Application note

Comprehensive genotyping with the PureTarget repeat expansion panel and HiFi sequencing

Introduction

Tandem repeats are regions of the genome consisting of repetitive units of specific DNA sequences. These regions are hypermutable: they can increase in length across generations and may have variable lengths within somatic tissues of an individual (Paulson 2018). Many tandem repeats become pathogenic when they exceed a length threshold, which varies from gene to gene, resulting in mutations called repeat expansions (REs, Ibañez et al., 2022). Repeat expansions have been linked to dozens of diseases and cancer, most notably neuromuscular disorders like Huntington's disease, Fragile-X disorder, spinocerebellar ataxia, and myotonic dystrophies (Depienne & Mandel, 2021). Disease severity and age of onset of these conditions are often associated with their repeat length (Ibañez et al., 2022).

Though common, these regions are challenging to characterize and as such, the majority of patients with rare neurological diseases remain undiagnosed (Ibañez et al., 2022). Repeat expansions have historically been profiled with Southern blotting and PCR-based assays (Tarleton, 2003), and more recently short-read bioinformatics methods (Dolzhenko et al., 2020; Dashnow et al., 2022), but these methods are limited by throughput and read length, respectively. Given that many repeat expansions are longer than the typical length of short reads, comprehensive genotyping of REs requires a high-throughput long-read sequencing approach that can reliably manage the high structural variability of these regions. Powered by the exceptional accuracy of HiFi sequencing and the Tandem Repeat Genotyping Tool (TRGT), the PacBio® PureTarget™ repeat expansion panel offers more comprehensive genotyping for 20 of the most important repeat expansions for human health. This application note demonstrates the performance of the PureTarget repeat expansion panel, and presents the PureTarget panel as a scalable, more comprehensive solution for profiling repeat expansions, compared to legacy genotyping and next-generation sequencing methods.

Scalable, convenient, and more comprehensive genotyping

The PureTarget repeat expansion panel enables genotyping of critical pathogenic repeat expansion loci at scale. Up to 48 samples can be sequenced per Revio™ SMRT® Cell and up to 24 samples per Sequel® II and IIe SMRT Cell for the fixed 20 gene panel. Library preparation can be completed in 1 day using between 1 and 4 µg of starting DNA per sample. With 24-hour movies on the Revio system and automated analysis in SMRT® Link, users can go from sample to answer in 3 days. A single Revio system can process about 70,000 samples per year with full system utilization.

The kit includes a panel of 20 repeat expansions loci (see table 1 for target list) with the panel capturing ~2 kb upstream and downstream of the repeat. For normal alleles, resulting sequences are 4–5 kb in length but reads for expanded alleles may be longer. Sequencing of samples with large expansions indicate that it is possible to span repeats up to 35 kb in length in a single read.

Gene	Disease
ATN1, ATXN1, ATXN2, ATXN3, ATXN7, ATXN8, ATXN10, CACNA1A, PPP2R2B, TBP	Spinocerebellar Ataxias
FXN	Friedreich Ataxia
FMR1	Fragile X-related Disorders (FXDs)
HTT	Huntington's Disease
RFC1	CANVAS Syndrome
DMPK, CNBP	Myotonic Dystrophy (DM1, DM2)
C9orf72	Amyotrophic lateral sclerosis (ALS) Frontotemporal dementia (FTD)
AR	Spinal bulbar muscular atrophy
PABPN1	Oculopharyngeal Muscular Dystrophy
TCF4	Fuchs Endothelial Corneal Dystrophy

Table 1. 20 genes included in the PureTarget repeat expansion panel.

HiFi data from PureTarget libraries can be analyzed with the Tandem Repeat Genotyping Tool (TRGT) to deliver comprehensive genotypes, including:

- Accurate sizing of both alleles, including large expansions (Figure 1).
- Single-base resolution of repeat sequences (Figures 2 and 3).
- Simultaneous detection of methylation (Figure 4).
- Deep coverage to profile mosaicism (Figure 5).

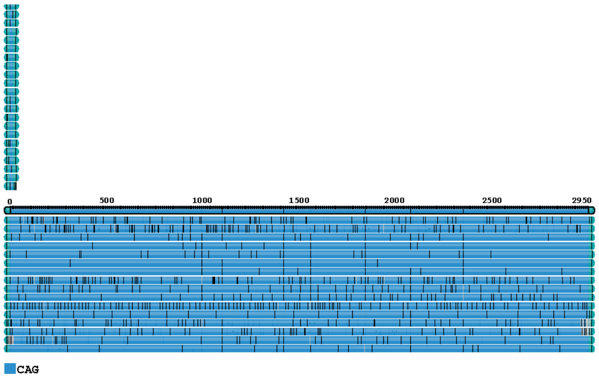

Figure 1. Accurate sizing of expanded *DMPK* allele with 2950 repeat motifs (8,850 bp).

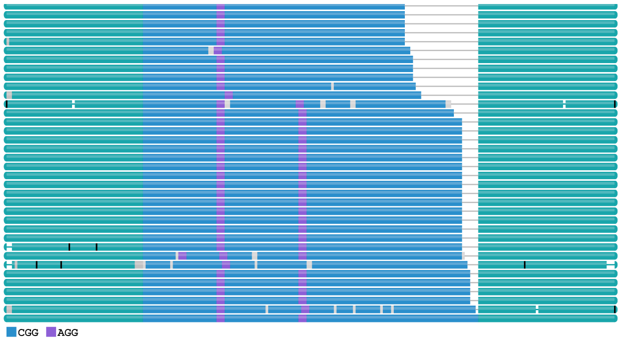

Figure 2. AGG interruption sequences in *FMR1* repeat.

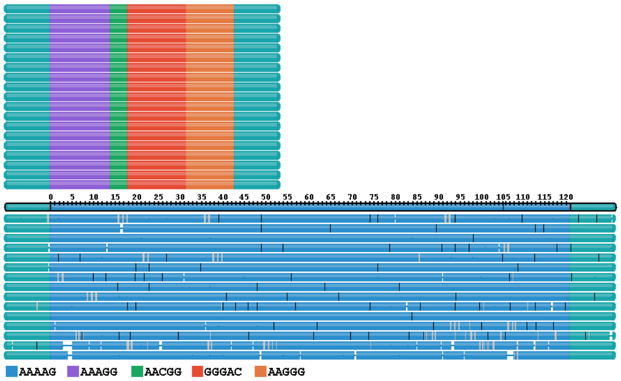

Figure 3. Single-base resolution of a complex *RFC1*including pathogenic "AAGGG" repeat motif (orange).

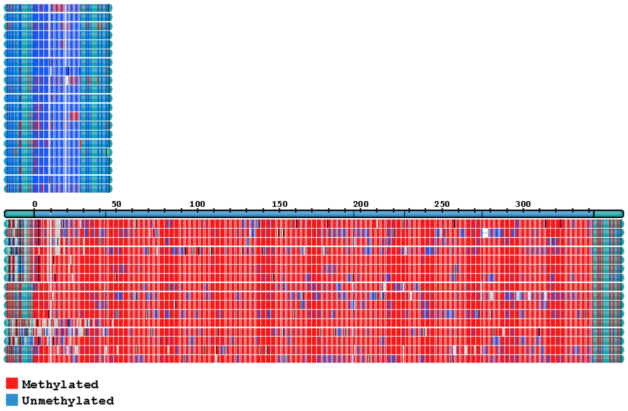

Figure 4. Profile at *FMR1* shows consistent methylation of expanded allele in female carrier (NA07537).

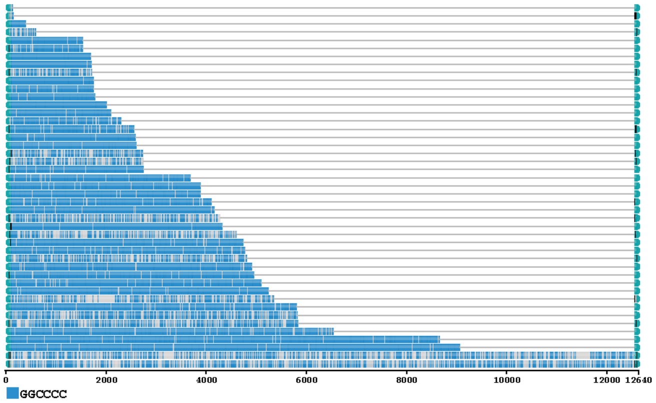

Figure 5. Deep coverage captures distribution of repeat lengths within a sample at *C9orf72*.

Comparison to other genotyping methods

PureTarget can deliver a more comprehensive and scalable solution for repeat expansion genotyping than other methods. Traditional methods like PCR and fragment length analysis, repeat-primed PCR, southern blots, and newer kits from Asuragen (AmplideX product line) only interrogate a single target at a time. For studies of ataxia, for example, multiple targets must be genotyped, which leads to long turn-around times, high labor cost, and operational complexity.

Sequencing methods like Illumina whole genome sequencing (WGS) or ReadUntil from Oxford Nanopore allow the simultaneous interrogation of multiple loci but each has disadvantages compared to PureTarget. For Illumina WGS, short reads limit the length of a repeat that can be genotyped accurately and as a result, underestimate repeat length at important targets like *C9orf72*, *DMPK*, or *FXN* (Ibañez et al., 2022). For *FMR1*, false positive calls may result from the inability to distinguish fully expanded *FMR1* alleles from pre-mutations (Ibañez et al., 2022). The Oxford Nanopore ReadUntil method can size long expansions and has flexible panel content. However, current implementations show low enrichment rates and high ff-target rates, making the path to scaling up to large sample numbers unclear (Stevanoski et al., 2022).

PureTarget performance and product specs

Optimal performance is obtained using PacBio Nanobind® extraction kits with human blood or cell line samples as the officially supported sample type for the kit. See "Sample types and DNA extraction methods" section below for more details on recommended DNA quality, quantity, sample types, and extraction kits. Typical target coverages for different sample multiplex levels on the Sequel IIe and Revio systems are shown in Figure 6. Complete product specs are shown in table 2.

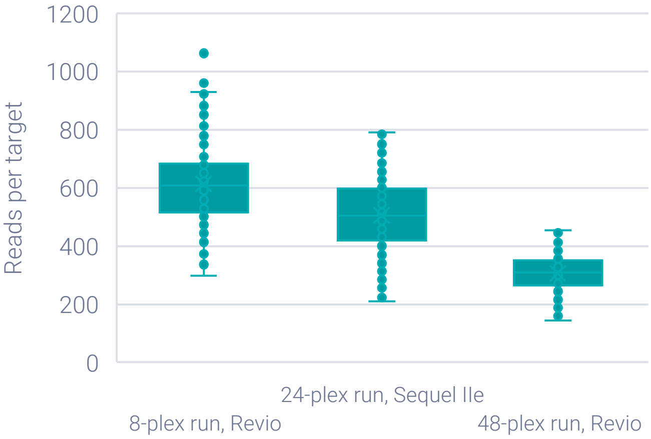

Figure 6. Deep coverage in multiplexed samples across 20 panel targets. Samples were prepared from 2.0–2.5 µg of high molecular weight (HMW) DNA extracted with Nanobind PanDNA from whole blood.

Spec	Metric
DNA input[1]	2 µg/sample
DNA quality[2]	GQN at 30 kb >5
Mean sample coverage[3]	>200-fold
Minimum sample coverage[3]	50-fold
Sample multiplexing[4]	Up to 48 - Revio System Up to 24 - Sequel II Systems
Library size[5]	4–5 kb
Methylation[6]	Detected

Table 2. PureTarget repeat expansion panel product specs.

Target enrichment and library prep

PureTarget libraries use CRISPR-Cas9 for target enrichment (Tsai et al., 2018, Tsai et al., 2022). The advantages of this approach over other methods that use PCR include:

- Retention of methylation signal in the resulting library molecules
- Less size-bias leading to better coverage in expanded alleles
- Coverage in regions of high GC%
- Absence of replication errors and other artifacts

The PureTarget repeat expansion kit contains reagents needed for target enrichment and SMRTbell® library prep, including the pre-mixed pool of guide RNAs for DNA digestion with Cas9 for the 20-target panel. Barcoded adapters can be purchased with the SMRTbell adapter index plate 96A. Typical turnaround time for library prep is less than 8 hours with average hands-on time of 3.5 hours.

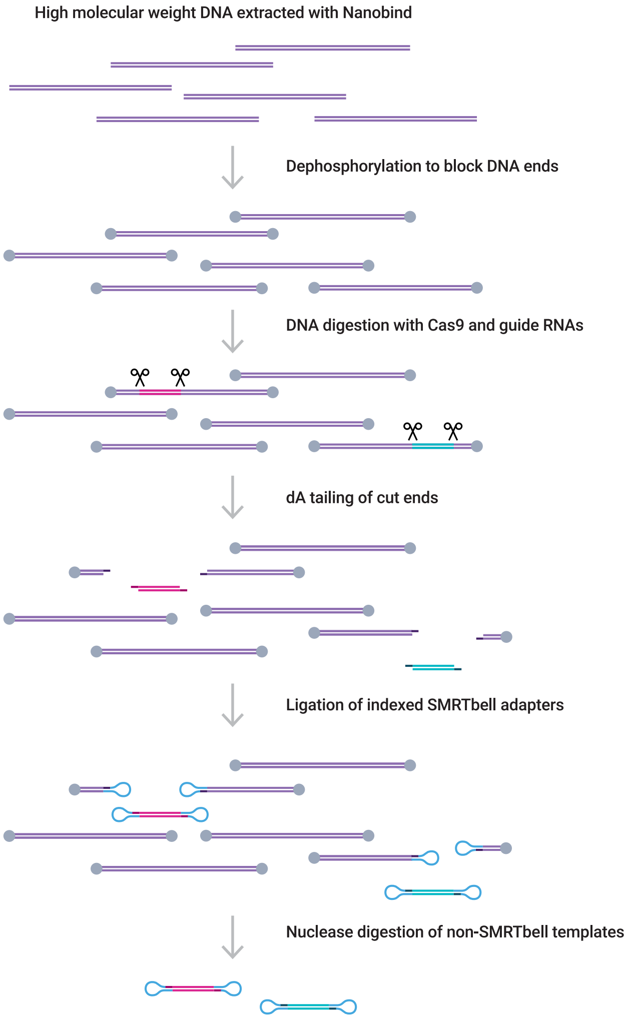
High molecular weight DNA extracted with Nanobind
Dephosphorylation to block DNA ends
DNA digestion with Cas9 and guide RNAs
dA tailing of cut ends
Ligation of indexed SMRTbell adapters
Nuclease digestion of non-SMRTbell templates

Figure 7. Target enrichment with CRISPR-Cas9.

Figure 7 illustrates the major steps of target enrichment with CRISPR-Cas9. Briefly, high molecular weight (HMW) DNA is dephosphorylated to block the 5’ and 3’ ends. Next, a complex of Cas9 nuclease enzymes and guide RNAs is used to digest the DNA at precise cut sites upstream and downstream of the repeat regions of interest. Then, barcoded SMRTbell adapters are ligated following dA tailing of the newly cut ends. Finally, nuclease digestion removes non-SMRTbell molecules. The barcoded SMRTbell libraries can then be pooled before a final bead wash to prepare the sample for annealing, binding, and cleanup and subsequent sequencing.

[1] 1–4 µg input DNA supported. [2] 50% of mass of DNA molecules longer than 30 kb as measured on the Femto Pulse system (Agilent). [3] Mean coverage for 2 µg DNA of supported sample types (Nanobind-extracted human blood and cell line) for unexpanded repeat alleles. [4] The PureTarget kit supports smaller batches of samples in multiples of 8. [5] Inserts with expanded alleles will be longer. [6] Methylation probabilities for CpG sites encoded in BAM.

Data analysis

Analysis of PureTarget repeat expansion libraries can be performed in SMRT Link using the PureTarget repeat expansion analysis workflow or at the command line. Both options utilize the Tandem Repeat Genotyping Tool (TRGT, Dolzhenko et al., 2024,) to call repeat expansion alleles. TRGT was developed to analyze whole genome HiFi datasets but can also be applied to targeted sequencing datasets. In addition to producing consensus repeat sizes for both alleles, TRGT profiles repeat sequence composition, CpG methylation of each analyzed repeat, and mosaicism. The companion tool, TRVZ, can be used for visualization of reads overlapping the repeats, with visual reports available in SMRT Link for download. The genotypes provided by TRGT are intended for research only.

Figure 8 shows the major steps of the analysis workflow for SMRT Link or the command line. Samples are demultiplexed using lima, 5mC methylation probabilities for CpG sites are called with jasmine, reads are mapped to the hg38 reference genome with pbmm2, repeat genotypes are called with TRGT, and visualizations are produced with TRVZ.

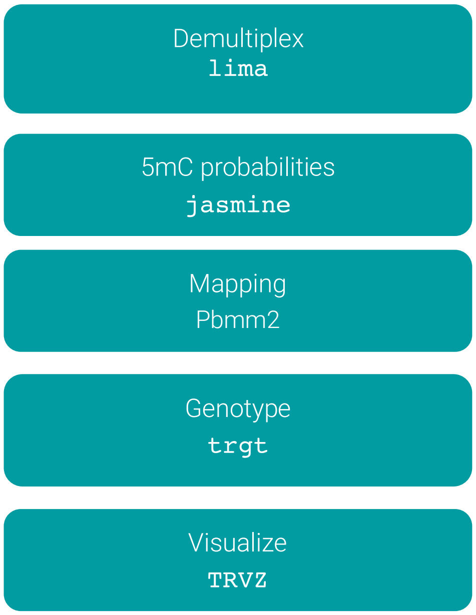

Figure 8. Major steps of the PureTarget bioinformatics workflow.

Configuration in SMRT Link

Users have two options for analysis when setting up sequencing runs in SMRT Link. For the fastest turnaround time and seamless analysis, users can include the PureTarget repeat expansion analysis in their run design and analysis will be automatically performed when sequencing is complete. Alternatively, users who prefer command line analysis may configure SMRT Link to do automatic demultiplexing only. Demultiplexed BAM files may then be transferred for command line analysis starting at the mapping step.

Sample types and DNA extraction methods

It is recommended that users obtain high-quality genomic DNA with Nanobind extraction kits from PacBio. Officially supported sample types are whole blood extracted using red blood cell (RBC) lysis method and the Nanobind PanDNA kit, PBMCs extracted with Nanobind CBB kit, or human cell lines extracted with Nanobind PanDNA or Nanobind CBB kit. When using sample types and extraction methods other than the above we recommend users:

- First, demonstrate success using supported sample types, starting with an 8-plex and increasing sample quantity thereafter.
- Introduce new sample types or extraction methods in limited numbers, for example, 3 or fewer new sample types in an 8-plex of otherwise controls.
- See table 3 for more information about samples that are officially supported, have been tested, or are not recommended.

Human Sample type	Extraction method	Category
Whole blood	Nanobind PanDNA kit (PacBio 103-260-000) Extracted using RBC-lysis method	Supported
Peripheral blood mononuclear cells (PBMC)	Nanobind PanDNA kit*	Supported
Commercial lymphoblastoid cell lines	Nanobind PanDNA kit Nanobind HT CBB kit – automated (PacBio 102-762-700)	Supported
Skeletal muscle	Nanobind PanDNA kit (PacBio 103-260-000)	Tested in low plex
Brain tissue		Tested in low plex
Myoblasts		Tested in low plex
Whole blood	FlexiGene DNA Whole Blood Kit– automated (AutoGen AGKT-FG-640)	Tested in high plex
Whole blood	Qiagen Genomic-Tip	Tested in high plex
Whole blood	QIAsymphony (Qiagen)	Tested in high plex
Whole blood	Bionano SP Blood and Cell Culture DNA Isolation Kit (80042)	Tested
Corneal endothelial (CEC) cell culture	Bionano SP Blood and Cell Culture DNA Isolation Kit (80042)	Tested
Whole blood	Monarch® HMW DNA Extraction Kit for Cells & Blood New Englad Biolabs T3050S/T3050L	Tested
Whole blood	Gentra Puregene Blood Kit Qiagen 158467/158389	Tested
Fibroblasts	Qiagen Genomic-Tip	Tested
Whole blood	chemagic DNA blood kit (Revity)	Not recommended

Table 3. Guidance on sample types and extraction methods. Low plex means that fewer than 8 samples extracted with this method were pooled with other sample extraction types and run on a SMRT Cell at 8-plex or higher. High plex means 8 or more samples extracted with the method were pooled and run on a SMRT Cell. *Also supported by Nanobind CBB kit (PacBio 102-301-900).

DNA quantity and read coverage

PureTarget libraries do not use amplification to enrich targets but rather retain targets of interest and deplete off-target molecules. As such, the library quantity loaded on the SMRT Cell is lower than other library types like WGS and the sequencing yield for a sample can be influenced by how much starting DNA is used in the library prep. Figure 9 shows that sample coverage for the target regions increases with the quantity of DNA used in the library prep for a give multiplex level.

The standard DNA mass recommended for PureTarget libraries is 2 µg. Users who wish to increase coverage of a sample may use up to 4 µg of DNA in library prep. If higher coverage is needed, we recommend multiple preps of 4 µg for the sample and combining the data during analysis. Note: coverage is lower at higher sample multiplex (Figure 6); if you wish to generate high coverage data we recommend running an 8-plex. It is not recommended to exceed a total cumulative mass of 75 µg per Sequel II run or 100 µg mass per Revio run.

If users have limited DNA quantity or do not need high coverage, they may use as little as 1 µg of DNA in the prep.

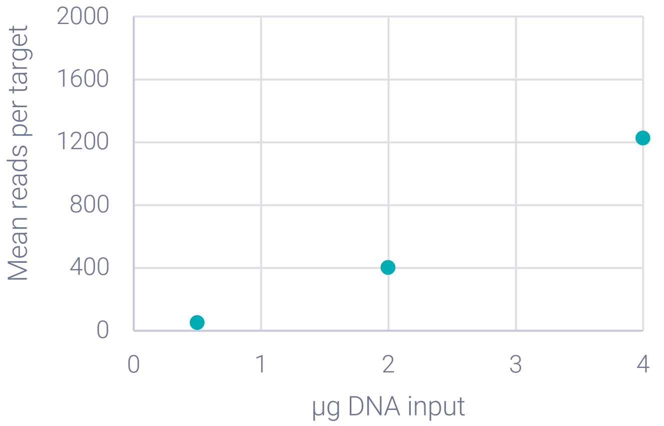

Figure 9. DNA input quantity versus coverage. DNA was extracted from whole blood using the Nanobind PanDNA kit and run in an 8-plex.

DNA quality and read coverage

PureTarget libraries contain SMRTbell templates that each span the full target region. This places requirements on the length of input DNA; minimally, the length of the input DNA must exceed the length of the target. Longer expansions therefore require higher quality, longer DNA as input to ensure full expansions are contained within single fragments of DNA. For best performance, HMW gDNA should be used with GQN (30 kb) >5 as measured by the Femto Pulse system. If the DNA is lower quality, fewer spanning regions may be contained in single reads, and low coverage will be observed, see Figure 10.

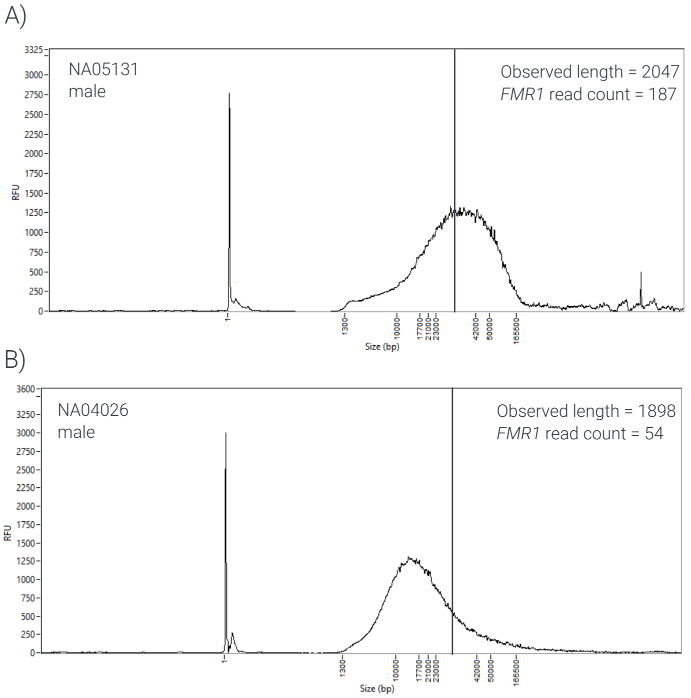

Figure 10. Comparison of a) high- and b) low-quality DNA samples illustrate that more on-target reads are observed at an *FMR1* repeat expansion for the sample with higher molecular weight DNA. Vertical line designates 30 kb.

Custom panels

Custom PureTarget panels are not officially supported by PacBio. Users must design and order their guide RNAs and optimize their custom panels accordingly. PacBio can offer limited guidance on guide RNA design software and strategies for optimizing custom panels. We recommend first demonstrating success on the PureTarget repeat expansion panel using supported sample types before adding new guide RNAs or testing a custom set of guides. Adding a small number of repeat expansion targets is relatively low risk provided fragments are of similar size to panel provided in the kit (4–5 kb). Success has been shown for adding up to 5 pairs of guides for additional repeat expansion targets. The SMRT Link PureTarget repeat expansion analysis or command line analysis with TRGT can be used with an updated target BED file that includes the new coordinates.

Adding a small number of gene targets with fragment sizes of ~5 kb is also low risk but will require different analysis tools. It is recommended to use Deep Variant or pbsv in place of TRGT for small variant calling or structural variant calling for non-repeat targets. Any other design strategy is untested and will require optimization. This includes tiled designs where targeted fragments overlap to target larger regions in the panel. The PureTarget repeat expansion panel is in total ~100 kb in length so performance in panels that are much smaller or much larger is unknown and may require optimization of the wet lab protocol or reagents.

Please consult with PacBio field application scientist or tech support for more specific details on library prep and see TRGT documentation for example formats and characterized repeats.

References

Dashnow, H., et al. (2022). STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. Genome Biology, 23(1), 1-20. https://doi.org/10.1186/s13059-022-02826-4

Depienne, C., & Mandel, J. L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? The American Journal of Human Genetics, 108(5), 764-785. https://doi.org/10.1016/j.ajhg.2021.03.011.

Dolzhenko, E., et al. (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome biology, 21, 1-14. https://doi.org/10.1186/s13059-020-02017-z

Dolzhenko, E., et al. (2024). Characterization and visualization of tandem repeats at genome scale. Nature Biotechnology, 1-9. https://doi.org/10.1038/s41587-023-02057-3

Ibañez, K., et al. (2022). Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. The Lancet Neurology, 21(3), 234-245. https://doi.org/10.1016/S1474-4422(21)00462-2.

Paulson, H. (2018). Repeat expansion diseases. Handbook of Clinical Neurology, 147, 105-123. https://doi.org/10.1016/B978-0-444-63233-3.00009-9.

Tarleton, J. (2003). Detection of FMR1 trinucleotide repeat expansion mutations using Southern blot and PCR methodologies. Neurogenetics: Methods and Protocols, 29-39. https://doi.org/10.1385/1-59259-330-5:29

Tsai, Y. C., et al. (2017). Amplification-free, CRISPR-Cas9 targeted enrichment and SMRT sequencing of repeat-expansion disease causative genomic regions. BioRxiv, 203919. https://doi.org/10.1101/203919

Tsai, Y. C., et al. (2022). Multiplex CRISPR/Cas9-Guided No-Amp targeted sequencing panel for spinocerebellar ataxia repeat expansions. In Genomic Structural Variants in Nervous System Disorders (pp. 95-120). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2357-2_6