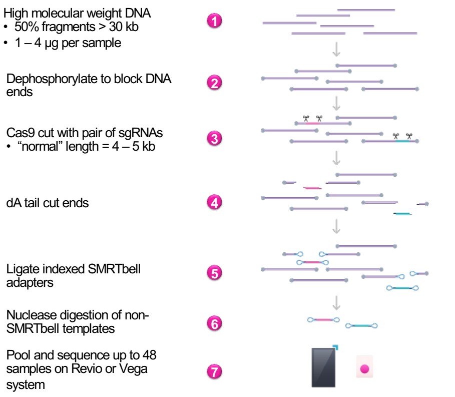
# Targeted long-read sequencing of native DNA for genetic disease diagnostic and screening research

Jocelyne Bruand, Sarah B. Kingan, Jeff Zhou, Davy Lee, Heather Ferrao, Ian McLaughlin, Sijie Wei, Richa Pathak, Ravi Dalal, Tom Mokveld, Guilherme De Sena Brandine, Egor Dolzhenko, Nat Echols, Michael A. Eberle, Duncan Kilburn
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

## Introduction

Short tandem repeats (STRs) are DNA sequences composed of repetitions of 2 – 6 bp motifs. Expansions of STRs are the cause of over 60 monogenic diseases, including Huntington's disease, fragile X syndrome, and amyotrophic lateral sclerosis[1]. In addition to their length, the pathogenicity of these STRs can be impacted by sequence composition, methylation status and mosaicism. One such example is the FMR1 repeat whose CGG repeat expansions are typically hypermethylated and where AGG interruption sequences can stabilize the repeat. Detecting all the characteristics associated with pathogenic repeat expansions traditionally required multiple assays, however high-accuracy long-read sequencing of unamplified DNA can resolve all these features in a single assay.

## Scalable amplification-free workflow



1. High molecular weight DNA
   - 50% fragments > 30 kb
   - 1 – 4 µg per sample
2. Dephosphorylate to block DNA ends
3. Cas9 cut with pair of sgRNAs
   - "normal" length = 4 – 5 kb
4. dA tail cut ends
5. Ligate indexed SMRTbell adapters
6. Nuclease digestion of non-SMRTbell templates
7. Pool and sequence up to 48 samples on Revio or Vega system

**Figure 1.** PureTarget repeat expansion panel is a robust amplification-free approach to generate long-read HiFi sequencing libraries containing loci associated with 20 pathogenic STR expansions. Starting with high molecular weight DNA from blood or cell line extracted with Nanobind PanDNA kit, the workflow employs Cas9 and a single pair of guide RNAs to target each repeat region and ~3-5 kb of flanking sequence. Comprehensive genotyping of consensus repeat size, motif analysis and methylation is performed with Tandem Repeat Genotyping Tool (TRGT)[2] in SMRT Link software.

| Gene(s) | Associated disease |
|---|---|
| *ATN1, ATXN1, ATXN2, ATXN3, ATXN7, ATXN8, ATXN10, CACNA1A, PPP2R2B, TBP* | Spinocerebellar ataxia |
| *FMR1* | Fragile X-associated disorders |
| *C9orf72* | Amyotrophic lateral sclerosis and Frontotemporal dementia |
| *DMPK, CNBP* | Myotonic dystrophy (DM1, DM2) |
| *FXN* | Friedreich's ataxia |
| *RFC1* | CANVAS |
| *HTT* | Huntington's disease |
| *AR* | Spinal-bulbar muscular atrophy |
| *PABPN1* | Oculopharyngeal muscular dystrophy |
| *TCF4* | Fuchs endothelial corneal dystrophy |

**Table 1.** PureTarget Repeat Expansion Panel Content. The PureTarget product line includes a panel of 20 repeat expansion loci relevant to neurodegenerative disease.

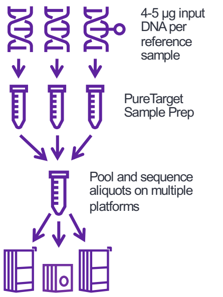
## References and Resources

1. Leitão, E., et al. (2024). Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. Rev Neurol (Paris). 180(5):383-392. doi: 10.1016/j.neurol.2024.03.005.
2. Dolzhenko, E., et al. (2024). Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol. 2024 doi: 10.1038/s41587-023-02057-3.

PureTarget website

## Datasets description

To assess the accuracy of PureTarget across sequencing platforms and chemistries, we sequenced panels of reference samples with validated pathogenic expansions. By including the same sample across multiple panels, and sequencing each pool across multiple platforms, we allow for both technical replicates and robust comparison across the sequencing technologies.



4-5 µg input DNA per reference sample

PureTarget Sample Prep

Pool and sequence aliquots on multiple platforms

**Figure 2 (left).** Diagram of the experimental set up. We prepared 4-5 µg of input DNA with PureTarget for each sample. Libraries were pooled and sequenced on multiple technologies to allow for comparison of both technical and sequencing replicates.

**Table 2 (below).** List of panels generated for these results. Sample pools were composed of 16 or 24 samples, with potential technical replicates within a panel. As shows on Figure 2, pooled libraries were aliquoted across sequencers. We aliquoted such that Revio v1 and Vega received 2µg input DNA per sample equivalent, while Revio system with SPRQ chemistry received 1 µg of input DNA, thanks to ability to load lower input samples.

| Panel | Plex | Vega | Revio v1 | Revio SPRQ |
|---|---|---|---|---|
| Coriell A | 24* | 2µg | 2µg | 1µg |
| Coriell B | 24 | 2µg | 2µg | N/A |
| Coriell C | 16 | 2µg | 2µg | 1µg |

*Two samples from Coriell A are excluded from this report due to insufficient input material (< 2µg instead of 5µg).

## Sample coverage across platforms

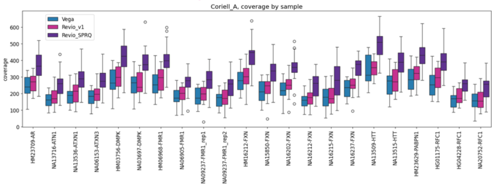


**Figure 3**. Target coverage by sample for reference Coriell samples with known repeat expansions (N=22) across different sequencing technologies. Target coverage is similar between Vega and Revio v1, while Revio SPRQ has higher coverage.
Full dataset available at https://downloads.pacbcloud.com/public/2024Q4/Vega/PureTargetCoriell24/

## Complex motif and mosaicism at HTT

### A. Genotypes

**Expected genotype:**
Short allele – (CAG)15
Long allele – (CAG)70

**Observed genotype:**
Short allele – (CAG)15(CAA)(CAG)(CCG)(CCA)(CCG)10
Long allele – (CAG)73(CAA)(CAG)(CCG)(CCA)(CCG)7

### B. Coverage and mosaicism

**Vega Coverage**
Short allele – 74-fold
Long allele – 95-fold

**Revio Coverage**
Short allele – 248-fold
Long allele – 247-fold

**Vega Mosaicism**
Short allele - 83-91
Long allele - 237-274

**Revio Mosaicism**
Short allele – 84-97
Long allele - 234-282

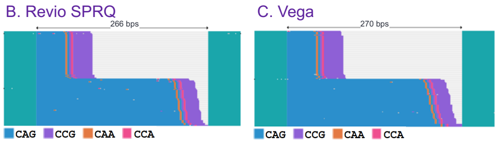


**Figure 4**. Six technical replicates of sample NA13509, from an individual clinically affected by Huntington's disease, were sequenced giving consistent genotypes (A) across platforms. Waterfall plots down sampled to 50 reads per allele for Revio SPRQ (C) and Vega (D) reveal high depth and low level of mosaicism in the long allele, ranging from ~234 – 282 bps on Revio (B).

## Mosaicism at *FXN*

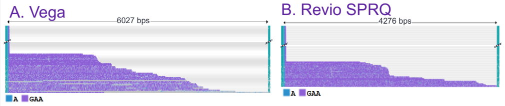


**Figure 5**. Waterfall plots for NA16237 at the FXN locus, sequenced on Vega (A) and Revio SPRQ (B). The short allele does not show mosaicism and was partially cropped from the waterfall plots. The long allele shows high mosaicism with alleles ranging from ~2000 to ~5000 bps. Coverage (long/short) for Revio SPRQ was 200 / 29 reads and for Vega was 131 / 37.

## Expanded alleles in technical replicates

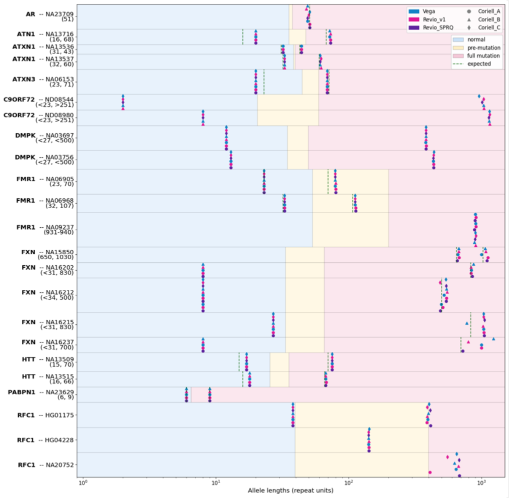
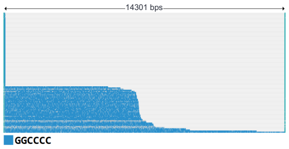


**Figure 6**. Swimlane plot showing the genotype calls for replicates of the reference samples. Samples with 3 or more replicates are shown. Color indicates the sequencing platform, while shape represents the technical replicate. Points with the same shape come from the same original aliquot. Allele length is reported as the number of pathogenic repeat units. When available, the expected genotype is specified on the y-axis label. Background colors indicate whether the length is considered normal, pre-mutation or full mutation[1].

## Genotyping at *C9orf72* on Vega



| Sample | Panel | Alleles (bps) | # Reads |
|---|---|---|---|
| ND08544 | Coriell B | 12 / 6196 | 71 / 54 |
| ND08544 | Coriell C | 12 / 5715 | 64 / 40 |
| ND08980 | Coriell B | 48 / 6831 | 70 / 41 |
| ND08980 | Coriell C | 48 / 6865 | 65 / 41 |

**Figure 7**. Genotyping calls and read coverage for C9orf72 expansions on Vega. Technical replicates were sequenced on Vega for two samples with known C9orf72 expansions, ND08544 and ND08980. For both samples, we see consistent coverage of the expanded allele, which contains repeats > 5000 bps for ND08544 and > 6500 bps for ND08980. The long allele shows mosaicism. Read coverage was ≥ 40 for both allele on all samples and replicates.

## Methylation and AGG motif at *FMR1*

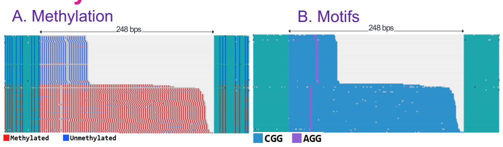


**Figure 8**. *FMR1* waterfall plots on Vega. Coverage is downsampled to 100 reads. When representing methylation (A), we note that the short allele is unmethylated, while the long allele is hypermethylated. Looking at the motif sequence, we can see that both alleles contain the protective AGG interruptive motif.

## Conclusions

- PureTarget is a **complete solution** to accurately characterize **lengths**, **repeat sequence** and **methylation** status **of repeat expansions relevant for human disease**.
- The PureTarget repeat expansion panel protocol and analysis in SMRT Link can deliver **sample to answer in 3 days**.
- PureTarget delivers **consistent results across technical replicates and HiFi sequencing platforms** for reference samples with known repeat expansions.