bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Long-read genome sequencing for the diagnosis of neurodevelopmental disorders

Susan M. Hiatt[1], James M.J. Lawlor[1], Lori H. Handley[1,] Ryne C. Ramaker[1], Brianne B. Rogers[1], E. Christopher Partridge[1], Lori Beth Boston[1,] Melissa Williams[1,] Jerry Jenkins[1], David E. Gray[1], Kevin M. Bowling[1], E. Martina Bebin[2], Jane Grimwood[1], Jeremy Schmutz[1], Gregory M. Cooper[1]*

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA, 35806

[2]University of Alabama at Birmingham, Birmingham, AL, USA, 35924

*gcooper@hudsonalpha.org, 256-327-9490

Conflicts of Interest

The authors all declare no conflicts of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Exome and genome sequencing have proven to be effective tools for the diagnosis of neurodevelopmental disorders (NDDs), but large fractions of NDDs cannot be attributed to currently detectable genetic variation. This is likely, at least in part, a result of the fact that many genetic variants are difficult or impossible to detect through typical short-read sequencing approaches. Here, we describe a genomic analysis using Pacific Biosciences circular consensus sequencing (CCS) reads, which are both long (>10 kb) and accurate (>99% bp accuracy). We used CCS on six proband-parent trios with NDDs that were unexplained despite extensive testing, including genome sequencing with short reads. We identified variants and created *de novo* assemblies in each trio, with global metrics indicating these data sets are more accurate and comprehensive than those provided by short-read data. In one proband, we identified a likely pathogenic (LP), *de novo* L1-mediated insertion in *CDKL5* that results in duplication of exon 3, leading to a frameshift. In a second proband, we identified a *de novo* translocation affecting *DGKB* and *MLLT3*, which we show disrupts *MLLT3* transcript levels. We consider this a variant of uncertain significance (VUS). The breadth and quality of variant detection coupled to finding variants of clinical and research interest in two of six probands with unexplained NDDs strongly support the value of long-read genome sequencing for understanding rare disease.

Key Words

Long read sequencing, Clinical sequencing, neurodevelopmental disorder, structural variation, mobile element insertion

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Neurodevelopmental disorders (NDDS) are a heterogeneous group of conditions that lead to a range of physical and intellectual disabilities and collectively affect 1-3% of children[1]. Many NDDs result from large-effect genetic variation, which often occurs *de novo*[2], with hundreds of genes known to associate with disease[3]. Owing to this combination of factors, exome and genome sequencing (ES/GS) have proven to be powerful tools for both clinical diagnostics and research on the genetic causes of NDDs. However, while discovery power and diagnostic yield of genomic testing have consistently improved over time[4], most NDDs cannot be attributed to currently detectable genetic variation[5].

There are a variety of hypotheses that might explain the fact that most NDDs cannot be traced to a causal genetic variant after ES/GS, including potential environmental causes and complex genetic effects driven by small-effect variants[6]. However, one likely possibility is that at least some NDDs result from highly penetrant variants that are missed by typical genomic testing. ES/GS are generally performed by generating millions of "short" sequencing reads, often paired-end 150 bp reads, followed by alignment of those reads to the human reference assembly and detection of variation from the reference. Various limitations of this process, such as confident alignment of variant reads to a unique genomic location, make it difficult to detect many variants, including some known to be highly penetrant contributors to disease. Examples of NDD-associated variation that might be missed include low-complexity repeat variants[7], small to moderately-sized structural variants (SVs)[4,8], and mobile element insertions (MEIs)[9,10]. Indeed, despite extensive effort from many groups, detection of such variation

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

remains plagued by high error rates, both false positives and false negatives, and it is likely that many such variants are simply invisible to short read analysis[11].

One potential approach to overcome variant detection limitations in ES/GS is to use sequencing platforms that provide longer reads, which allow for more comprehensive and accurate read alignment to the reference assembly, including within and near to repetitive regions, and *de novo* assembly[12]. Recently, Pacific Biosciences released an approach, called Circular Consensus Sequencing (CCS), or "HiFi", in which fragments of DNA are circularized and then sequenced repeatedly[13]. This leads to sequence reads that are both long (>10 kb) and accurate at the basepair level (>99%). In principle, such an approach holds great potential for more comprehensive and accurate detection of human genetic variation, especially in the context of rare genetic disease.

We have used CCS to analyze six proband-parent trios affected with NDDs that we previously sequenced using a typical Illumina genome sequencing (IGS) approach but in whom no causal genetic variant, or even potentially causal variant, was found. The CCS data were used to detect variation within each trio and generate *de novo* genome assemblies, with a variety of metrics indicating that the results are more comprehensive and accurate, especially for complex variation, than those seen in short-read datasets. In one proband, we identified a complex L1-mediated *de novo* insertion within *CDKL5* that leads to a duplicated coding exon and is predicted to lead to a frameshift and loss-of-function. Transcript analyses confirm that the duplicated exon is spliced into mRNA in the proband. We have classified this variant as likely pathogenic using American College of Medical Genetics (ACMG) standards[14]. In a second proband, we found a *de novo* translocation that leads to disruption of *MLLT3*. Transcript

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

analyses indicate reduced *MLLT3* levels in the proband. We have classified this as a variant of uncertain significance (VUS). At a high level, these data strongly support the value of long-read genome analysis for the detection of NDD-associated variation, and more broadly for the analysis of human genetic disease.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Materials and Methods

Illumina sequencing, variant calling and analysis

Six probands and their unaffected parents were enrolled in a research study aimed at identifying genetic causes of NDDs[15], which was monitored by Western IRB (20130675). All six of these families underwent trio Illumina genome sequencing (IGS) between four and five years ago, which was performed as described[15]. Briefly, whole blood genomic DNA was isolated using the QIAsymphony (Qiagen), and sequencing libraries were constructed by the HudsonAlpha Genomic Services Lab. Sequencing was performed on the Illumina HiSeqX using paired end reads with a read length of 150 base pairs. Each genome was sequenced at an approximate mean depth of 30X, with at least 80% of base positions reaching 20X coverage. While originally analyzed using hg37, for this study reads were aligned to hg38 using DRAGEN version 07.011.352.3.2.8b. Variants were discovered (in gvcf mode) with DRAGEN and joint genotyping was performed across six trios using GATK version 3.8-1-0-gf15c1c3ef. Structural variants (SVs) were called using a combination of Delly[16], CNVnator[17], ERDS[18], and Manta[19], followed by heuristic merging of SVs from the different callers based on breakpoint proximity and SV type. SVs are also annotated with gene features and allele frequencies from dbVar[20], relevant developmental delay publications[21,22], and an internal SV database. Mobile element insertions (MEIs) were called using MELT[23] run in MELT-SINGLE mode. Variant analysis and interpretation was performed using ACMG guidelines[14], similar to that which we previously performed[4,15]. None of the probands had a Pathogenic (P), Likely Pathogenic (LP), or Variant of Uncertain Significance (VUS) identified by IGS, either at the time of original analysis or after a reanalysis performed at the time of generation of long-read data. In all trios, expected relatedness was

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

confirmed[24]. IGS data for Probands 1-5 are available via dbGAP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001089.v3.p1). Project Accession Number: phs001089. Complete IGS data for proband 6 is not available due to consent restrictions.

Long-Read sequencing, variant calling, analysis and de novo assemblies

Long-read sequencing was performed using Circular Consensus Sequence (CCS) mode on a PacBio Sequel II instrument (Pacific Biosciences of California, Inc.). Libraries were constructed using a SMRTbell Template Prep Kit 1.0 and tightly sized on a SageELF instrument (Sage Science, Beverly, MA, USA). Sequencing was performed using a 30 hour movie time with 2 hour pre-extension and the resulting raw data was processed using either the CCS3.4 or CCS4 algorithm, as the latter was released during the course of the study. Comparison of the number of high-quality indel events in a read versus the number of passes confirmed that these algorithms produced comparable results. Probands were sequenced to an average CCS depth of 30X (range 25 to 35), while parents were covered at an average depth of 16x (range 10 to 22, see Table 1, Supplemental Table 2). CCS reads were aligned to the complete GRCh38.p13 human reference. For SNVs and indels, CCS reads were aligned using the Sentieon v.201808.07 implementation of the BWA-MEM aligner (https://www.biorxiv.org/content/10.1101/396325v1), and variants were called using DeepVariant v0.10[25] and joint-genotyped using GLNexus v1.2.6 (https://www.biorxiv.org/content/10.1101/343970v1). For structural variants (SVs), reads were aligned using pbmm2 1.0.0 (https://github.com/PacificBiosciences/pbmm2) and SVs were

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

called using pbsv v2.2.2 (https://github.com/PacificBiosciences/pbsv). Candidate *de novo* SVs required a proband genotype of 0/1 and parent genotypes of 0/0, with ≥ 6 alternate reads in the proband and 0 alternate reads, ≥5 reference reads in the parents. *De novo* assemblies were generated by assembling an average of 29.67x CCS coverage for probands and 16.08x CCS coverage for parents (Supplemental Table 2) using the Canu v1.8[26] assembler. Insertion/deletion errors in the consensus sequence were not polished. This produced assemblies consisting of an average of 22,560 scaffolds for probands and 17,760 scaffolds for parents, with a contig N50 of 10.02 Mb for probands and 4.68 Mb for parents, and an average assembled size for probands of 3,515.5 Mb and for parents 3,372.6 Mb (Supplemental Table 2). Dot plots illustrating sequence differences were created using Gepard[27].

PacBio CCS data for Probands 1-5 will be submitted to dbGAP.

QC Statistics

SNV and indel concordance and *de novo* variant counts (Supplemental Tables 1A, 1C) were calculated using bcftools v1.9 and rtg-tools vcfeval v3.9.1. "High-quality *de novo"* variants were defined as PASS variants (IGS/GATK only) on autosomes (on primary contigs only) that were biallelic with DP ≥ 7 and genotype quality (GQ) ≥ 35. Additional requirements were a proband genotype of 0/1, with ≥ 2 alternate reads and an allele balance ≥ 0.3 and ≤ 0.7. Required parent genotypes were 0/0, with alternate allele depth of 0. Mendelian error rates were also calculated using bcftools (Supplemental Table 1D). "Rigorous" error rates were restricted to PASS variants (IGS/GATK only) on autosomes with GQ>20, and total allele depth (DP) >5. Total

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variant counts per trio (Supplemental Table 1B) were calculated using VEP (v98), counting multi-allelic sites as one variant. SV counts (Supplemental Table 1E) were calculated using bcftools and R. Counts were restricted to calls designated as "PASS", with an alternate AD ≥2. Candidate SV *de novos* required proband genotype of 0/1 and parent genotypes of 0/0, with ≥ 6 alternate reads in the proband and 0 alternate reads, ≥5 reference reads in the parents *De novo* MELT calls (Supplemental Table 5B) in IGS data were defined as isolated proband calls where the parent did not have the same-type (ALU, L1, or SVA) of call within 1 kb as calculated by bedtools closest v.2.25.0. These calls were then filtered (using bcftools) for "PASS" calls and varying depths, defined as the number of read pairs supporting both sides of the breakpoint (LP, RP). To create a comparable set of *de novo* mobile element calls in CCS data (Supplemental Table 5D), individual calls were extracted from the pbsv joint-called VCF using bcftools and awk and isolated proband calls were defined as they were for the IGS data and filtered (using bcftools) for PASS calls and varying depths, defined as the proband alternate allele depth (AD[1]).

Simple repeat and low mappability regions

We generated a bed file of disease-related low-complexity repeat regions in 35 genes from previous studies[7,28]. Most regions (25) include triplet nucleotide repeats, while the remainder include repeat units of 4-12 bp (Supplemental Table 3A). Reads aligning to these regions were extracted from bwa-mem-aligned bams and visualized using the Integrated Genomics Viewer (IGV[29]). Proband depths of MAPQ60 reads spanning each region (Supplemental Table 4A) were calculated using bedtools multicov v2.28.0. For the depth calculations, regions were expanded

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

by 15 bp on either side (using bedtools slop) to count reads anchored into non-repeat sequence. The mean length of these regions was 83 bp, with a max of 133 bp.

Low mappability regions were defined as the regions of the genome that do not lie in Umap k100 mappable regions (https://bismap.hoffmanlab.org/)[30]. Regions ≥ 100,000 nt long and those on non-primary contigs were removed, leaving a total of 242,222 difficult-to-map regions with average length 411 bp. Proband depths of MAPQ60 reads spanning each region were calculated using bedtools multicov v2.28.0 (Supplemental Table 4B). High quality protein-altering variants (Supplemental Table 4C) in probands were defined using VEP annotations, and counted using bcftools v1.9. Requirements included a heterozygous or homozygous genotype in the proband, with ≥4 alternate reads, an allele balance ≥ 0.3 and ≤ 0.7, GQ>20, and DP>5. Reads supporting 57 loss-of-function variants (high-quality and low-quality) in Proband 5 were visualized with IGV and semi-quantitatively scored to assess call accuracy. Approximate counts of reads were recorded and grouped by mapping quality (MapQ=0 and MapQ≥1), along with subjective descriptions of the reads (Supplemental Table 4D). The total evidence across CCS and IGS reads was used to estimate truth and score each variant call as true positive (TP), false positive (FP), true negative (TN), false negative (FN), or undetermined (UN), see Supplemental Table 4D and 4E).

CDKL5 cDNA Amplicon Sequencing

Total RNA was extracted from whole blood in PAXgene tubes using a PAXgene Blood RNA Kit

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

version 2 (PreAnalytiX, #762164) according to the manufacturer's protocol. cDNA was generated with a High Capacity Reverse Transcription Kit (Applied Biosystems, #4368814) using 500 ng of extracted RNA from each individual as input. Primers were designed to *CDKL5* exons 2, 5, and 6 to generate two amplicons spanning the potentially disrupted region of *CDKL5* mRNA. Select amplicons were purified and sent to MCLAB (Molecular Cloning Laboratories, South San Francisco, CA, USA) for Sanger sequencing. See Supplemental Methods for additional details, including primers.

CDKL5 Genomic DNA PCR

We performed PCR to amplify products spanning both junctions of the insertion, in addition to the majority of the insertion using the genomic DNA (gDNA) of the proband and parents as template. Select amplicons were purified and sent to MCLAB (Molecular Cloning Laboratories, South San Francisco, CA, USA) for Sanger sequencing. See Supplemental Methods for additional details, including primers.

DGKB/MLLT3 qPCR

Total RNA was extracted from whole blood using a PAXgene Blood RNA Kit version 2 (PreAnalytiX, #762164) and cDNA was generated with a High Capacity Reverse Transcription Kit (Applied Biosystems, #4368814) in an identical fashion as described for *CDKL5* cDNA amplicon sequencing. For qPCR, Two TaqMan probes targeting the MLLT3 exon 3-4 and exon 9-10 splice junctions (ThermoFisher, Hs00971092_m1 and Hs00971099_m1) were used with cDNA diluted 1:5 in dH2O to perform qPCR for six replicates per sample on an Applied Biosystems Quant

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Studio 6 Flex. Differences in CT values from the median CT values for either an unrelated family or the proband's parents were used to compute relative expression levels. See Supplemental Methods for additional details, including primers.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Results

Affected probands and their unaffected parents were enrolled in a research study aimed at identifying genetic causes of NDDs[15]. All trios were originally subject to standard Illumina genome sequencing (IGS) and analysis using ACMG standards[14] to find pathogenic (P) or likely pathogenic (LP) variants, or variants of uncertain significance (VUS). Within the subset of probands for which no variants of interest (P, LP, VUS) were identified either originally or after subsequent reanalyses[4,15], six trios were selected for sequencing using the PacBio Sequel II Circular Consensus Sequencing (CCS) approach (Table 1). These trios were selected for those with a strong suspicion of a genetic disorder, in addition to diversifying with respect to gender and ethnicity. Parents were sequenced, at a relatively reduced depth, to facilitate identification of *de novo* variation.

QC of CCS data

Variant calls from CCS data and IGS data were largely concordant (Supplemental Table 1A). When comparing each individuals' variant calls in the Genome in a Bottle (GIAB) high confidence regions[31] between CCS and IGS, concordance was 94.63%, with higher concordance for SNVs (96.88%) than indels (75.96%). Concordance was slightly higher for probands only, likely due to the lower CCS read-depth coverage in parents. While CCS data showed a consistently lower number of SNV calls than IGS (mean = 7.0 M vs. 7.45 M, per trio), more *de novo* SNVs at high QC stringency were produced in CCS data than IGS (mean SNVs= 89 vs. 38, Supplemental Table 1B, 1C). CCS yielded far fewer *de novo* indels at these same thresholds (mean indels 11 vs. 148), with the IGS *de novo* indel count being much higher than biological

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

expectation[32] and likely mostly false positive calls (Supplemental Table 1C). In examining reads supporting variation that was uniquely called in each set, we found that CCS false positive *de novos* were usually false negative calls in the parent, due to lower genome-wide coverage in the parent and the effects of random sampling (i.e., sites at which there were 7 or more CCS reads in a parent that randomly happened to all derive from the same haplotype, Supplemental Table 1C). Mendelian error rates in autosomes were noticeably lower in CCS data relative to IGS (harmonic mean of high-quality calls 0.18% vs. 0.34%, Supplemental Table 1D), suggesting the CCS SNV calls are of higher accuracy, consistent with previously published data[13].

Each trio had an average of ~56,000 SVs among all three members, including an average of 59 candidate *de novo* SVs per proband (Supplemental Table 1E). Trio SVs mainly represent insertions (48%) and deletions (43%), followed by duplications (6%), single breakends (3%), and inversions (<1%).

De novo genomes were assembled for each individual. The average N50 for all 18 individuals was 6.5 Mb, with noticeably higher values for probands (10.0 Mb) than parents (4.7 Mb), reflecting the increased sequencing depth in the probands (Supplemental Table 2).

Variation in Simple Repeat regions

Accurate genotyping of simple repeat regions like trinucleotide repeat expansions presents a challenge in short read data where the reads are often not long enough to span variant alleles. We assessed the ability of CCS to detect variation in these genomic regions, and compared that to IGS. We first examined variation in *FMR1* (MIM: 309550). Expansion of a trinucleotide repeat in the 5' UTR of *FMR1* is associated with Fragile X syndrome (MIM: 300624), the second-most

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

common genetic cause of intellectual disability[33]. Visualization of this region in all 18 individuals indicated insertions in all but two samples in the CGG repeat region of *FMR1* relative to hg38, with a range of insertion sizes from 6-105 bp (Supplemental Table 3, Supplemental Figure 1). When manually inspecting these regions, while one or two major alternative alleles are clearly visible, there are often minor discrepancies in insertion lengths, often by multiples of 3. It is unclear if this represents true somatic variation, or if this represents inaccuracy of sequencing.

Like that for *FMR1*, manual curation of 34 other disease-causal repeat regions in each proband indicated that alignment of CCS reads provides a more accurate assessment of variation in these regions compared to IGS. When looking at region-spanning reads with high quality alignment (mapQ=60), 97% (34 of 35) of the regions were covered by at least 10 CCS reads in all six probands, as compared to 11% (4 of 35) of regions with high-quality IGS reads (Supplemental Table 4A). While all query regions measured ≤ 144 bp (which includes an extension of 15 bp on either end of the repeat region), seven query regions were ≥100 bp. When considering only regions of interest <100 bp, 14% (4 of 28 regions) are covered by at least 10 high-quality IGS reads in each proband. Mean coverage of high-quality, region-spanning reads across probands was higher in CCS data than in IGS (29 vs. 11, Supplemental Table 4A). Of all repeat regions studied, none harbored variation classified as P/LP/VUS.

We also compared coverage of high-quality CCS and IGS reads in low mappability regions of the genome, specifically those that cannot be uniquely mapped by 100 bp kmers[30]. While over half of these regions (62.5%) were fully covered by at least 10 high quality CCS reads (mapQ=60) in all six probands, only 19.3% of the regions met the same coverage metrics in the IGS data (Supplemental Table 4B). The average CCS read depth in these regions was 26 reads,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

vs. 8 reads in IGS. Within these regions, CCS yielded twice as many high quality, protein-altering variants in each proband when compared to IGS (182 in CCS vs. 85 in IGS) (Supplemental Table 4C). Outside of the low mappability regions, counts of protein-altering variants were similar (6,627 in CCS vs. 6,759 in IGS).

To assess the accuracy of the protein-altering variant calls in low-mappability regions, we visualized reads for 57 loss-of-function variants detected by CCS, IGS, or both in Proband 5 and used the totality of read evidence to score each variant as TP, FP, TN, FN, or undetermined. Six of these were "high-quality" calls (see Methods), and all of these were correctly called in CCS (TPs, 100%); in IGS, two were correctly called (TPs, 33%) and four were undetected (FNs, 67%) (Supplemental Table 4D). Among all 57 unfiltered variant calls, most CCS calls were correct (29 TP, 15 TN, total 77%) while most IGS calls were incorrect (16 FP, 22 FN, total 67%) (Supplemental Table 4E).

Mobile Element Insertions

We searched for mobile element insertions (MEIs) in these six probands within the IGS data using MELT (Supplemental Table 5A, 5B)[23] and within CCS data using pbsv (see Methods, Supplemental Table 1E, 5C, 5D). Our results suggest that CCS detection of MEIs is far more accurate. For example, it has been estimated that there exists a *de novo* Alu insertion in ~1 in every 20 live births (mean of 0.05 per individual)[34,35]. However, at stringent QC filters (i.e., ≥5 read-pairs at both breakpoints, PASS, and no parental calls of the same MEI type within 1kb), a total of 82 candidate de novo Alu insertions (average of 13.7) were called across the six probands using the IGS data (Supplemental Table 5B), a number far larger than that expected.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Inspection of these calls indicated that most were bona fide heterozygous Alu insertions in the proband that were inherited but undetected in the parents. Filtering changes to improve sensitivity come at a cost of elevated false positive rates; for example, requiring only 2 supporting read pairs at each breakpoint leads to an average of ~55 candidate *de novo* Alu insertions per proband (Supplemental Table 5B). In contrast, using the CCS data and stringent QC filters (≥5 alternate reads, PASS, and no parental calls within 1kb) we identified a total of only 6 candidate *de novo* Alu MEIs among the 6 probands (Supplemental Table 5D), an observation that is far closer to biological expectation. We retained 4 candidate *de novo* Alu MEIs after further inspection of genotype and parental reference read depth (Supplemental Table 1E). One of these 4 appears genuine, while the other three appear to be correctly called in the proband but missed in the parents owing to low read-depth such that the Alu insertion-bearing haplotype was simply not covered by any CCS reads (Supplemental Figure 9).

A likely pathogenic de novo structural variant in CDKL5

Analysis of CCS data in proband 6 indicated a *de novo* structural variant within the *CDKL5* gene (MIM: 300203, Figure 1A). Given the *de novo* status of this event, the association of *CDKL5* with early infantile epileptic encephalopathy 2 (EIEE2, MIM: 300672), and the overlap of disease with the proband's phenotype, which includes intellectual disability, developmental delay, and seizures, we prioritized this event as the most interesting candidate variant in this proband.

Inspection of the region surrounding *CDKL5* in the proband's *de novo* assembly indicated a "primary" ~21 Mb contig and an "alternate" ~20 kb contig, each containing *CDKL5*

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(Supplemental Figure 2A). The "primary" 21 Mb contig represents a contiguous region that is well-supported by overlapping CCS reads. The smaller 20 kb alternate contig is a region where the probands' reads support a region of deviation from the primary contig. Alignment of the proband's contigs to one another and to GRCh38 identified a heterozygous 6993 bp insertion in an intron of *CDKL5* (GRCh38:chrX:18,510,871-18,510,872_ins6993, Figure 1, Supplemental Figures 2B, 3). Each parent's *de novo* assembly has only one contig in this region which aligns to the reference with no gaps. Analysis of SNVs in the region surrounding the insertion indicate that it lies on the proband's paternal allele. However, mosaicism is suspected, as there exist paternal haplotype reads within the proband that do not harbor the insertion (4 of 6 paternal reads without the insertion at the 5' end of the event, and 7 of 16 paternal reads without insertion at the 3' end of the event; Supplemental Figure 4).

Annotation of the insertion indicated that it contains three distinct segments: 4272 bp of a retrotransposed, 5' truncated L1HS mobile element (including a polyA tail), 2602 bp of sequence identical to an intron of the nearby *PPEF1* gene (NC_000023.11:g.18738310_18740911; NM_006240.2:c.235+4502_235+7103), and a 119 bp region that includes a duplicated exon 3 of *CDKL5* (35 bp) and surrounding intronic sequence (GRCh38:chrX:18510753-18510871; NM_003159.2:c.65-67 to NM_003159.2:c.99+17; 119 bp total)(Figure 1B,C). The 2,602 bp copy of PPEF1 intronic sequence includes the 5' end (1953 bp) of an L1PA5 element that is ~6.5% divergent from its consensus L1, an AluSx element, and additional repetitive and non-repetitive intronic sequence. The size and identity of this insert in the proband, and absence in both parents, was confirmed by PCR amplification and Sanger sequencing (see Supplemental Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Exon 3 of *CDKL5*, which lies within the target-site duplication of the L1-mediated insertion, is a coding exon that is 35 bp long; inclusion of a second copy of exon 3 into *CDKL5* mRNA is predicted to lead to a frameshift (Thr35ProfsTer52, Figure 2B). To determine the effect of this insertion on *CDKL5* transcripts, we performed RT-PCR from RNA isolated from each member of the trio. Using primers designed to span from exon 2 to exon 5, all three members of the trio had an expected amplicon of 240 bp. However, the proband had an additional amplicon of 275 bp (Figure 2A). Sanger sequencing of this amplicon indicated that a duplicate exon 3 was spliced into this transcript (Figure 2B). The presence of transcripts with a second copy of exon 3 strongly supports the hypothesis that the variant leads to a *CDKL5* loss-of-function effect in the proband.

A de novo structural variant affecting DGKB and *MLLT3*

Analysis of variant calls in proband 4 indicated a *de novo* structural variant affecting both *DGKB* (MIM: 604070) on chromosome 7 and *MLLT3* (MIM: 159558) on chromosome 9 (Figure 3). CCS reads and contigs from the proband's *de novo* assembly support the existence of at least three breakpoints, suggesting that a ~250 kb fragment harboring three coding exons of *DGKB* are removed from chromosome 7 and inserted into an intron of *MLLT3* on chromosome 9. As our original sequencing and analysis was unable to assemble a haplotype containing complete representation of the altered chromosomes, we generated a second, longer-insert library from this proband and generated an additional 15x coverage. While the second library and sequencing analysis again revealed the same SV breakpoints, the exact structure of these loci in this proband still cannot not be resolved. We have also thus far not been able to Sanger

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

confirm the breakpoints, as they include L1s and other repetitive sequence. However, the existence of numerous, consistent long reads generated from multiple independent library preparations strongly support the presence of a *de novo* SV disrupting these loci. Analysis of SNVs in the surrounding affected regions indicate that the breakpoints all lie on the paternal allele (Supplemental Figures 5-7).

To determine if *MLLT3* transcripts are disrupted in this proband, we performed qPCR using RNA from each member of the trio, in addition to three unrelated individuals (Family 3). Using two validated TaqMan probes near the region of interest (exons 3-4 and exons 9-10), we found that proband 4 showed a ~35-39% decrease in *MLLT3* compared to her parents and a 38-45% decrease relative to unrelated individuals (Figure 4, Supplemental Table 6). Expression of *DGKB* was not examined, as the gene is not expressed at appreciable levels in blood[36].

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Discussion

Here we describe CCS long-read sequencing of six probands with NDDs who had previously undergone extensive genetic testing with no variants found to be relevant to disease. Generally, the CCS genomes appeared to be highly comprehensive and accurate in terms of variant detection, facilitating detection of a diversity of variant types across many loci, including those that prove challenging to analysis with short reads. Detection of simple-repeat expansions and variants within low-mappability regions, for example, was far more accurate in CCS data than that seen in IGS, and many complex SVs were plainly visible in CCS data but missed by IGS.

Given the importance of *de novo* variation in rare disease diagnostics, especially for NDDs, it is also important to note the qualities of discrepant *de novo* calls between the two technologies. We found that most of the erroneously called *de novo* variants in the CCS data were correctly called as heterozygous in the proband but missed in the parents due to lower coverage and random sampling effects such that the variant haplotype was simply not covered by any reads in the transmitting parent. Such errors could be mitigated by sequencing parents more deeply. In contrast, *de novo* variants unique to IGS were enriched for systematic artifacts that cannot be corrected for with higher read-depth. Indels, for example, are a well-known source of error and heavily enriched among IGS de novo variant calls.

In one proband we identified a likely pathogenic, *de novo* L1-mediated insertion in *CDKL5*. *CDKL5* encodes cyclin-dependent kinase-like 5, a serine-threonine protein kinase that plays a role in neuronal morphology, possibly via regulation of microtubule dynamics[37]. Variation in *CDKL5* has been associated with EIEE2 (MIM: 300672), an X-linked dominant

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

syndrome characterized by infantile spasms, early-onset intractable epilepsy, hypotonia, and variable additional Rett-like features[38,39]. *CDKL5* is one of the most commonly implicated genes identified by ES/GS in epilepsy cases[40]. Single nucleotide variants (SNVs), small insertions and deletions, copy-number variants (CNVs) and balanced translocations have all been identified in affected individuals, each supporting a haploinsufficiency model of disease[41]. We also note that *de novo* SVs, including deletions and at least one translocation, have been reported with a breakpoint in intron 3, near the breakpoint identified here[41–44] (Supplemental Table 7, Supplemental Figure 8).

The variant harbors two classic marks of an L1HS insertion, including the preferred L1 EN consensus cleavage site (5'-TTTT/G-3'), and a 119-bp target-site duplication (TSD) which, in this case, includes exon 3 of *CDKL5*. Although TSDs are often fewer than 50 bp long, TSDs up to 323 bp have been detected[45]. The variant appears to be a chimeric L1 insertion, consisting of retrotransposition of an active L1HS mobile element, with 5' truncation, and duplication of a 2.6 kb intronic segment of *PPEF1*, which itself includes a partial L1 sequence, that lies about 230 kb downstream of *CDKL5*. This mechanism has been described previously, and has been proposed to result from a combination of retrotransposition and a synthesis-dependent strand annealing (SDSA)-like mechanism[45].

Using ACMG variant classification guidelines, we classified this variant as Likely Pathogenic. The variant was experimentally confirmed to result in frameshifted transcripts due to exon duplication, and was shown to be *de novo*, allowing for use of both the PVS1 (loss of function)[46] and PM2 (*de novo*)[47] evidence codes. Use of Likely Pathogenic, as opposed to Pathogenic, reflects the uncertainty resulting from the intrinsically unusual nature of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variant and its potential somatic mosaicism, in addition to the fact that its absence from population variant databases is not in principle a reliable indicator of true rarity. Identification of additional MEIs and other complex structural variants will likely aid in disease interpretation by both facilitating more accurate allele frequency estimation and by improving interpretation guidelines.

More generally, MEIs have been previously described as a pathogenic mechanism of gene disruption, but the contribution to developmental disorders has been limited to a modest number of cases in a few studies, each of which report P/LP variation lying within coding exons[9,10]. However, the MEI observed here in *CDKL5* would likely be missed by exome sequencing as the breakpoints are intronic, and in fact was also missed in our previous short-read genome sequencing analysis[15]. Global analyses of MEIs, such as our assessment of *de novo* Alu insertion rates (Supplemental Table 5), also support the conclusion that MEI events are far more effectively detected within CCS data compared to that seen in short read genomes. We find it likely that long-read sequencing will uncover MEIs that disrupt gene function and lead to NDDs in many currently unexplained cases.

CCS also led to detection of a complex, *de novo* rearrangement in proband 4. Complex chromosomal rearrangements leading to gene disruption have been reported in individuals with NDDs or other congenital anomalies[48–50]. The variant identified in Proband 4 appears to impact two genes. One of the affected genes, *DGKB,* encodes diacylglycerol kinase beta which appears to be somewhat tolerant to loss-of-function variation (pLI = 0.53, o/e = 0.22 (0.13 - 0.37)[51]; RVIS= 21.1%[52]). While some studies have suggested a role for this gene in neuronal spine formation[53], it is not clear whether this gene is relevant to NDDs. The second

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gene affected by rearrangement in Proband 4 is *MLLT3*, which is predicted to be moderately intolerant to loss-of-function variation (pLI = 1, o/e = 0 (0 - 0.13)[51]; RVIS = 21.1%[52]). *MLLT3*, also known as *AF9*, undergoes somatic translocation with the *MLL* gene, also known as *KMT2A* (MIM: 159555), in patients with acute leukemia; pathogenicity in these cases results from expression of an in-frame KMT2A-MLLT3 fusion protein and subsequent deregulation of target HOX genes[54]. Balanced translocations between chromosome 4 and chromosome 9, resulting in disruption of *MLLT3,* have been previously reported in two individuals, each with NDDs including intractable seizures[55,56]. Although proband 4 does not exhibit seizures, she does have features that overlap the described probands, including speech delay, hypotonia, and fifth-finger clinodactyly. In sum, there remains considerable uncertainty about the nature of this variant, as additional data will be required to fully resolve the structure of these loci in this proband. Further, the disease relevance of the two affected genes is unclear. Nevertheless, the fact that it occurred *de novo*, associates with a reduction in *MLLT3* expression, and overlaps with an SV previously reported in two NDD probands make it a highly intriguing VUS.

Here we describe an analysis of six NDD-affected probands using PacBio CCS. These data facilitated more comprehensive and accurate detection of variation across a spectrum of categories, including low-complexity repeats, mobile element insertions, and complex structural variation. Among these newly detected variants, we have identified one likely pathogenic variant and one VUS. While the sample size is far too small to facilitate precise estimates of future yields in NDDs, the fact that there exist two compelling hits from only six probands is consistent with the hypothesis that the ultimate yield among previously tested but unsolved NDDs is substantial. This is likely also true for individuals suspected to have rare

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

congenital disease more generally. Further, as CCS can capture complex variation in addition to essentially all variation detectable by short-read sequencing, it is likely that it will become a powerful front-line tool for research and clinical testing within rare disease genetics.

Acknowledgements

This work was supported by a grant from The National Human Genome Research Institute, UM1HG007301. Some reagents were provided by PacBio as part of an early-access testing program. We thank our colleagues at HudsonAlpha who provided advice and general support, including Amy Nesmith Cox, Greg Barsh, Kelly East, Whitley Kelley, David Bick, and Elaine Lyon, in addition to the HudsonAlpha Genomic Services Laboratory and Clinical Services Laboratory. We also thank the clinical team at North Alabama Children's Specialists. Finally, we are grateful to the families who participated in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Ropers HH. Genetics of intellectual disability. *Curr Opin Genet Dev*. 2008;18(3):241-250. doi:10.1016/j.gde.2008.07.008

2. Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet*. 2010;42(12):1109-1112. doi:10.1038/ng.712

3. Wellcome Sanger Institute DDD. Development Disorder Genotype - Phenotype Database. https://decipher.sanger.ac.uk/ddd#ddgenes.

4. Hiatt SM, Amaral MD, Bowling KM, et al. Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin Genet*. 2018;94(1):174-178. doi:10.1111/cge.13259

5. Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Med*. 2018;3(1). doi:10.1038/s41525-018-0053-8

6. 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. doi:10.1038/s41586-018-0566-4

7. McMurray CT. Expansions in simple DNA repeats underlie ~20 severe neuromuscular and neurodegenerative disorders. *Nat Publ Gr*. 2010;11(11):786-799. doi:10.1038/nrg2828

8. 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. doi:10.1136/jmedgenet-2014-102588

9. Torene RI, Galens K, Liu S, et al. Mobile element insertion detection in 89,874 clinical exomes. doi:10.1038/s41436-020

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

10. Gardner EJ, Prigmore E, Gallone G, et al. Contribution of retrotransposition to developmental disorders. doi:10.1038/s41467-019-12520-y

11. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: The long and the short of it. *Genome Biol*. 2019;20(1). doi:10.1186/s13059-019-1828-7

12. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*. 2019;10(MAY):426. doi:10.3389/fgene.2019.00426

13. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162. doi:10.1038/s41587-019-0217-9

14. 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. doi:10.1038/gim.2015.30

15. Bowling KM, Thompson ML, Amaral MD, et al. Genomic diagnosis for children with intellectual disability and/or developmental delay. *Genome Med*. 2017;9(1):43. doi:10.1186/s13073-017-0433-1

16. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333-i339. doi:10.1093/bioinformatics/bts378

17. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sequencing. *Genome Res*. 2011;21(6):974-984. doi:10.1101/gr.114876.110

18. Zhu M, Need AC, Han Y, et al. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet*. 2012;91(3):408-421. doi:10.1016/j.ajhg.2012.07.004

19. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

20. Lappalainen I, Lopez J, Skipper L, et al. dbVar and DGVa: public archives for genomic structural variation. doi:10.1093/nar/gks1213

21. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet*. 2014;46(10):1063-1071. doi:10.1038/ng.3092

22. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nat Genet*. 2011;43(9):838-846. doi:10.1038/ng.909

23. Gardner EJ, Lam VK, Harris DN, et al. The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res*. 2017;27(11):1916-1929. doi:10.1101/gr.218032.116

24. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867-2873. doi:10.1093/bioinformatics/btq559

25. Poplin R, Chang PC, Alexander D, et al. A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983. doi:10.1038/nbt.4235

26. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722-736. doi:10.1101/gr.215087.116

27. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. 2007;23(8):1026-1028. doi:10.1093/bioinformatics/btm039

28. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J Biol Chem*. 2020;295(13):4134-4170. doi:10.1074/jbc.REV119.007678

29. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. *Cancer Res*. 2017;77(21):e31-e34. doi:10.1158/0008-5472.CAN-17-0337

30. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res*. August 2018. doi:10.1093/nar/gky677

31. Zook JM, McDaniel J, Olson ND, et al. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol*. 2019;37(5):561-566. doi:10.1038/s41587-019-0074-6

32. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet*. 2014;46(9):944-950. doi:10.1038/ng.3050

33. Rousseau F, Rouillard P, Morel ML, Khandjian EW, Morgan K. Prevalence of carriers of premutation-size alleles of the FMR1 gene - and implications for the population genetics of the fragile X syndrome. *Am J Hum Genet*. 1995;57(5):1006-1018.

34. Xing J, Zhang Y, Han K, et al. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res*. 2009;19(9):1516-1526.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

doi:10.1101/gr.091827.109

35. Feusier J, Watkins WS, Thomas J, et al. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res*. 2019;29(10):1567-1577. doi:10.1101/gr.247965.118

36. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

37. Barbiero I, Peroni D, Siniscalchi P, et al. Pregnenolone and pregnenolone-methyl-ether rescue neuronal defects caused by dysfunctional CLIP170 in a neuronal model of CDKL5 Deficiency Disorder. *Neuropharmacology*. 2020;164. doi:10.1016/j.neuropharm.2019.107897

38. Bahi-Buisson N, Juliette Nectoux Ã, Haydee Ã, et al. Key clinical features to identify girls with CDKL5 mutations. 2008. doi:10.1093/brain/awn197

39. Kadam SD, Sullivan BJ, Goyal A, Blue ME, Smith-Hicks C. Rett syndrome and CDKL5 deficiency disorder: From bench to clinic. *Int J Mol Sci*. 2019;20(20). doi:10.3390/ijms20205098

40. Symonds JD, McTague A. Epilepsy and developmental disorders: Next generation sequencing in the clinic. *Eur J Paediatr Neurol*. 2020;24:15-23. doi:10.1016/j.ejpn.2019.12.008

41. Erez A, Patel AJ, Wang X, et al. Alu-specific microhomology-mediated deletions in CDKL5 in females with early-onset seizure disorder. *Neurogenetics*. 2009;10(4):363-369. doi:10.1007/s10048-009-0195-z

42. Bartnik M, Derwińska K, Gos M, et al. Early-onset seizures due to mosaic exonic deletions

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of CDKL5 in a male and two females. *Genet Med*. 2011;13(5):447-452. doi:10.1097/GIM.0b013e31820605f5

43. Cordova-Fletes C, Rademacher N, Muller I, et al. CDKL5 truncation due to a t(X;2)(p22.1;p25.3) in a girl with X-linked infantile spasm syndrome. *Clin Genet*. 2010;77(1):92-96. doi:10.1111/j.1399-0004.2009.01286.x

44. Sanchis-Juan A, Stephens J, French CE, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. 2018;10(1):95. doi:10.1186/s13073-018-0606-6

45. Gilbert N, Lutz S, Morrish TA, Moran J V. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Mol Cell Biol*. 2005;25(17):7780-7795. doi:10.1128/mcb.25.17.7780-7795.2005

46. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat*. 2018;39(11):1517-1524. doi:10.1002/humu.23626

47. Group SVIW. *ClinGen Sequence Variant Interpretation Recommendation for de Novo Criteria (PS2/PM6)-Version 1.0 Working Group Page: Https://Clinicalgenome.Org/Working-Groups/Sequence-Variant-Interpretation/ SVI Recommendation for De Novo Criteria (PS2 & PM6)-Versi*.; 2018. https://clinicalgenome.org/working-groups/sequence-variant-interpretation/. Accessed June 23, 2020.

48. Middelkamp S, Vlaar JM, Giltay J, et al. Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants. *Genome Med*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2019;11(1). doi:10.1186/s13073-019-0692-0

49. Plesser Duvdevani M, Pettersson M, Eisfeldt J, et al. Whole-genome sequencing reveals complex chromosome rearrangement disrupting <scp> *NIPBL* </scp> in infant with Cornelia de Lange syndrome. *Am J Med Genet Part A*. 2020;182(5):1143-1151. doi:10.1002/ajmg.a.61539

50. Lei M, Liang D, Yang Y, et al. Long-read DNA sequencing fully characterized chromothripsis in a patient with Langer–Giedion syndrome and Cornelia de Lange syndrome-4. *J Hum Genet*. April 2020:1-8. doi:10.1038/s10038-020-0754-6

51. Lek M, Karczewski KJ, Minikel E V, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057

52. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;9(8):e1003709. doi:10.1371/journal.pgen.1003709

53. Hozumi Y, Kakefuda K, Yamasaki M, Watanabe M, Hara H, Goto K. Involvement of diacylglycerol kinase β in the spine formation at distal dendrites of striatal medium spiny neurons. *Brain Res*. 2015;1594:36-45. doi:10.1016/j.brainres.2014.11.012

54. Krivtsov A V., Armstrong SA. MLL translocations, histone modifications and leukaemia stem-cell development. *Nat Rev Cancer*. 2007;7(11):823-833. doi:10.1038/nrc2253

55. Pramparo T, Grosso S, Messa J, et al. Loss-of-function mutation of the AF9/MLLT3 gene in a girl with neuromotor development delay, cerebellar ataxia, and epilepsy. *Hum Genet*. 2005;118(1):76-81. doi:10.1007/s00439-005-0004-1

56. Striano P, Elia M, Castiglia L, Galesi O, Pelligra S, Striano S. A t(4;9)(q34;p22)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Translocation Associated with Partial Epilepsy, Mental Retardation, and Dysmorphism.

Epilepsia. 2005;46(8):1322-1324. doi:10.1111/j.1528-1167.2005.64304.x

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure Legends

Figure 1. Proband 6 has a *de novo* insertion resulting in duplication of exon 3 of *CDKL5*. A. Alignment of CCS reads near exon 3 of *CDKL5* in IGV in Proband 6 and her parents. Unaligned portions of reads on either end of the 119 bp duplicated region are indicated with black triangles. The location of hard-clipped bases are designated with a black diamond. **B.** Gene structure of *CDKL5*, *RS1*, and *PPEF1*, indicating the location of the 6993 bp insertion in *CDKL5* and location of the duplicated *PPEF1* intronic sequence (red). **C.** Zoomed in view of the insertion. Black boxes indicate exons, gray boxes indicate the duplicated 119 bp segment, blue bar indicates a partial L1HS retrotransposon, and red indicates the duplicated *PPEF1* intronic sequence. Green boxes indicate RepeatMasker annotation of the proband's insertion-bearing, contig sequence.

Figure 2. The duplicated *CDKL5* exon 3 is present in a subset of the proband's *CDKL5* transcripts. A. RT-PCR using primers specific to exons 2-5 of *CDKL5* cDNA results in a 240 bp amplicon in proband (P), Dad (D), and Mom (M). An additional 275 bp amplicon is present only in the proband (asterisk). **B.** Sanger sequencing of both amplicons from the proband confirmed that the 240 bp amplicon includes the normal, expected sequencing and inclusion of a duplicated exon 3 in the upper, 275 bp band. This is predicted to lead to a frameshift (red circle) and downstream stop, p.(Thr35ProfsTer52). Yellow outlined box, exon 3 sequence; orange outlined box, duplicated exon 3 sequence.

Figure 3. CCS reads support three breakpoints within *DGKB* and *MLLT3*. A. Schematic of *DGKB* and *MLLT3* showing exons near the three observed breakpoints (red triangles). Potentially rearranged exons are shown as gray boxes. Red numbers are coordinates on the chromosome 7

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.02.185447; this version posted July 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

reference, while red letters indicate which panels show reads supporting the breakpoint. **B-D.** Black text at the top of each panel indicates the genomic location shown, and smaller gray text indicates the alternate location where unaligned portions of the reads (multicolored) align.

Figure 4. *MLLT3* shows decreased expression in Proband 4 (P4). qRT-PCR using TaqMan probes targeting the *MLLT3* exon 3-4 (**A, C**) and exon 9-10 (**B, C**) splice junctions, normalized to either the median of Family 3 values (**A, B**), or the median of Family 4 Parent values (**C, D**). Samples include Proband (P), Dad (D), and Mom (M) from two different trios in this study, 3 and 4.

Table 1. Probands selected for PacBio sequencing.

Family ID	Proband Gender	Race	Major Phenotypic Features	Previous Genetic Testing				PacBio CCS Coverage (P/D/M)
				Array	Single Gene Test(s) or Panel(s)[a]	ES/GS	Other	
1	F	C	Seizures, facial dysmorphism, hypotonia	Normal	Normal x2	No Findings (both)	Karyotype - normal	25x/10x/11x
2	F	AA	ID, seizures, hypotonia	Normal	Normal x7	No Findings (both)	Mito - normal	26x/16x/12x
3	M	C	ID, seizures	VUS dup	Normal x3	No Findings (GS)	Fragile X - normal	35x/19x/22x
4	F	C/AA	ID, facial dysmorphism, hypotonia	Normal	Normal x1	No Findings (GS)	Fragile X - normal	29x/14x/20x
5	M	C	ID, seizures, speech delay, brain MRI abnormalities	Normal	Normal x4	No Findings (GS)	Mito - normal	30x/16x/20x
6	F	C	ID, seizures, speech delay	Normal	NP	No Findings (GS)	NP	33x/19x/14x

ES/GS, exome sequencing/genome sequencing; P, proband; D, dad; M, mom; F, female; M, male; C, Caucasian; AA, African American; ID, intellectual disability; NP, not performed. [a] Some VUS SNVs have been reported in these probands.

available under aCC-BY-ND 4.0 International license.

A

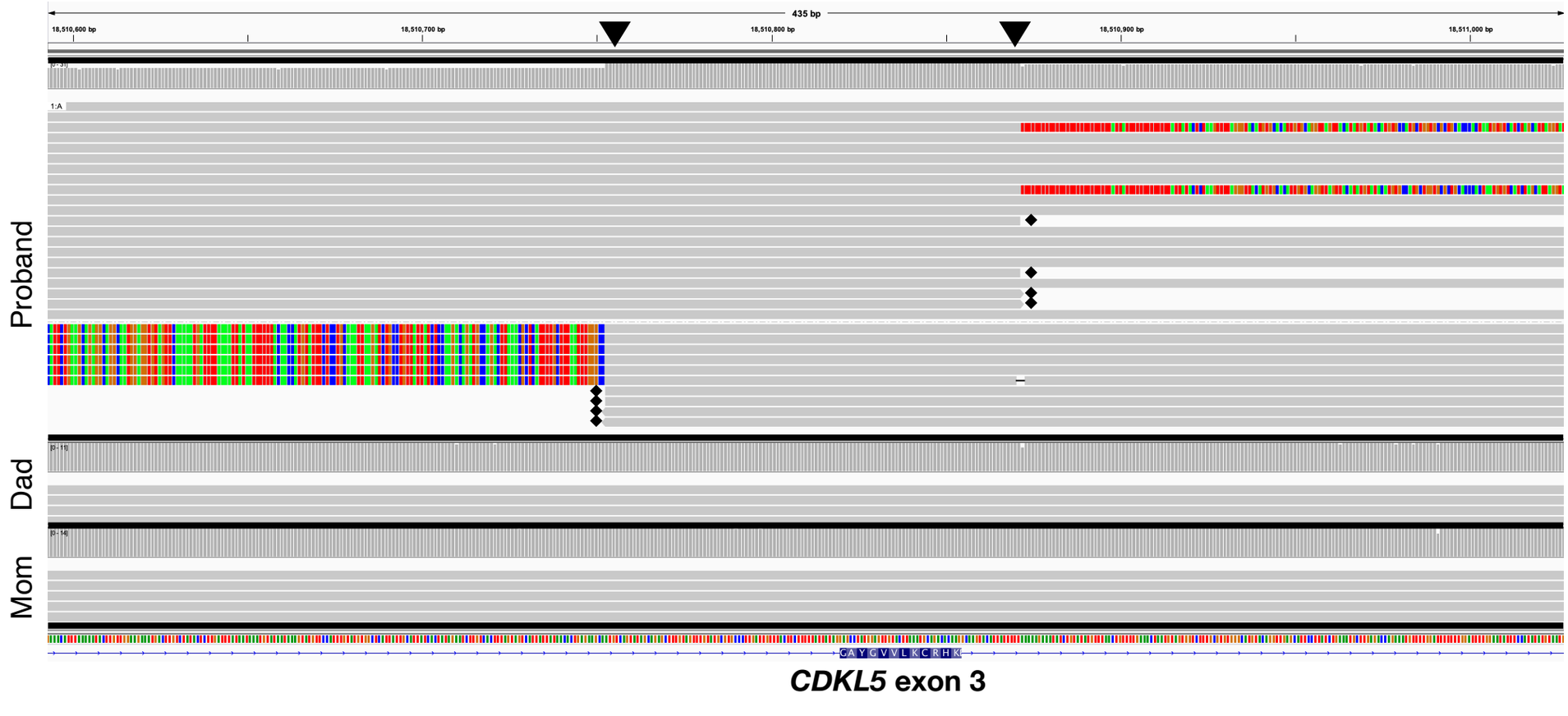
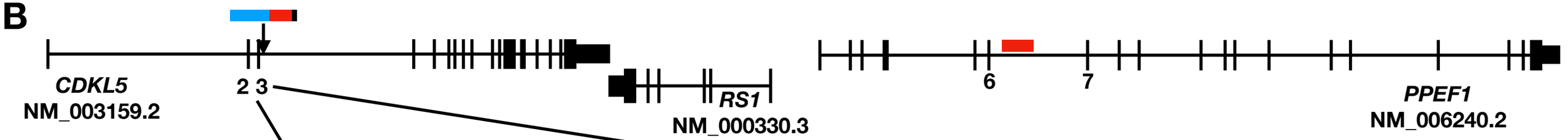
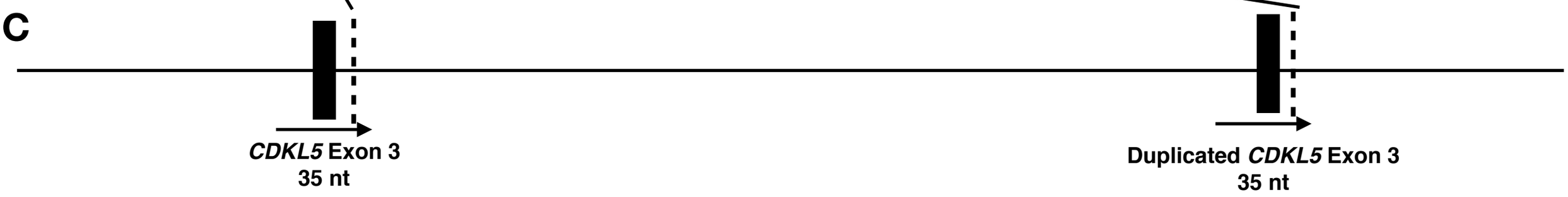


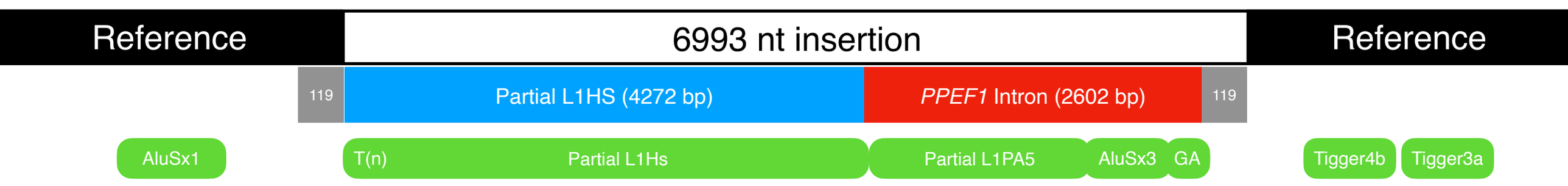
Figure 1

B



C





available under aCC-BY-ND 4.0 International license.

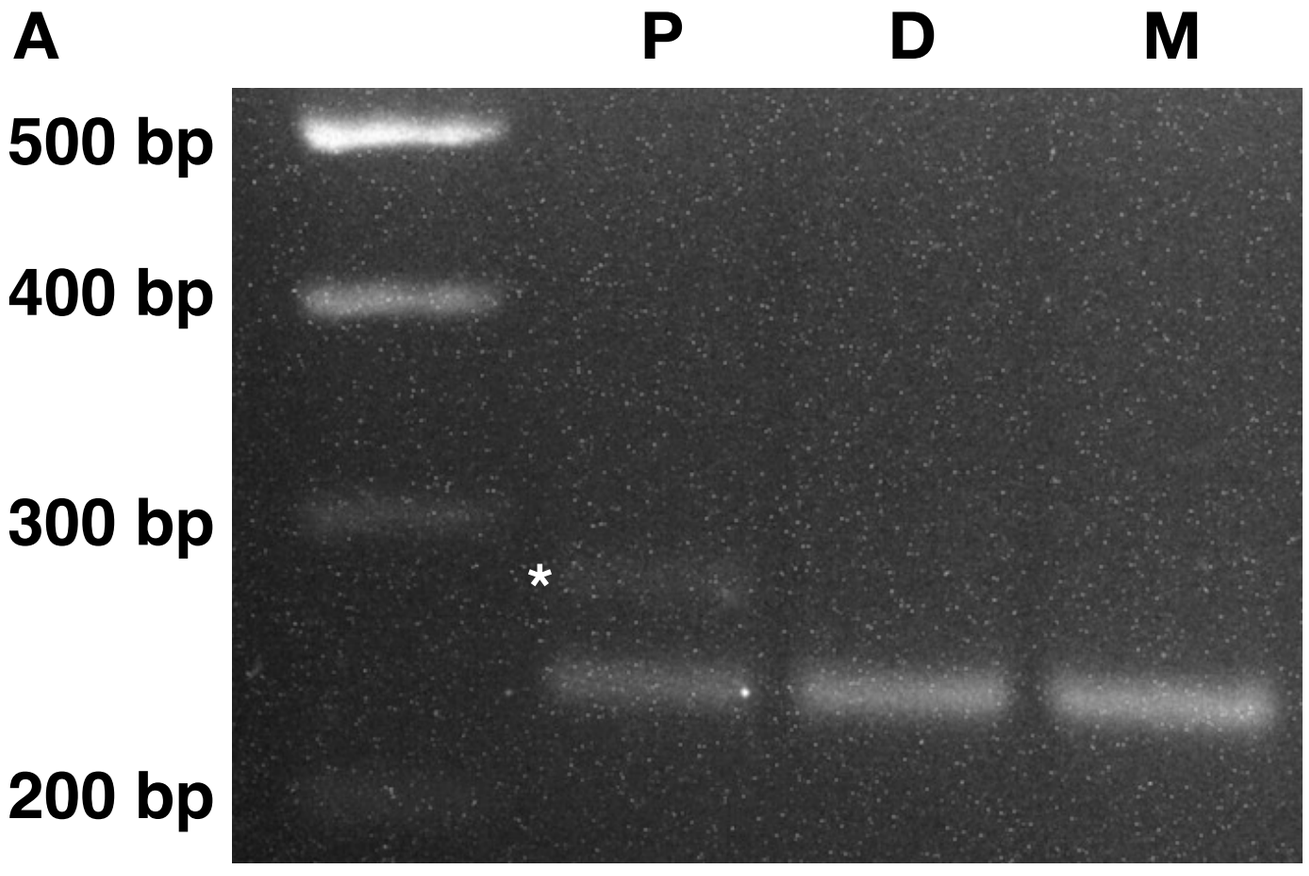
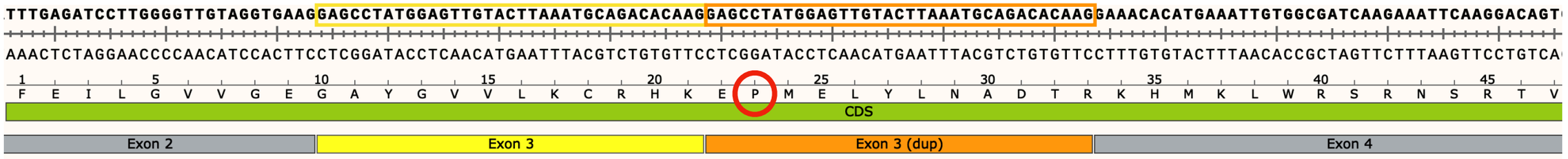


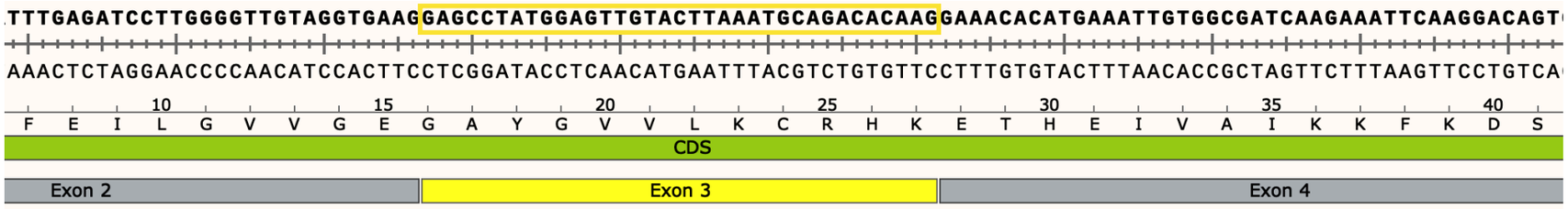
Figure 2

B

275 nt Amplicon



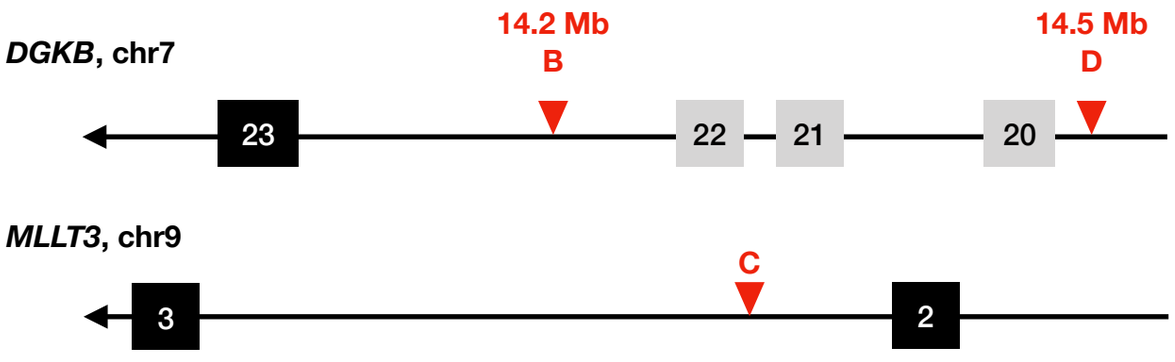
240 nt Amplicon



available under aCC-BY-ND 4.0 International license.

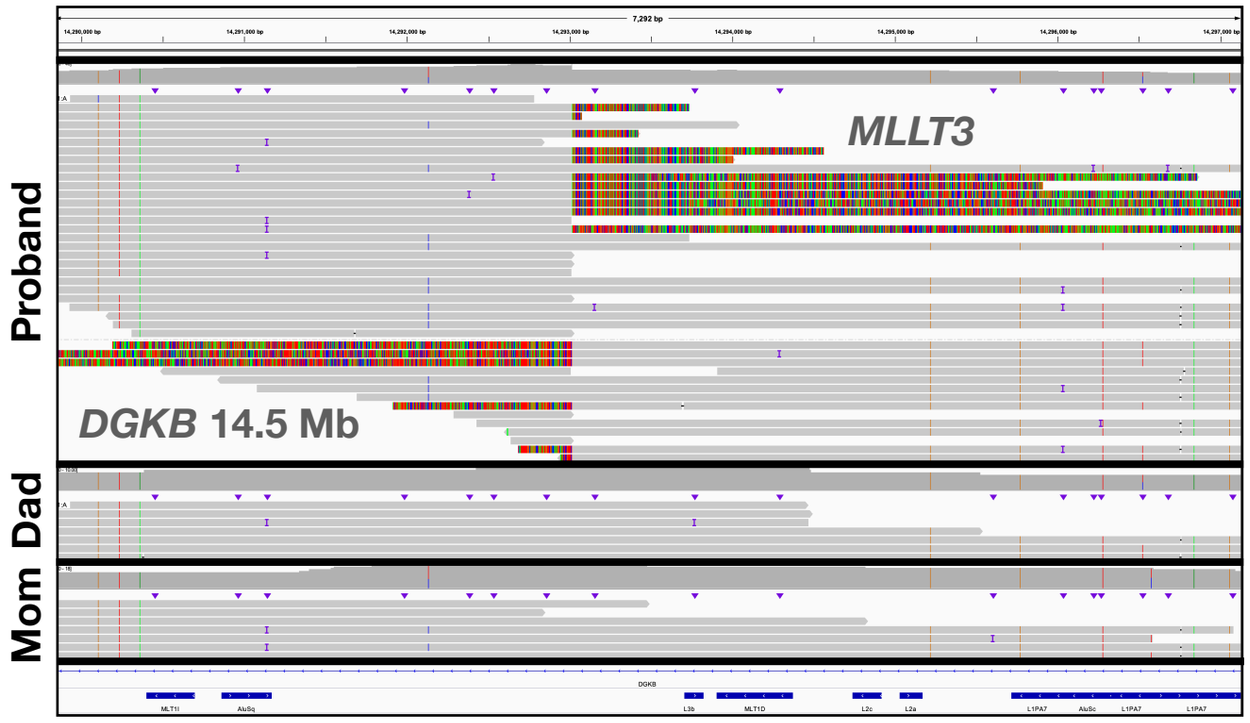
Figure 3

A



B *DGKB* 14.2 Mb

Chr7:14,289,831-14,297,152

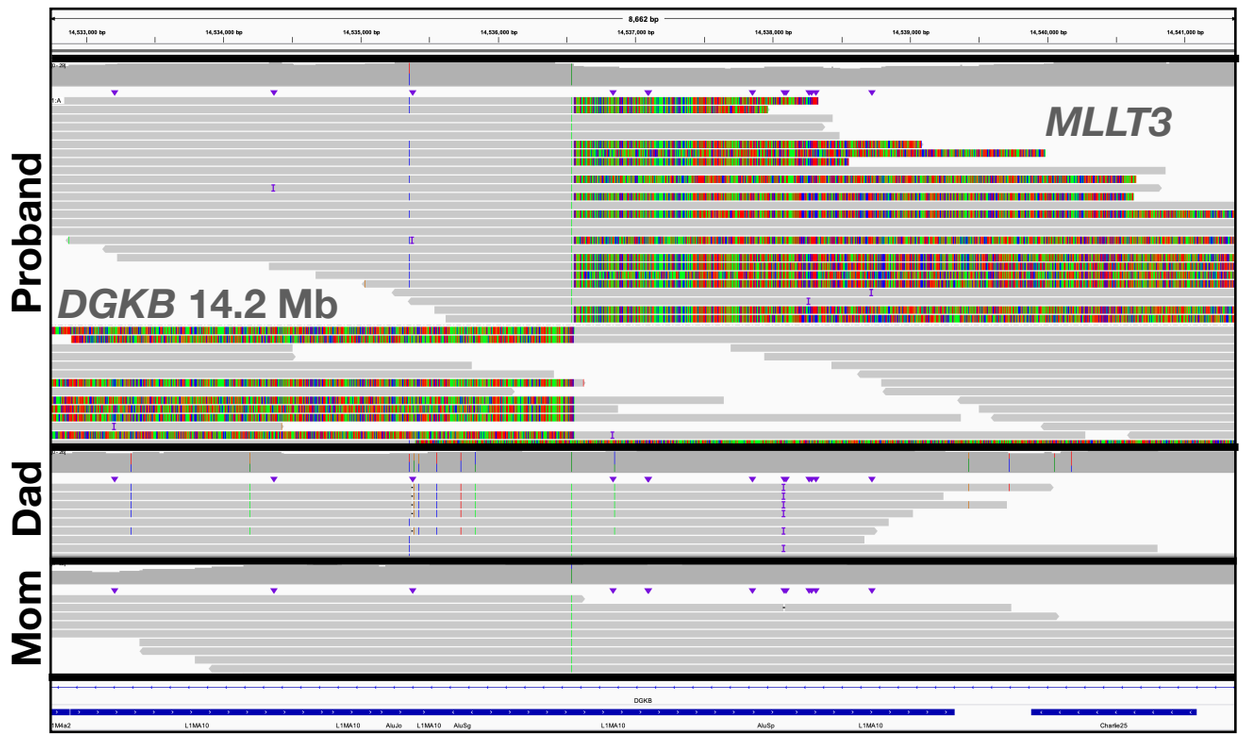


C *MLLT3*

chr9:20,584,500-20,592,609

D *DGKB* 14.5 Mb

chr7:14,532,724-14,541,421



available under aCC-BY-ND 4.0 International license.

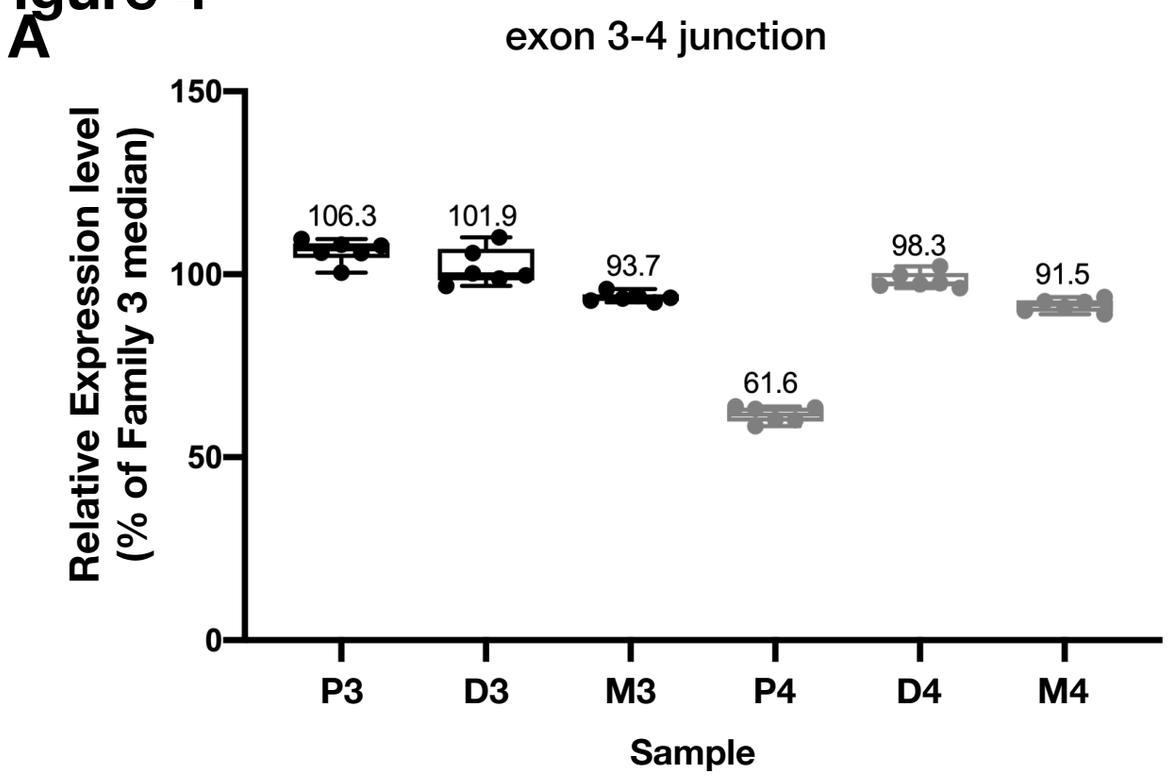
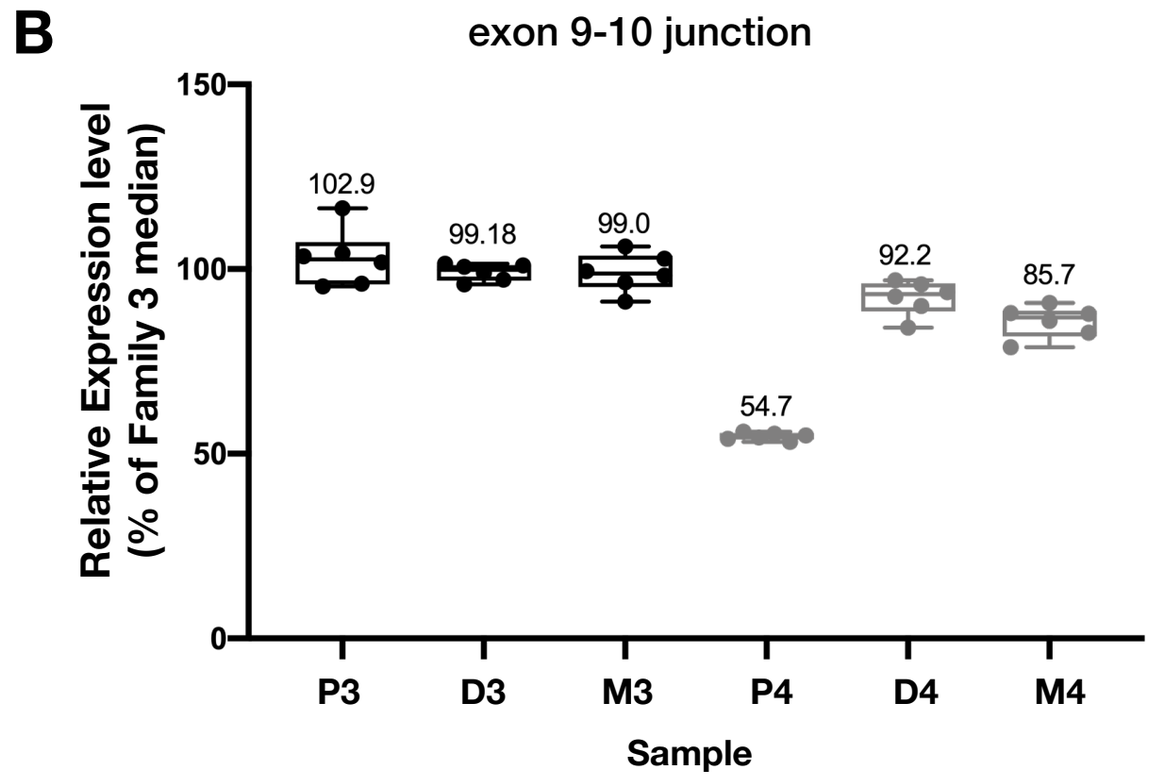
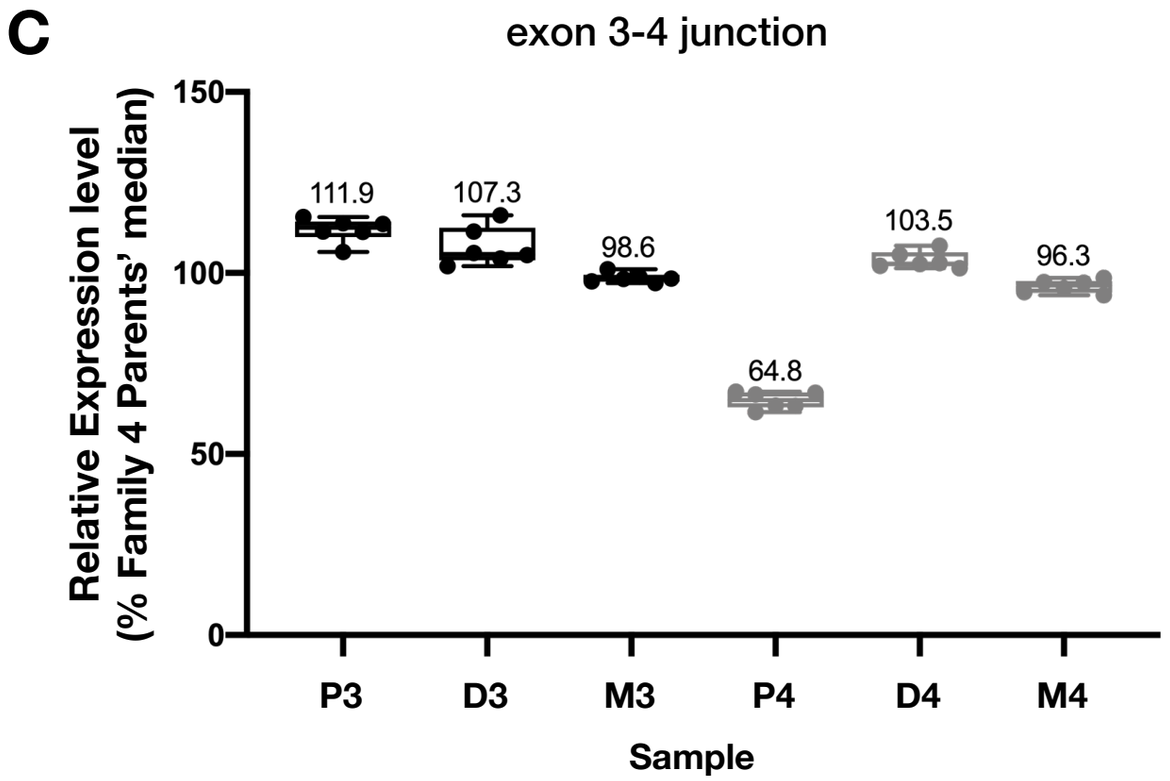
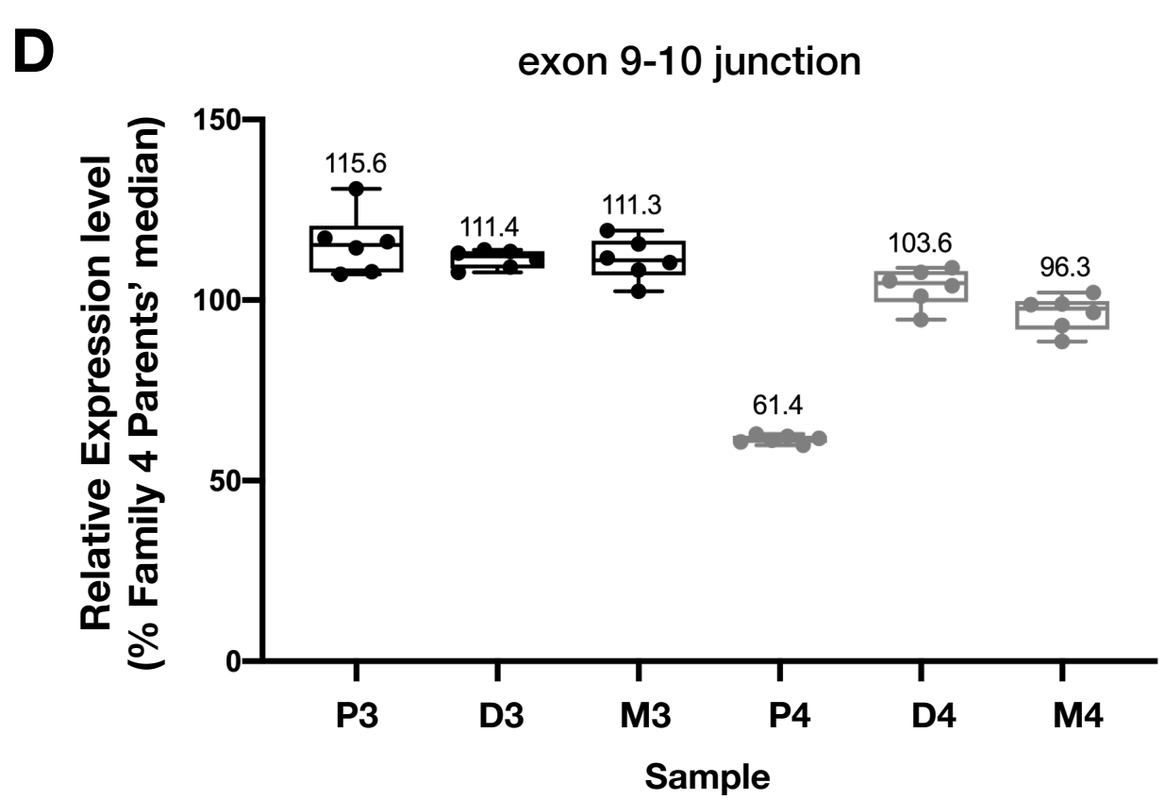



Figure 4