Targeted long-read sequencing of native DNA for comprehensive characterization of repeat expansions

Abstract # 1875257

Sarah B Kingan[1], Guilherme De Sena Brandine[1], Jocelyne Bruand[1],Jeff Zhou[1], Valeriya Gaysinskaya[1], Janet Aiyedun[1], Julian Rochar , Duncan Kilburn[1], Egor Dolzhenko[1], Zoi Kontogeorgiou[1], Anita Szabo[1], Christina Zarouchlioti[1], Robert Thaenert[1], Pilar Alvarez Jerez[1], Kimberley Billingsley[1], Sonia Lameiras[1], Sylvain Baulande[1], Alice Davidson[1], Georgios Koutsis[2], Georgia Karadima[2], Stéphanie Tomé[3], Michael A Eberle[1]
1. Pacific Biosciences (PacBio), Menlo Park, United States, 2. National and Kapodistrian University of Athens, 1st Department of Neurology, Athens, Greece, 3. University College London, Institute of Ophthalmology, United Kingdom,
4. Quest Diagnostics, Marlborough, United States, 5. National Institutes of Health, Center for Alzheimer's and Related Dementias, National Institute on Aging, Bethesda, United States, 6. Institut Curie, PSL Research University, ICGex Next-Generation Sequencing Platform, Paris, France,
7. National and Kapodistrian University of Athens, Neurogenetics Unit, 1st Department of Neurology, Eginition Hospital, School of Medicine, Athens, Greece 8. Sorbonne Université, Inserm, Institut de Myologie, Centre de Recherche en myologie, Paris, France

Introduction

Short tandem repeats (STRs) are DNA sequences composed of repetitions of 1 – 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 long-read sequencing of unamplified DNA holds the promise to resolve all these features in a single assay.

Scalable amplification-free workflow with PureTarget

High molecular weight DNA
- 50% fragments > 30 kb
- 1 – 4 µg per sample

1

Dephosphorylate to block DNA ends

2

Cas9 cut with pair of sgRNAs
- "normal" length = 4 – 5 kb

3

dA tail cut ends

4

Ligate indexed SMRTbell adapters

5

Nuclease digestion of non-SMRTbell templates

6

Pool and sequence up to 48 samples on Revio or Vega system

7

Figure 1. PureTarget 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 regions and ~1-2 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

Results in reference and positive samples

To assess the accuracy of this method, we sequenced 129 samples with validated pathogenic expansions at *CNBP*, *DMPK*, *RFC1*, *C9orf72*, and 16 other loci. Combined, we tested 2580 sample-expansion combinations, including technical replicates, for expansions spanning between 66 bp and >10 kb. Our assay correctly categorized all (129/129) expansions, detected hypermethylation in the *FMR1* expansion, and identified the pathogenic AAGGG motif in the *RFC1* repeat. We discovered additional expansions of the *TCF4* repeats and *FXN*, *RFC1* (not shown), which is consistent with these loci having carrier frequencies between 1:50 and 1:20.

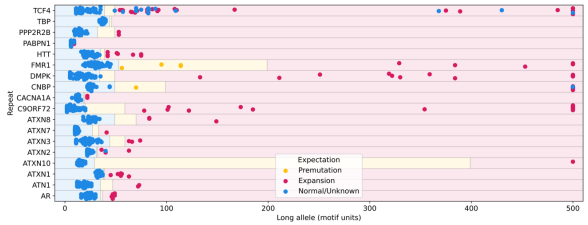

Figure 2. "Swim lane plot" showing the long allele length of 150 samples for 18 autosomal and X-linked dominant loci. Dots are colored by expected genotype.

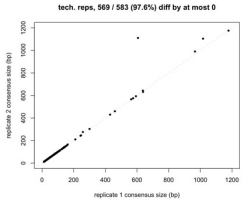

Figure 3. Reproducibility of consensus length in 15 pairs of technical replicates, 8 males and 7 females. 570/584 have identical consensus sequences, 577 are at most off by 1, and all (584/584) have concordant ranges, meaning the range of allele sizes overlap between replicates.

Concordant genotypes for Friedreich's ataxia

		Long allele			Short allele	
FXN Sample	Coverage	Observed motif count	Expected motif count	Coverage	Observed motif count	Expected motif count
HM16212	92	471	500	146	8	<30
NA16202	49	817	830	86	8	<30
NA16212	68	515	500	99	8	<30
NA16237	49	699	700	134	8	<30

Table 1. Expansions in FXN repeat, associated with Friedreich's ataxia, are concordant with expectations. Samples were prepared with 2 µg DNA each and sequenced on the Vega system in a 24-plex of samples. Note the read coverage is high for both short and long alleles giving confidence in the call.

Detecting AGG interruption and methylation in a single assay for *FMR1*

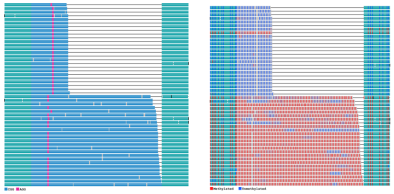

Figure 4. TRGT waterfall plots of Coriell reference sample NA06905. Allele plot on left shows AGG interruption sequence in both short (23 repeats) and long (79 repeats) allele. Expected genotype is 23 and 70 repeats by southern blot and PCR. Methylation plot on right shows 5mC hypermethylation at CpG sites in the long allele and hypomethylation on short allele. Reads are filtered for Q20 and randomly down-sampled.

Detecting pathogenic expansions in *RFC1*

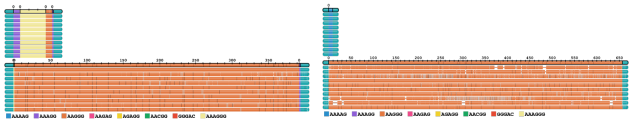

Figure 5. TRGT motif allele plots of HG01175 (left) and NA20752 (right) at *RFC1* repeat showing consensus allele for the normal (top) and pathogenic expanded allele (bottom) with phased reads aligned to each consensus sequence. In HG1175, 393 "AAGGG" motifs are observed in the expanded allele with consensus length of 1948 bp. Note the diversity of repeat motifs which are distinguishable by HiFi reads. In NA20752, 657 "AAGGG" motifs are observed in the expanded allele with consensus length of 3253 bp.

Coverage of PureTarget panel on HiFi systems

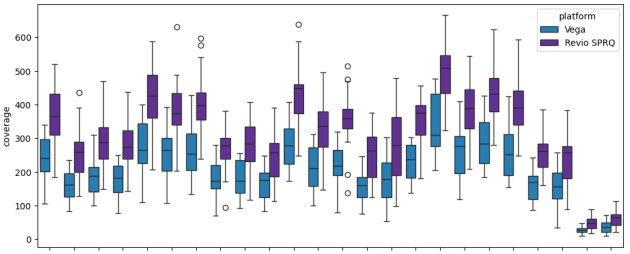

Figure 6. Reference Coriell samples with known repeat expansions (N=24) were prepared with 1 µg DNA input on Revio SPRQ chemistry and 2 ug DNA input on Vega. Full dataset available at https://downloads.pacbcloud.com/public/2024Q4/Vega/PureTargetCoriell24/

Conclusion

- PureTarget is complete solution to accurately characterize lengths, repeat sequence and methylation status of repeat expansions relevant for human disease
- PureTarget repeat expansion panel, protocol, and analysis in SMRT Link can deliver sample to answer in 3 days.

Resources

PureTarget website

PureTarget app note

A talk on resolving complex tandem repeats

Tandem repeat genotyping tool (TRGT) publication

References

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.

Acknowledgements

The authors would like to thank Jonas Korlach, Mozhgan Novbakhtian, and Kristin Robertshaw for thoughtful comments on the poster and assistance with graphics.