein-coding genes, 6 of which are known disease genes: *KIF22* (HGNC:6391), associated with a spondyloepimetaphyseal dysplasia form (OMIM 603546); *PRRT2* (HGNC:30500), whose inactivating variants causes dominant forms of infantile seizures (OMIM 605751) and episodic kinesigenic dyskinesia (OMIM 128200); *TLCD3B* (HGNC:25295), linked to a recessive form of cone-rod dystrophy (OMIM 619531); *ALDOA* (HGNC:414), associated with a recessive form of aldolase A deficiency (OMIM 611881); *TBX6* (HGNC:11605), implicated in the spondylocostal dysostosis (OMIM 122600); *CORO1A* (HGNC:2252), associated with a recessive form of immunodeficiency (OMIM 615401). Although the pathogenetic mechanism of the 16p11.2 duplication in NDDs remains unclear, it is assumed that dosage imbalance of 1 or more genes within this genomic region contributes to disease susceptibility.[55] Indeed, affected individuals with the 16p11.2 duplication, often carry a second large CNV, which may increase the penetrance of the disorder.[56] Notably, none of the genes within the duplication appear to be involved in DDR pathways.

Patient F82A also harbors a nonsense variant in *PIF1* (Figure 1A), which encodes a DNA-dependent ATP-metabolizing enzyme. *PIF1* functions as a 5′ to 3′ DNA helicase, and it is essential for maintaining genome stability. The nonsense variant affects the last exon of *PIF1* and is predicted by Ensembl Variant Effect Predictor to escape nonsense-mediated decay, potentially producing a truncated protein lacking the last 20 amino acids of the region (aa 167-641, UniProt) deputed to hydrolyze ATP. In vitro experiments have confirmed that *PIF1* exhibits ATP-dependent helicase activity, as well as 5′→3′ translocase activity, enabling it to unwind duplex DNA structures, G quadruplexes, and RNA-DNA hybrids (reviewed in Meir and Greene[57]). *PIF1* plays multiple roles in DNA replication, including break-induced replication, a specialized homologous recombination pathway for DNA DSB repair.[57] Finally, protein-protein interactions analysis using STRING revealed a possible interaction in DNA repair pathways between *FAN1* and *PIF1* mediated through *RAD51* (HGNC:9817) (Supplemental Figure 2C and D).

Patient F19A carries a paternal missense variant in *FAAP100* (HGNC:26171) (NM_025161.6:c.26G>A p.(Arg9His)) and a maternal missense variant in *INO80E* (NM_173618.3:c.454G>A p.(Glu152Lys)) (Figure 1A). *FAAP100* is a component of the FA core complex, which plays a key role in the DNA damage response network associated with the FA pathway. When *FAAP100* is repressed, it induces chromosomal instability and hypersensitivity to DNA cross-linking agents.[58] Faap100 is also required for the monoubiquitination of fancd2, playing a role similar to that of fan1 in fancd2-fanci monoubiquitination and activation of the FA pathway. The missense variant affects the nuclear pore localization protein npl4c (IPR007717). The arginine residue at position 9 is highly conserved (conservation = 0.9 from 52 aligned protein sequences), indicating that it is likely crucial for the protein's function (VarSite "disease propensity" value: 1.45).

Proband F19A also carries a de novo frameshift variant inactivating *SHANK3* (HGNC:14294) (NM_001372044.2:c.3855dupG p.(Val1211ArgfsTer85)). Shank3 is a multidomain scaffold protein of the postsynaptic density that connects neurotransmitter receptors, ion channels, and other membrane proteins to the actin cytoskeleton and G-protein-coupled signaling pathways. Variants in this gene are causative of ASD, and are implicated in the 22q13.3 deletion syndrome (Phelan-McDermid syndrome, OMIM 606232).[59] Although the shank3 product is not directly linked to DNA repair, it may be involved in some aspects of this patient's phenotype.

Patient F20A has also a variant in a gene of the FA pathway (*FANCA*; HGNC:3582), highlighting the potential contribution of this pathway to the pathogenetic mechanisms underlying these patients' conditions. Additionally, F20A carries a missense variant in *EXO1* (HGNC:3511), a gene involved in DNA mismatch repair.

Proband F66A has 2 variants, each in genes implicated in DDR through homologous recombination; *MMS19* (HGNC:13824) and *INO80E*, where *INO80E* is also implicated in F74A and F19A probands.

Interestingly, among the 95 probands without regression, only 1 patient carried variants in DDR genes (F69A). This patient has heterozygous variants in 2 genes of the DNA mismatch repair (MMR) pathway: *MSH4* (HGNC:7327) and *PMS2* (HGNC:9122). Specifically, this patient is heterozygous for a frameshift variant of *PMS2*, a gene thought to promote repeat expansions. Expanded repeat sequences can form secondary structures that might lead the mismatch repair complex to act erroneously, potentially inducing somatic expansions. Accordingly, MMR gene silencing is being considered as a potential therapeutic approach (reviewed in Pengo and Squitieri[60]).

FAN1 expression is reduced in probands F36A, F74A, and F82A

To better characterize this selected group of individuals with variants in DDR genes who also exhibited a regressive phenotype, we investigated whether the variant in 1 of the *FAN1* alleles led to a decrease in protein levels in F36A, F74A, and F82A.

Proband F36A carries a heterozygous nonsense variant in *FAN1*, whereas probands F74A and F82A both carry a heterozygous deletion of chromosome 15 encompassing *FAN1*. In Supplemental Figure 3, immunoblotting of lymphoblasts from probands F36A and F74A showed that *FAN1* expression was reduced, as expected, to about half of the normal dosage, thus excluding compensatory mechanisms. Cell lysates of HEK293 cells and HEK293 cells in which *FAN1* was silenced by siRNA were used to validate the specificity of the anti-fan1 antibody. As expected from genome data, *FAN1* expression was reduced in the lymphoblasts of F36A and his mother (F36M) but not in those of his father (F36F). Similarly, *FAN1* expression was reduced in the lymphoblasts of proband F74A and his father but not in those of her mother (F74M), who does not carry the chromosome 15 deletion (Supplemental Figure 3C and D).

Lymphoblastoid cells of probands F36A, F19A, F74A, and F82A show increased sensitivity to MMC-mediated cytotoxicity

Alteration of genes responsible for maintaining genome stability may result in increased sensitivity of cells to the cytotoxic effects of InterCrossLink-inducing agents, such as MMC.[61] MMC treatment causes the cross-linking of double-stranded DNA and results in inhibition of DNA replication and ultimately in a decreased cell viability.[19] Thus, to further investigate the functional impact of DDR gene defects, we assessed the sensitivity of probands' cells to the MMC cytotoxicity. Briefly, lymphoblastoid cells from probands F36A, F19A, F74A, F82A, and a control individual were obtained. An equal number of cells for each individual was seeded and treated for 5 days with increasing concentration of MMC (0.075 µmol/L to 2 µmol/L). Then, cell cultures were arrested, and cell viability was assessed. As shown in Figure 1B, at 0.15 µmol/L of MMC, lymphoblasts from probands F36A, F82A, F74A, and F19A exhibited a statistically significant decrease in cell viability, with a reduction approaching 50% compared with the control cells, which maintained over 75% viability. At 0.30 µmol/L of MMC, cell viability of lymphoblasts from those probands dropped to about 30%, approximately half that of the control cells. The F74A and F82A probands showed a statistically significant decrease in cell viability at the lowest MMC dosage tested (0.075 µmol/L).

In summary, the increased sensitivity of the lymphoblasts of probands F36A, F19A, F74A, and F82A to MMC-induced cytotoxicity suggests a common functional impact of the combination of genetic variants in DDR genes carried by these 4 probands. We then examined whether the MMC sensitivity observed in the probands was also increased compared with their parents. Lymphoblastoid cells from the parents of probands F36A and F82A were obtained and MMC sensitivity test was repeated including cells from the 2 trios and from a control individual (Supplemental Figure 4A and B). The cell viability of the 2 probands was reduced compared with both the control individual and their parents.

Proband F36A carries a nonsense variant in *FAN1*, shared with his mother, and a nonsense variant in *SFR1* and a missense in *ERCC2* inherited from his father. These findings suggest that the increased susceptibility to MMC-mediated cytotoxicity in F36A is due to the combined effect of variants in 3 DDR genes and that the loss of a single allele of *FAN1* is not sufficient to confer an increased susceptibility to MMC cytotoxicity. A similar effect may be present in proband F82A, who carries a nonsense variant in *PIF1*, shared with his mother, and an additional de novo deletion of chromosome 15 containing *FAN1*.

Proband F36A shows an increased burden of somatic STR expansions

We hypothesized that the reduction in DNA repair capability in the 4 probands could lead to an accumulation of somatic genomic variants. Three of the 4 probands are heterozygous for *FAN1* loss-of-function variants (Figure 1A, Table 2[52,54]). Given that *FAN1* is thought to play a protective role against STR expansions, we exploited GS short-read sequencing data to quantify somatic variants related to STRs.

To this end, we analyzed our samples using the ExpansionHunter DeNovo tool. We calculated the total number of IRR pairs for each individual in a cohort consisting of the 12 members of our 4 trios and 132 healthy controls, all sequenced with the same technology. We did not observe significant differences in the amount of total IRR between any members of the F19, F74, and F82 families and the control group. However, the proband F36A exhibited a significantly higher total IRR compared with all other samples analyzed (Figure 2A, Supplemental Table 8). Although a few control samples showed high total IRR values, F36A was the only sample with a total IRR significantly higher than all others (z-score = 3.9, FDR < 0.01; Figure 2A, Supplemental Table 8).

Next, we sought to identify which repeat motifs were most enriched in F36A compared with the rest of the cohort. We performed an "outlier analysis" with ExpansionHunter DeNovo on our cohort following the tools recommendations. On a total of 6619 repeat motifs identified, 13 showed the highest enrichment z-score in the proband F36A (considering only z-scores above 1). Only one repeat motif ("*AATATATATATTATATAT*") had a z-score above 3 (Z = 3.63; Figure 2B, Supplemental Tables 9 and 10).

Taken together these results suggest that proband F36A presents an increased burden of total, and especially AT-rich, STR expansions compared with his parents and the cohort of healthy controls. Comparing patients and their parents is especially interesting given their shared genetic background. We acknowledge, however, that the parents' greater age could increase their somatic mutational burden,

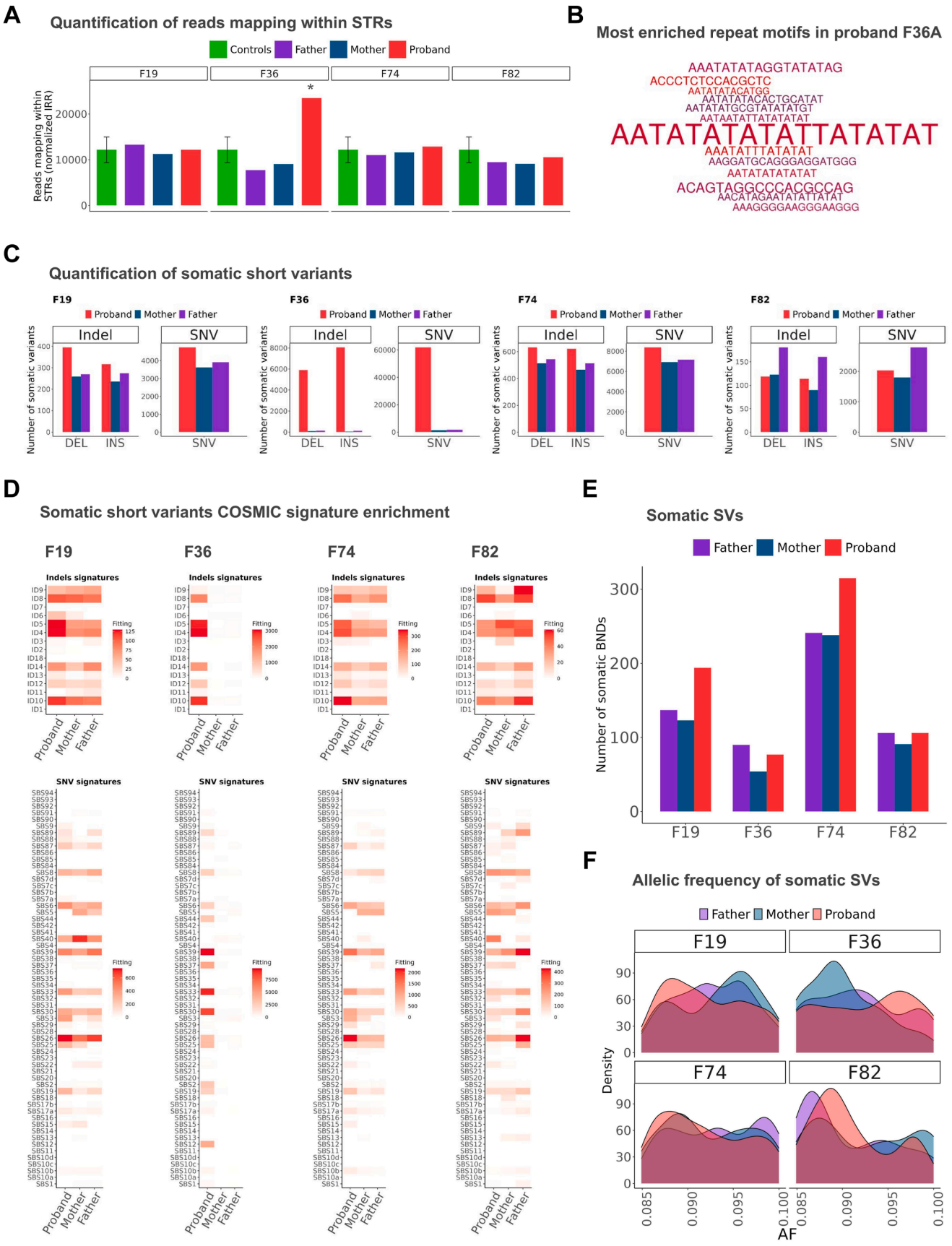

Figure 2 **GS short-reads analyses.** A. Normalized amount of in-repeat read pairs quantified in each individual of each analyzed trio, *z-score > 3, P value < .01. B. Word cloud representing the most enriched STR motifs in the proband F36A compared with all other samples of the analyzed cohort, word sizes are proportional to the z-score calculated by Expansion Hunter DeNovo. C. Total number of somatic SNVs and Indels identified in the analyzed trios. D. COSIMC signature enrichment of each analyzed trio; E. Number of somatic break-ends (BND) detected with svABA. F. Distribution of the allelic frequencies (AF) associated to somatic BNDs detected with svABA.

potentially masking a significant rise in their offspring's variants. Consequently, the proband exhibiting the greatest difference in somatic variant quantity holds significant importance.

Specific probands with regressive phenotypes show increased somatic mutational burden

Despite all 4 probands showing increased sensitivity to MCC, only 1 individual appears to carry an increased burden of STR expansions. To further investigate the presence of somatic variants in these individuals, we used Mutect2 to detect somatic short variants (specifically single-nucleotide variants, SNVs, and small insertions and deletions, Indels) within annotated genomic repeats. We developed a method to quantify the burden of short somatic variants in the members of our trios, as described in the Materials and Methods section. To minimize sequencing depth biases, we randomly downsampled our alignment files to ensure the same average sequencing coverage for all members of each trio (Supplemental Tables 2 and 3).

As expected, F36A exhibited a significantly higher amount of all identified classes of somatic variants compared with his parents (Figure 2C). Importantly, an overall increase was also shown in probands F19A and F74A (Figure 2C). Specifically, F19A displayed a marked increase in small deletions and SNVs compared with his parents, whereas F74A exhibited an increase in small deletions, insertions, and SNVs relative to her parents (Figure 2C). In contrast, proband F82A showed a comparable number of somatic short variants to her mother's and fewer somatic variants than her father's (Figure 2C).

We then plotted the allelic frequency (AF), defined as the ratio between the number of reads supporting the alternative and the reference allele, for each member of each analyzed trio. AF measures the percentage of cells carrying each somatic variant. Although we did not observe significant differences in AF distribution among members of trios F19, F74, and F82, somatic variants identified in the proband F36A showed an overall higher AF compared with his parents and to all members of the other trios (Supplemental Figure 5A). This result suggests that the increased rate of somatic short variants in F36A might have originated at an early developmental stage.

We also assessed the length distribution of somatic indels identified in all analyzed individuals. Over 99% of somatic indels were shorter than 10 bp. Interestingly, a higher proportion of 3-bp-long small insertions and deletions was observed, particularly in proband F36A and, to a lesser but significant extent, in proband F74A (Supplemental Figure 5B). This suggests that the mechanisms at the basis of the increased mutational burden in these individuals may favor the formation of 3-bp-long short variants.

Next, we investigated whether the somatic mutational burden in these 4 individuals showed enrichment for specific mutational signatures. Using the R package mutSigExtractor, we extracted the 96 trinucleotide complexes from the somatic short variants and fitted them to Pan-Cancer Analysis of Whole Genomes signature profiles to compute signature enrichments. Notably, probands F36A and F74A showed strong enrichment for the SBS26 signature, associated with defective DNA mismatch repair, and the ID8 signature, involved in NHEJ DNA repair. Furthermore, the proband F36A exhibited an enrichment for the SBS30 signature, associated with deficiency in base excision repair (Figure 2D). Additionally, probands F19A, F36A, and F74A showed enrichment of signatures ID4, ID5, and ID10, which had not yet been linked to specific molecular effects; however, ID10 is associated with a high percentage of large insertions at repeats (>5 bp). These results suggest that abnormalities in DDR pathways may lead to somatic variant accumulation in specific probands characterized by a regressive neurodevelopmental phenotype.

To determine whether our findings are common among probands with neurodevelopmental conditions, we performed the same short somatic variant burden analysis on a cohort of 10 trios, each consisting of a mother, father, and a proband affected by a nonregressive neurodevelopmental phenotype. These samples had been randomly downsampled to achieve uniform coverage for each trio member (Supplemental Table 3). We also repeated the short somatic variant burden analysis on the previously analyzed 4 trios after further downsampling to avoid coverage bias. Overall, all trios had an average coverage between 39X and 47X (Supplemental Table 3). To avoid sex bias, we calculated the average ratio of somatic variants between the proband and the mother, as well as between the proband and the father, separately for trios with male and female probands. Proband F74A exhibits the highest proband/parents ratio of somatic SNVs and small deletions among all analyzed trios with female probands (Supplemental Figure 5C), whereas probands F19A and F36A demonstrated the highest ratio across all classes of somatic variants identified compared with the trios with male probands (Supplemental Figure 5D).

Finally, we used the tool svABA[42] to estimate the number of somatic BNDs within annotated genomic repeats as described in the Materials and Methods section. We observed an increased number of BNDs in probands F19A and F74A compared with their parents (Figure 2E). Although proband F36A did not show a markedly increased number of BNDs, somatic BNDs detected in this proband exhibited a generally higher AF compared with his parents and all other trio members (Figure 2F). This finding suggests that BNDs may have originated during an early developmental stage.

Taken together, these results indicate that an increased burden of somatic variation may be exclusive to a subset of individuals with a regressive neurodevelopmental phenotype.

Long-read GS data reveals an increased STR expansions burden in proband F36A

Short-reads GS data revealed an increased burden of somatic small variants within genomic repeats in specific probands affected by a regressive neurodevelopmental phenotype. In particular, proband F36A showed an overwhelmingly higher number of variants compared with his parents. Notably, F36A presents the most severe phenotype among the 4 probands in our cohort.

Therefore, we focused our attention on the F36 trio, generating high-coverage, long-read Oxford Nanopore Technologies GS data for all 3 members.

To estimate the STR expansion burden in the individuals, we utilized Straglr, a tool that enables de novo genome-wide detection of STR expansions without requiring prior knowledge of known STR loci. We normalized the total number of reads supporting STR expansions (minimum size = 10 bp, read_status = full) to the total number of mapped reads for each individual. STR loci were then categorized into 2 groups based on allelic frequency: STR expansions supported by more than 10% of mapped reads at the locus were classified as germline, whereas all others were considered somatic. Over 99% of STR loci were classified as germline (Supplemental Table 11). The proband exhibited an increased normalized number of reads supporting both germline and somatic STR expansions compared with his parents and an unrelated control (Figure 3A). Interestingly, although the number of reads supporting germline STR expansions was similar between parents, the *FAN1* loss-of-function (LoF) heterozygous mother had slightly more somatic STR-supporting reads than the father (Figure 3A).

We next intersected the genomic coordinates of loci affected by germline and somatic STR expansions with repeat loci reported as expanded to ≥45 repeat units in ≥5 UKB participants, as recently described (Hujoel et al[23]). We identified 6, 8, and 4 germinal STR expansions in the proband, mother, and father of family F36, respectively (Supplemental Table 5). Overall, we detected STR expansions at 10 out of the 15 loci reported in Hujoel et al[23] supporting the idea that these are mutational hotspots.

Long-read sequencing can detect STR expansions up to several kilobases in length. We thus evaluated the length distribution of both germline and somatic STR expansions in our samples. Germline expansions were, on average, longer than somatic ones: ~25% of germline expansions exceed 100 bp, compared with just 4-7% of somatic expansions (Figure 3B). Notably, the average length of somatic expansions was slightly shorter in the mother and the proband than in the father (Figure 3B).

It has been demonstrated that STR expansions in regulatory regions can alter the expression of nearby genes and those in coding regions can disrupt protein function. To estimate the deleterious potential of the identified STR expansions, we calculated the proportion of germline and somatic STR expansions overlapping exons, introns, and regulatory regions. Exonic and intronic coordinates were extracted from the Gencode gtf file (version GRCh38); whereas genomic coordinates of regulatory regions were retrieved from Ensembl (Human Regulatory Features, GRCh38). We observed an overall increase in the proportion of somatic STR expansions overlapping functional genomic regions compared with germline: exonic overlaps increased approximately ~50-fold (from ~0.015% to ~0.8%), intronic ~30-fold (from ~0.4% to ~13.5%), and regulatory regions ~35-fold (from ~0.2% to ~7%) (Figure 3C and D, Supplemental Table 11). We also assessed the overlap with known STR loci from RepeatMasker (718,028 simple repeats, motif length 1 to ~13 bp). We noted a ~5-fold decrease in the frequency of somatic STR expansions compared with germline expansions within annotated STR loci (Supplemental Table 11). These findings suggest that although somatic STR expansions are fewer in number, they may disproportionately affect gene expression and function.

We next applied Tandem-genotypes, an algorithm that estimates STR copy number per read mapped to known STR loci (RepeatMasker, 5,683,690 tandem repeats with unit length 1 to ~13 bp) and uses RefGene (GRCh38) for gene annotation. We calculated the total number of normalized "outlier reads" (reads potentially supporting a somatic STR expansion) for each trio member as described in Materials and Methods. We identified 543,380 loci with >30 reads in all 3 individuals and normally distributed STR copy-number values (Shapiro test, $P < .05$) (Supplemental Table 12). The proband showed an overall higher number of normalized "outlier reads" compared with both parents (Figure 3E). The *FAN1* LoF heterozygous mother had an intermediate amount of somatic STR expansions, between the proband and the father (Figure 3E).

To evaluate the functional relevance of these events, we performed GO enrichment analysis using DAVID on the genes overlapping somatic STR expansions (Supplemental Table 13). In total, somatic STR expansions overlapped 509, 302, and 131 genes in individuals F36A, F36M, and F36P, respectively (based on conversion from Gene_Name to Ensembl_ID). As a background, we used all genes overlapping STR loci annotated in RepeatMasker (motif length 2-13 bp) and covered by at least 30 reads in all members of the trio (n = 21,020). Importantly, in F36A, genes overlapping somatic STR expansions were enriched for pathways relevant to neurodevelopmental processes, including "postsynaptic cell membrane," "calcium channel," and "axon guidance" (Figure 3F and G, Supplemental Table 13). GO analysis revealed enrichment in metabolism-related pathways in F36P and no enrichment for F36M (Supplemental Table 13).

It is important to note that most of the somatic STR expansion events detected in this analysis are associated with loci characterized by a large number of reads supporting the reference allele (STR copy number = 0), and

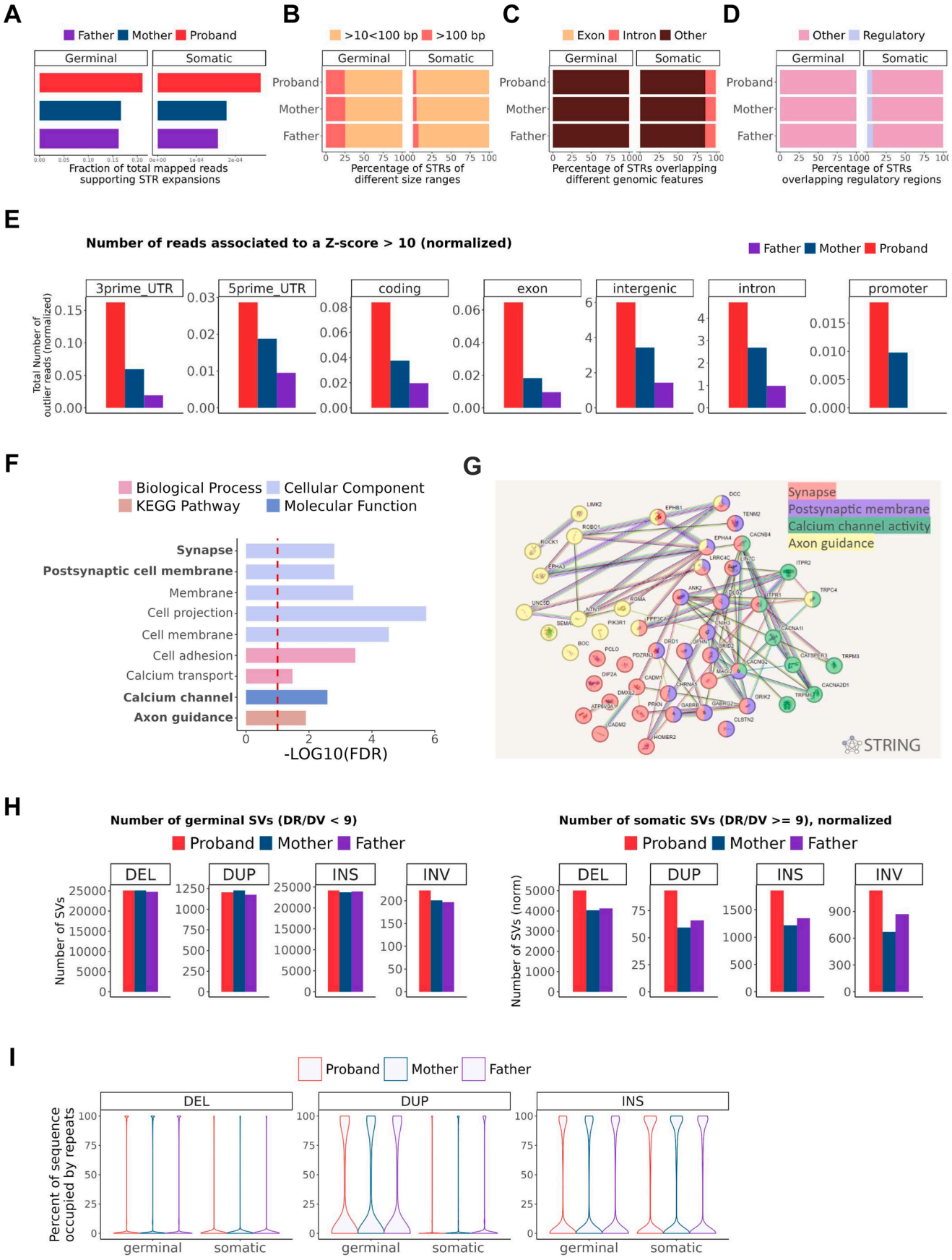

Figure 3 **Long-read sequencing of family F36.** A. Normalized amount of in-repeat read pairs quantified in each individual of each analyzed trio with Straglr, *z-score > 3, P value < .01. B. Percent of long and short germinal and somatic STR expansions identified with Straglr in the members of family F36. C. Percentage of intronic and exonic germinal and somatic STR expansions identified with Straglr. D.

only a single read supporting an additional repeat unit (STR copy number = 1) (Supplemental Table 12). Therefore, the majority of identified loci do not represent a large number of haplotypes for a single somatic STR expansion. Indeed, even with quality score analysis, a single observed STR unit cannot be excluded as an artifact. Yet, because this phenomenon affects all samples equally, the false-positive rate is expected to be consistent across them. Despite this uniform background noise, the proband consistently emerges as distinct in all analyses.

We also applied our custom pipeline for detecting somatic STR expansions to 2 additional data set: (1) 59 known pathogenic STR loci and (2) 67 loci previously found to be expanded in individuals with ASD but not in healthy controls, as described in Materials and Methods. However, we did not detect any read supporting somatic STR expansion at these loci.

Taken together these results suggest that the proband exhibits an elevated rate of somatic STR expansions compared with his parents. Furthermore, these expansions may functionally affect genes involved in neurodevelopment.

Long-read GS data reveal an increased number of somatic duplications and insertions in proband F36A

To estimate the number of somatic and germline SVs in the proband F36A, we ran CuteSV on the alignment files of the trio. For each SV, including deletions, insertions, duplications, and inversions, we calculated the ratio between the number of reads supporting the reference allele (DR) and those supporting the SV (DV) and plotted the distribution of these ratios for each individual to assess the rate of somatic SVs (Supplemental Figure 6). Interestingly, the proband showed a higher number of insertions and duplications with elevated DR/DV ratios (~10), which may indicate somatic events (Supplemental Table 14). We classified SVs with a DR/SV ratio greater than 9 as somatic and those with a ratio below 2 as germline. Notably, although we did not observe substantial differences in the number of germline SVs among trio members, the proband showed a higher number of somatic SVs, especially INS and DUP, compared with his parents (Figure 3H).

Next, we then used Tandem-Repeat Finder[51] to determine whether the genomic sequences implicated in the detected SVs are composed of simple repeats. We calculated the proportion of deleted, duplicated, and inserted sequences consisting of simple repeats (1-13 bp motifs). Germinal deletions and duplications had a higher repeat content than their somatic counterparts. Conversely, somatic insertions exhibited a higher proportion of simple repeats than germinal insertions (Figure 3I, Supplemental Table 15). We defined sequences with more than 75% simple repeats as "repeat-rich." Deleted sequences had the lowest fraction of repeat-rich content (~4% for somatic and ~20% for germinal), whereas insertions had the highest (~35% for somatic and ~26% for germinal) (Supplemental Table 15). These findings suggest that a significant portion of somatic insertions may represent repeat expansions.

Taken together, these results further support an elevated burden of somatic structural variation in the proband compared with his parents.

Locus-specific somatic STR expansions in proband F36A

We next sought to identify specific genomic loci within genes linked to NDDs for which the proband shows an increased number of reads supporting somatic STR expansions relative to his parents.

To this end, we focused on 2 sets of loci: (1) a group of 59 known pathogenic STR loci, as reported by Ibañez et al[48] (2022), Mousavi et al[49] (2019), Trost et al[15] (2020), and Mitsuhashi et al[47] (2019) (Supplemental Table 16) and (2) a set of 57 loci corresponding to STR regions previously found to be expanded exclusively in ASD patients but not in healthy controls, as identified by Trost et al[15] (2020) (Supplemental Table 17). However, manual inspection of the plots representing the amount of STR copy number at each locus, found no evident increase in-repeat number in proband compared with his parents.

We therefore adopted an alternative strategy to identify genomic loci that might contribute to the proband's phenotype. Using Tandem-genotypes, we had previously quantified STR copy number across 5,683,690 annotated loci encompassing tandem repeats with unit lengths ranging from 1 to 13 bp. Leveraging the built-in Tandem-genotypes-join function, we ranked all STR loci according to changes in STR copy numbers between the proband and his parents. From this analysis, we selected the top 100 loci with the strongest proband-parent differences and examined them for overlap with NDD-relevant genes (Supplemental Table 18). Among them, 65 overlap with coding exons, potentially altering the structure or function of the encoded

Percentage of Somatic and germinal STR expansions identified with Straglr overlapping regulatory features. E. Normalized number of reads (identified with Tandem-genotypes) supporting STR expansions characterized by a significantly higher copy number (outlier reads, z-score > 10) compared with all other reads detected within each genomic locus annotated in RepeatMasker. F. Gene Ontology enrichments relative to the genes associated to outlier reads identified in the proband with Tandem-genotypes. G. Gene network relative to the genes associated to outlier reads identified in the proband with Tandem-genotypes. H. normalized number of germinal and somatic SVs identified with CuteSV. I. Percent of the genomic sequences relative to germinal and somatic SVs identified with CuteSV occupied by simple repeats (calculated with TRF).

proteins. Fourteen were classified as SFARI genes and 15 were OMIM disease genes associated with complex phenotypes, including neurodevelopmental features. These genes were enriched for GO biological processes potentially relevant for NDDs (Supplemental Figure 7).

Manual evaluation of these top-ranked loci revealed intronic/UTR STR expansions in genes such as *MORN1* (HGNC:25852), *FHIT* (HGNC:3701), *AGMO* (HGNC:33784), and *GALNT17* (HGNC:16347), as well as smaller exonic STR expansions in *FOXL3* (HGNC:54201), *CUL4B* (HGNC:2555), and *SHANK2* (HGNC:14295) (Supplemental Table 18). We hypothesize that these genes may represent examples of NDD-relevant loci potentially contributing to the probands phenotype. One locus of particular interest is located within exon 25 of *SHANK2*, a gene involved in the organization of the postsynaptic density. LoF variants in *SHANK2* have been previously implicated in ASD.[62] Tandem-genotypes result showed a significantly increased number of long-reads supporting additional copies of the repeat unit "TCG" within exon 25 of *SHANK2* (Supplemental Figure 8A, Supplemental Table 18). However, closer examination of these reads revealed a more complex mutational landscape, involving both small deletions and insertions within a 120 bp interval of *SHANK2* exon 25 (Supplemental Figure 8B). Despite this discrepancy, our findings suggest a potential increase in somatic genomic instability within this exon in proband F36A compared with his parents (Supplemental Figure 8A, Supplemental Table 18). Importantly, such instability may potentially lead to an elevated number of somatic frameshift variants in *SHANK2* potentially influencing the proband's phenotype.

In addition, we detected 2 long reads indicating a potential intronic STR expansion within *FHIT* (Supplemental Figure 8C, Supplemental Table 18), a gene previously associated with ASD, schizophrenia,[63] and depressive phenotypes.[64]

Taken together, our data support the hypothesis that deficiencies in DDR genes could lead to the accumulation of somatic variants that may affect NDD-relevant genes.

Discussion

In this study, we advance the understanding of the molecular basis of NDDs with regression by providing 2 lines of preliminary experimental evidence. First, probands were found to carry polygenic pathogenic variants in DDR genes, many of which are components of the FA pathway, although single variants can also be present in individuals without regression. These combinations impaired DDR efficiency. Second, probands with severe clinical phenotypes and regression exhibit an exceptional burden of somatic variants and STR instability, which are enriched in genes implicated in neurodevelopment.

Developmental regression is defined as the pathological loss of previously acquired neurodevelopmental skills, such as language, motor abilities, or social engagement, after a period of apparently typical development. It represents a clinically significant feature observed in a subset of individuals with NDDs and is often associated with more severe phenotypic outcomes. Although the molecular mechanisms underlying regression remain largely unknown, recent studies have identified ultrarare variants in DDR genes in individuals with NDDs and/or developmental regression. In particular, 7 probands were found to carry pathogenic variants in 5 FA pathway genes (*FANCE* HGNC:3586, *FANCP* HGNC:23845, *FANCA* HGNC:3582, *FANCI* HGNC:25568, *FANCC* HGNC:3584) highlighting a potential role for this DDR pathway in the pathophysiology of regression.[25]

In this study, we identified a total of 6 probands with NDDs and developmental regression who carried combinations of variants in DDR genes. To this end, we first analyzed a cohort of 105 individuals with NDD, of whom 11 were clinically characterized by a regressive phenotype. Through ES followed by integrative analysis using STRING and ORVAL, we identified 6 probands harboring combinations of variants in DDR genes. Importantly, 4 of these 6 were among the 11 individuals with regression, 1 displayed a severe phenotype that precluded assessment of regression, and 1 proband (F69A) exhibited an NDD without evidence of regression. We subsequently analyzed a second, larger cohort of 982 individuals with NDD, selected based on the presence of nonbenign CNVs identified by CMA during routine diagnostics. In this cohort, we identified 9 cases with recurrent deletion on chromosome 15 encompassing *FAN1*. Among the 4 patients for whom follow-up data were available, 2 presented with a regressive phenotype.

Sequencing data from the 6 individuals with NDD and regression revealed that all carry digenic or polygenic combinations of pathogenic variants in DDR genes, including *FAN1*, *ERCC2*, *SFR1*, *FANCI*, *INO80E*, and *FAAP100*. Notably, 5 patients present variants in genes encoding for proteins that are directly or indirectly involved in the FA pathway. The FA pathway is a critical DDR mechanism responsible for the recognition and repair of DNA interstrand crosslinks, which are highly toxic lesions that impede replication and transcription. The core FA complex, composed of at least 8 proteins, initiates the repair process by monoubiquitinating fancd2 and fanci in response to DNA damage. Once ubiquitinated, the fancd2-fanci complex is recruited to sites of DNA damage, where it binds to the nuclease fan1, which cleaves the crosslinked DNA strands and facilitates repair.

Probands F36A, F74A, and F82A all exhibit loss-of-function heterozygosity in *FAN1*. In F36A, the *FAN1* variant is inherited with a missense variant in *ERCC2* and a nonsense variant in *SFR1*, both genes involved in DDR.

F74A and F82A carry a heterozygous deletion of chromosome 15, which includes an overlapping region that encompasses *FAN1* along with additional protein-coding genes. Notably, F74A inherited 2 missense variants from her healthy mother, 1 in *FANCI* and 1 in *INO80E*. *FANCI* variants have previously been associated with NDDs and regression.[25]

In fact, cells expressing the p.(Leu605Phe) variant[53] of *FANCI* show severely reduced localization of the fanci-fancd2 complex at sites of DNA damage. In turn, this mislocalization affects *FAN1* ability to carry out its DDR function.

F19A carries a paternal missense variant in *FAAP100* and a maternal missense variant in *INO80E*. Faap100 is essential for the monoubiquitination of fancd2, playing a critical role in the monoubiquitination of fancd2-fanci and activation of the FA pathway.

F20A carries a pathogenic variant in *FANCA*, another critical gene in the FA pathway that has previously been implicated in individuals with NDDs regression.[25]

Interestingly, probands F66A, F74A, and F19A present a pathogenic variant in *INO80E*. *INO80E* encodes a subunit of the INO80 chromatin remodeling complex, which plays a key role in regulating DNA accessibility by repositioning or evicting nucleosomes. This regulation is crucial for controlling gene expression, DNA repair, and replication. CNVs involving *INO80E* have been linked to NDDs, including ASD, schizophrenia, and developmental delay. Further studies are needed to fully elucidate the role of *INO80E* in NDDs with regression.

Importantly, we observed a significant decrease in cell survival rates in proband-derived lymphoblastoid cell lines compared with those from healthy individuals upon treatment with MMC. This result suggests a decreased capability of proband-derived cells to repair MMC-induced DNA damages, supporting the harmful impact of the identified variants. Increased sensitivity to MMC, indicative of genome instability, may be even useful in uncovering the failure of DNA repair pathways underlying neurological diseases, such as Huntington disease.[65]

Indeed, increased sensitivity to MMC was also observed in cells of an NDD patient with a complex phenotype proposed to originate from a synergistic contribution between DDR variants affecting both the FA/BRCA pathway and the HDAC family.[66]

Equally informative were individuals with NDDs who presented variants in DDR genes but without regression. By analyzing a cohort of 10 trios with nonregressive NDDs, we identified a proband (NED034_P) with mild ID and epilepsy, who was heterozygous for a deletion on chromosome 15, del(15)(q13.2q13.3), encompassing *FAN1*. GS did not reveal any detectable variants in other DDR genes, suggesting that cooperation between at least 2 DDR genes may be required for regression to occur.

Hemideletion of the region encompassing *FAN1* alone is in fact insufficient to cause regression, as evidenced by 2 additional individuals carrying this variant, identified by CMA, who did not present the regression phenotype. Moreover, although the chromosome 15 deletion was de novo in patient F82A, in patients F36A and F74A, the variant was maternally and paternally inherited, respectively. Both parents carrying the single variant did not show regression.

The need for polygenic/digenic variants is further supported by the comparison of susceptibility to MMC treatments in F36A-derived lymphoblastoid cell lines. F36A proband, heterozygote for polygenic variants, shows an increased susceptibility to MMC-mediated cytotoxicity when compared with F36M and F36F.

A similar effect may be present in proband F82A, who carries a de novo deletion of chromosome 15 containing *FAN1* and a nonsense variant in *PIF1*.

Interestingly, among the 95 probands without regression, we identified an individual (F69A) carrying a combination of variants in 2 DDR genes, specifically in *MSH4* and *PMS2*. The proband is heterozygous for a frameshift variant of *PMS2*, which is especially interesting in the context of PMS2's role as a promoter of repeat expansions in HD. However, recent evidence suggests that *PMS2* can act both as a promutagenic and antimutagenic factor in the repeat expansion diseases by either promoting or suppressing repeats number (Jimenez DA, Miller CJ, Walker A, et al. Reconciling the effects of PMS2 in different repeat expansion disease models supports a common therapeutic strategy. *bioRxiv*. 2025. https://doi.org/10.1101/2024.08.13.607839). Moreover, pms2 and msh4 are crucial proteins in the MMR system, with a mechanism of action distinct from that of fan1 and the Fanconi anemia (FA) pathway.

Cumulative analysis of STR expansions using srGS data with ExpansionHunter *DeNovo* revealed a significant increase in the normalized number of reads mapping within STRs exclusively in proband F36. Conversely, Mutect2 analysis showed an elevated number of somatic SNVs and indels within genomic repeats in probands F19 and F74, with proband F36 exhibiting a greatly higher number of SNVs and indels within these regions compared with their parents. Notably, in our cohort, proband F36A exhibited the most severe phenotype. These findings support the hypothesis that DDR genes dysfunction may lead to an accumulation of somatic variants, which could contribute to the severity of NDDs. Furthermore, the observation that somatic variants are more prevalent in probands with a regressive phenotype compared with those with nonregressive forms of NDDs suggests that somatic mutational burden might be a distinguishing feature of the regressive subtype. Recent literature has highlighted the significance of STR expansions in ASD[15,16]; however, the contribution of increased somatic STR burden in the regressive NDD phenotype is still poorly explored.

We thus performed high-coverage long-read sequencing on the 3 members of the family with the proband showing the most severe phenotype and carrying deleterious variants within genes *FAN1*, *SFR1*, and *ERCC2*. We observed an

increase in both germinal and somatic STR expansions in the proband compared with his parents. Interestingly, somatic expansions were more likely to be intragenic and to affect regulatory regions compared with germinal ones. This observation suggests that somatic STR expansions could have a more pronounced functional impact on gene expression compared with their germline counterparts. The association of somatic STR expansions with genes involved in postsynaptic functions, calcium channels, and axon guidance further implicates their role in neurodevelopmental phenotypes. Overall, the increased burden of STRs may alter gene expression and function, contributing to the observed neurodevelopmental abnormalities.

Our study highlights the importance of incorporating long-read sequencing technologies to accurately detect and quantify STR expansions. The limitations of short-read sequencing in identifying and characterizing STR expansions emphasize the need for advanced sequencing approaches in future studies of NDDs. Moreover, our findings suggest that therapeutic strategies aimed at enhancing DNA repair mechanisms could be beneficial for individuals with NDDs, particularly those exhibiting developmental regression.

Investigating potential interventions to modulate the activity of *FAN1* and other DNA repair genes of the FA pathway could offer novel therapeutic avenues.

The primary limitation of this study is the small sample size of our cohort, which may restrict the generalizability and statistical power of our findings. To address this, future studies should be conducted on larger cohorts to validate and potentially extend our results. Additionally, the innovative nature of our study prevented us from using well-established bioinformatics protocols. Although this novelty presents an exciting opportunity to explore new methodologies, it also emphasizes the need for further refinement and validation to ensure the robustness and reproducibility of our findings in broader contexts. To address this issue, we used several different algorithms to comprehensively assess the increase in STR expansions in the analyzed probands. By utilizing a diverse set of tools, we aimed to cross-validate our findings and ensure a more robust analysis of somatic STR expansions across the probands in our study.

The detection of somatic variants is strictly constrained by sequencing depth. Given the coverage used in this study, the frequency of alternative alleles is close to the threshold of detection. Future studies should incorporate both short- and long- reads sequencing at higher depth (eg, >100× coverage). Moreover, our analysis was performed on blood samples. Considering the propensity of postmitotic neurons to accumulate somatic variants, we hypothesize that the burden of genomic variations in postmortem brain tissue will be higher, warranting further investigation.

In summary, our study provides new insights into the role of DDR genes and somatic STR expansions in NDDs, particularly in cases with developmental regression. The association between DDR genes dysfunction, increased somatic variant burden, and severe neurodevelopmental phenotypes underscores the importance of DDR pathways in the pathogenesis of these disorders. Although variants in DDR genes may not be the only cause of NDD with regression and they may not be exclusively associated with it given they are also present in NDD probands without regression, the detection of combinations of variants in DDR genes in NDD individuals with regression is very intriguing and strongly calls for additional epidemiological and molecular studies to prove a causal association. Further research into the mechanisms through which *FAN1* and other DDR genes influence neurodevelopment is crucial for developing targeted therapies and improving our understanding of NDDs.

Data Availability

All data supporting this study are included with the article and the supplemental material. Clinical data and genetic variants details are openly available in the Decipher database at https://www.deciphergenomics.org/, with the following accession numbers: 539013 (F36A); 539393 (F74A); 539396 (F82A); 539397 (F19A).

Acknowledgments

The authors wish to thank the families and referring nurses of the Child Neuropsychiatry Unit, IRCCS Istituto Giannina Gaslini, for their help and collaboration. The authors are also indebted to Eva Ferri, Emilie Jacquemod, Rosa Cossu, and Alessandra Gastaldi (IIT) for their scientific project manager assistance and technical editing of the manuscript.

Funding

This research received funding from: the European Union-NextGeneration EU PNRR-MR1-2022-12376067 ''Multiomic strategies to implement the diagnostic workflow of rare diseases'' Missione M6/componente: C2 Investimento: 2.1 ''Valorizzazione e potenziamento della ricerca biomedica del SSN'' del Piano nazionale di Ripresa e Resilienza (PNRR) CUP G33C22001060001 to A.P.; the Italian Ministry of Health (Ricerca Corrente); the Italian Ministry for Education, University and Research (Ministero dell'Istruzione, dell'Universita` e della Ricerca - MIUR) PRIN2020 code 20203P8C3X to F.Z. and A.P.; by the Project 5000genomi@VdA, co-founded by "Fondo Europeo di Sviluppo Regionale (FESR CUPB68H19005520007) to AA, AC, and SG; ST and MV were supported by "Fondo Europeo di Sviluppo Regionale (FESR CUPB68H19005520007); GS was recipient of a fellowship from Programma Investimenti per la crescita e l'occupazione 2014/20" (European Social Fund, ESF CUP

B65F19001200009) and continued his activity with a fellowship from Programma Regionale Fondo sociale europeo Plus 2021/2027 (European Social Fund, ESF CUP J51B24000170002).

Author Contributions

Conceptualization: G.S., M.C., S.G., A.P.; Data Curation: S.B., S.C., G.R., M.M., L.P., M.T.D., E.D.G., L.O., L.N.; Methodology: F.C., M.D.D., S.T., D.V.; Investigation: M.C., G.S., L.P., L.F., M.F., I.C., A.B.; Formal Analysis: G.S., M.C., A.P.; Supervision: R.S., A.P., A.C., S.G.; Resources: A.P., S.G., F.Z.; Funding Acquisition: A.P., F.Z., A.A., A.C., S.G.; Writing-original draft: M.C., G.S., S.G., A.P.; Writing-review and editing: M.V., A.A., F.Z., A.C.

ORCIDs

Maria Cerminara: https://orcid.org/0000-0002-2011-6479
Giovanni Spirito: https://orcid.org/0000-0003-1391-4218
Luca Pandolfini: https://orcid.org/0000-0003-1444-8167
Silvia Boeri: https://orcid.org/0000-0002-8398-8936
Giulia Rosti: https://orcid.org/0000-0002-5949-803X
Margherita Mancardi: https://orcid.org/0000-0001-5122-4029
Livia Pisciotta: https://orcid.org/0000-0001-7653-981X
Loretta Ferrera: https://orcid.org/0000-0003-0528-1158
Francesco Caroli: https://orcid.org/0000-0003-1828-6900
Marco Di Duca: https://orcid.org/0000-0001-8456-7706
Andrea Cavalli: https://orcid.org/0000-0003-4063-4502
Maria Teresa Divizia: https://orcid.org/0000-0002-8776-0015
Elisa De Grandis: https://orcid.org/0000-0002-9733-9216
Silvia Casabona: https://orcid.org/0009-0006-4735-2902
Sara Trova: https://orcid.org/0000-0001-9270-4977
Diego Vozzi: https://orcid.org/0000-0002-2902-8184
Antonio Amoroso: https://orcid.org/0000-0002-9437-9407
Remo Sanges: https://orcid.org/0000-0002-5047-9713
Manuela Vecchi: https://orcid.org/0000-0002-2750-9160
Lino Nobili: https://orcid.org/0000-0001-9317-5405
Federico Zara: https://orcid.org/0000-0001-9744-5222

Ethics Declaration

This study was reviewed and approved by the Ethics Committee of the Italian Regione Liguria (P.R. 399REG2017). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) and minor(s) legal guardian/next of kin for the publication of data included in this article.

Conflict of Interest

The authors declare no conflicts of interest.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2025.101661) contains supplemental material, which is available to authorized users.

Affiliations

[1]Medical Genetics Unit, IRCCS Istituto Giannina Gaslini, Genoa, Italy; [2]Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health (DiNOGMI), University of Genoa, Genoa, Italy; [3]Non-coding RNAs and RNA-based therapeutics, Italian Institute of Technology (IIT), CMP[3]VdA, Aosta, Italy; [4]Genomic Facility, Italian Institute of Technology (IIT), Genova, Italy; [5]Child Neuropsychiatry Unit, ERN EpiCARE Centre, IRCCS Istituto Giannina Gaslini, Genoa, Italy; [6]Clinical Genetics and Genomics Unit, IRCCS Istituto Giannina Gaslini, Genoa, Italy; [7]Child Neuropsychiatry Unit, ASST Fatebenefratelli Sacco, Milan, Italy; [8]University of Milan, Milan, Italy; [9]Computational and chemical biology, Italian Institute of Technology, CMP[3]VdA, 11100 Aosta, Italy; [10]Centre Européen de Calcul Atomique et Moléculaire (CECAM), Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [11]Center for Human Technologies, Computational and Chemical Biology, Fondazione Istituto Italiano di Tecnologia, Genoa, Italy; [12]Pediatric Neuropsychiatry, Healthcare agency of the Aosta Valley, Beauregard Hospital, Aosta, Italy; [13]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Area of Neuroscience, Trieste, Italy

References

1. Morris-Rosendahl DJ, Crocq MA. Neurodevelopmental disorders-the history and future of a diagnostic concept. *Dial Clin Neurosci*. 2020;22(1):65-72. http://doi.org/10.31887/DCNS.2020.22.1/macrocq
2. Thapar A, Cooper M, Rutter M. Neurodevelopmental disorders. *Lancet Psychiatry*. 2017;4(4):339-346. http://doi.org/10.1016/S2215-0366(16)30376-5
3. Asadollahi R, Oneda B, Joset P, et al. The clinical significance of small copy number variants in neurodevelopmental disorders. *J Med Genet*. 2014;51(10):677-688. http://doi.org/10.1136/jmedgenet-2014-102588
4. Brunet T, Jech R, Brugger M, et al. De novo variants in neurodevelopmental disorders-experiences from a tertiary care center. *Clin Genet*. 2021;100(1):14-28. http://doi.org/10.1111/cge.13946
5. Alibutud R, Hansali S, Cao X, et al. Structural variations contribute to the genetic etiology of autism spectrum disorder and language

impairments. *Int J Mol Sci*. 2023;24(17):13248. http://doi.org/10.3390/ijms241713248
6. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev*. 2012;22(3):229-237. http://doi.org/10.1016/j.gde.2012.03.002
7. Leblond CS, Cliquet F, Carton C, et al. Both rare and common genetic variants contribute to autism in the Faroe Islands. *npj Genom Med*. 2019;4:1. http://doi.org/10.1038/s41525-018-0075-2
8. Niemi MEK, Martin HC, Rice DL, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018;562(7726):268-271. http://doi.org/10.1038/s41586-018-0566-4
9. Leblond CS, Le TL, Malesys S, et al. Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Mol Cell Neurosci*. 2021;113:103623. http://doi.org/10.1016/j.mcn.2021.103623
10. Abrahams BS, Arking DE, Campbell DB, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*. 2013;4(1):36. http://doi.org/10.1186/2040-2392-4-36
11. Qing X, Zhang G, Wang ZQ. DNA damage response in neurodevelopment and neuromaintenance. *FEBS Journal*. 2023;290(13):3300-3310. http://doi.org/10.1111/febs.16535
12. Khokhlova EV, Fesenko ZS, Sopova JV, Leonova EI. Features of DNA repair in the early stages of mammalian embryonic development. *Genes*. 2020;11(10):1138. http://doi.org/10.3390/genes11101138
13. Chow ML, Pramparo T, Winn ME, et al. Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. *PLOS Genet*. 2012;8(3):e1002592. http://doi.org/10.1371/journal.pgen.1002592
14. Stoner R, Chow ML, Boyle MP, et al. Patches of disorganization in the neocortex of children with autism. *N Engl J Med*. 2014;370(13):1209-1219. http://doi.org/10.1056/NEJMoa1307491
15. Trost B, Engchuan W, Nguyen CM, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*. 2020;586(7827):80-86. http://doi.org/10.1038/s41586-020-2579-z
16. Mitra I, Huang B, Mousavi N, et al. Patterns of de novo tandem repeat mutations and their role in autism. *Nature*. 2021;589(7841):246-250. http://doi.org/10.1038/s41586-020-03078-7
17. Liu T, Ghosal G, Yuan J, Chen J, Huang J. FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. *Science*. 2010;329(5992):693-696. http://doi.org/10.1126/science.1192656
18. Deshmukh AL, Porro A, Mohiuddin M, et al. FAN1, a DNA repair nuclease, as a modifier of repeat expansion disorders. *J Huntingtons Dis*. 2021;10(1):95-122. http://doi.org/10.3233/JHD-200448
19. Kim KH, Hong EP, Shin JW, et al. Genetic and functional analyses point to FAN1 as the source of multiple Huntington disease modifier effects. *Am J Hum Genet*. 2020;107(1):96-110. http://doi.org/10.1016/j.ajhg.2020.05.012
20. McAllister B, Donaldson J, Binda CS, et al. Exome sequencing of individuals with Huntington's disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. *Nat Neurosci*. 2022;25(4):446-457. http://doi.org/10.1038/s41593-022-01033-5
21. Phadte AS, Bhatia M, Ebert H, et al. FAN1 removes triplet repeat extrusions via a PCNA- and RFC-dependent mechanism. *Proc Natl Acad Sci U S A*. 2023;120(33):e2302103120. http://doi.org/10.1073/pnas.2302103120
22. Zhao X, Lu H, Usdin K. FAN1's protection against CGG repeat expansion requires its nuclease activity and is FANCD2-independent. *Nucleic Acids Res*. 2021;49(20):11643-11652. http://doi.org/10.1093/nar/gkab899
23. Hujoel MLA, Handsaker RE, Kamitaki N, et al. Insights into the causes and consequences of DNA repeat expansions from 700,000 biobank participants. *bioRxiv*. 2024. http://doi.org/10.1101/2024.11.25.625248
24. Furley K, Mehra C, Goin-Kochel RP, et al. Developmental regression in children: current and future directions. *Cortex*. 2023;169:5-17. http://doi.org/10.1016/j.cortex.2023.09.001
25. Cunningham JL, Frankovich J, Dubin RA, et al. Ultrarare variants in DNA damage repair genes in pediatric acute-onset neuropsychiatric syndrome or acute behavioral regression in neurodevelopmental disorders. *Dev Neurosci*. 2024;47(4):231-250. http://doi.org/10.1159/000541908
26. Barthelemy J, Hanenberg H, Leffak M. FANCJ is essential to maintain microsatellite structure genome-wide during replication stress. *Nucleic Acids Res*. 2016;44(14):6803-6816. http://doi.org/10.1093/nar/gkw433
27. Cerminara M, Spirito G, Pisciotta L, et al. Case report: whole exome sequencing revealed disease-causing variants in two genes in a patient with autism spectrum disorder, intellectual disability, hyperactivity, sleep and gastrointestinal disturbances. *Front Genet*. 2021;12:625564. http://doi.org/10.3389/fgene.2021.625564
28. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat*. 2021;43(8):1012-1030. http://doi.org/10.1002/humu.24309
29. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(database issue):D789-D798. http://doi.org/10.1093/nar/gku1205
30. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(database issue):D980-D985. http://doi.org/10.1093/nar/gkt1113
31. Baldarelli RM, Smith CL, Ringwald M, Richardson JE, Bult CJ. Mouse Genome Informatics Group. Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse. *Genetics*. 2024;227(1):iyae031. http://doi.org/10.1093/genetics/iyae031
32. Richards S, Aziz N, Bale S, et al. 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*. 2015;17(5):405-424. http://doi.org/10.1038/gim.2015.30
33. Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105-1117. http://doi.org/10.1038/gim.2017.37
34. Renaux A, Papadimitriou S, Versbraegen N, et al. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res*. 2019;47(W1):W93-W98. http://doi.org/10.1093/nar/gkz437
35. Lee JM, Huang Y, Orth M, et al. Genetic modifiers of Huntington disease differentially influence motor and cognitive domains. *Am J Hum Genet*. 2022;109(5):885-899. http://doi.org/10.1016/j.ajhg.2022.03.004
36. Rao X, Huang X, Zhou Z, Lin X. An improvement of the 2(-delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. *Biostat Bioinforma BIOMATH*. 2013;3(3):71-85.
37. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006. http://doi.org/10.1101/gr.229102
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. http://doi.org/10.1093/bioinformatics/btp324
39. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. http://doi.org/10.1093/bioinformatics/btp352
40. Dolzhenko E, Bennett MF, Richmond PA, et al. ExpansionHunter DeNovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol*. 2020;21(1):102. http://doi.org/10.1186/s13059-020-02017-z
41. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947. http://doi.org/10.1093/nar/gky1015
42. Wala JA, Bandopadhayay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res*. 2018;28(4):581-591. http://doi.org/10.1101/gr.221028.117

43. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11. http://doi.org/10.1186/s13100-015-0041-9
44. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100. http://doi.org/10.1093/bioinformatics/bty191
45. Chiu R, Rajan-Babu IS, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol*. 2021;22(1):224. http://doi.org/10.1186/s13059-021-02447-3
46. Jiang T, Liu Y, Jiang Y, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020;21(1):189. http://doi.org/10.1186/s13059-020-02107-y
47. Mitsuhashi S, Frith MC, Mizuguchi T, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol*. 2019;20(1):58. http://doi.org/10.1186/s13059-019-1667-6
48. Ibañez K, Polke J, Hagelstrom RT, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol*. 2022;21(3):234-245. http://doi.org/10.1016/S1474-4422(21)00462-2
49. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res*. 2019;47(15):e90. http://doi.org/10.1093/nar/gkz501
50. Sherman BT, Hao M, Qiu J, et al. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. http://doi.org/10.1093/nar/gkac194
51. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573-580. http://doi.org/10.1093/nar/27.2.573
52. Riggs ER, Andersen EF, Cherry AM, 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*. 2020;22(2):245-257. http://doi.org/10.1038/s41436-019-0686-8
53. Fierheller CT, Guitton-Sert L, Alenezi WM, et al. A functionally impaired missense variant identified in French Canadian families implicates FANCI as a candidate ovarian cancer-predisposing gene. *Genome Med*. 2021;13(1):186. http://doi.org/10.1186/s13073-021-00998-5
54. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492. http://doi.org/10.1126/science.adg7492
55. Leone R, Zuglian C, Brambilla R, Morella I. Understanding copy number variations through their genes: a molecular view on 16p11.2 deletion and duplication syndromes. *Front Pharmacol*. 2024;15:1407865. http://doi.org/10.3389/fphar.2024.1407865
56. Girirajan S, Rosenfeld JA, Coe BP, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N Engl J Med*. 2012;367(14):1321-1331. http://doi.org/10.1056/NEJMoa1200395
57. Meir A, Greene EC. Srs2 and Pif1 as model systems for understanding Sf1a and Sf1b helicase structure and function. *Genes*. 2021;12(9):1319. http://doi.org/10.3390/genes12091319
58. Ling C, Ishiai M, Ali AM, et al. FAAP100 is essential for activation of the fanconi anemia-associated DNA damage response pathway. *EMBO J*. 2007;26(8):2104-2114. http://doi.org/10.1038/sj.emboj.7601666
59. Betancur C, Buxbaum JD. SHANK3 haploinsufficiency: a “common” but underdiagnosed highly penetrant monogenic cause of autism spectrum disorders. *Mol Autism*. 2013;4(1):17. http://doi.org/10.1186/2040-2392-4-17
60. Pengo M, Squitieri F. Beyond CAG repeats: the multifaceted role of genetics in Huntington disease. *Genes*. 2024;15(6):807. http://doi.org/10.3390/genes15060807
61. Fouquet B, Pawlikowska P, Caburet S, et al. A homozygous FANCM mutation underlies a familial case of non-syndromic primary ovarian insufficiency. *eLife*. 2017;6:e30490. http://doi.org/10.7554/eLife.30490
62. Zaslavsky K, Zhang WB, McCready FP, et al. SHANK2 mutations associated with autism spectrum disorder cause hyperconnectivity of human neurons. *Nat Neurosci*. 2019;22(4):556-564. http://doi.org/10.1038/s41593-019-0365-8
63. Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism*. 2017;8:21. http://doi.org/10.1186/s13229-017-0137-9
64. Direk N, Williams S, Smith JA, et al. An analysis of two genome-wide association meta-analyses identifies a new locus for broad depression phenotype. *Biol Psychiatry*. 2017;82(5):322-329. http://doi.org/10.1016/j.biopsych.2016.11.013
65. Goold R, Flower M, Moss DH, et al. FAN1 modifies Huntington’s disease progression by stabilizing the expanded HTT CAG repeat. *Hum Mol Genet*. 2019;28(4):650-661. http://doi.org/10.1093/hmg/ddy375
66. Maia N, Nabais Sá MJ, Oliveira C, et al. Can the synergic contribution of multigenic variants explain the clinical and cellular phenotypes of a neurodevelopmental disorder? *Genes*. 2021;13(1):78. http://doi.org/10.3390/genes13010078