Direct Detection and Counting of Trinucleotide Repeat Expansion Using CRISPR-Cas9 and PacBio HiFi Sequencing

Bhaven Mehta[1], David Corney[1], Wei Wang[1], Yu Qiu[1], Jian Chen[2], Hoda Bahgat[2], Faraz Zaidi[2], Christopher Mozdzierz[1], Haythem Latif[1]

[1]GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

[2]CHDI Foundation, New York, NY 10001

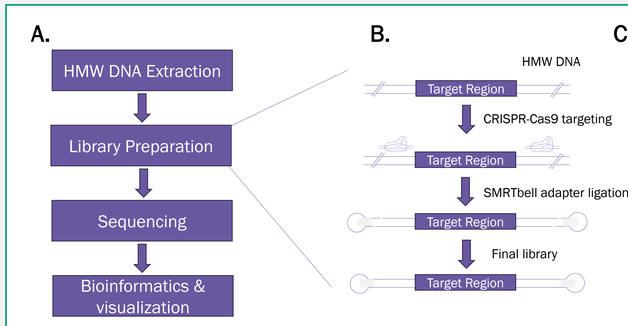
Abstract

Trinucleotide repeat expansions are a major cause of neurological and developmental disorders, including Huntington's disease, fragile X syndrome, and various spinocerebellar ataxias. When the number of repeats exceeds a certain pathogenic threshold length, they can lead to toxic gain-of-function or loss-of-function effects. Historically, PCR and Southern blot approaches have been used to identify repeat expansions. However, there are significant technical challenges associated with these methods, including successfully amplifying through hundreds or thousands of base pairs of repetitive sequences and failed PCR due to high GC content. As a result, these approaches suffer from poor accuracy and resolution of the number of repeats. Consequently, the loci containing these genes are commonly referred to as “dark” regions of the genome and are poorly characterized.

To circumvent these challenges, we have tested a novel PCR-free CRISPR-Cas9 approach to enrich 20 genes commonly involved in repeat expansion disorders followed by long-read PacBio® sequencing to successfully sequence and align repetitive sequence. High molecular weight DNA was extracted from frozen EDTA whole blood or fresh frozen tissue, followed by PureTarget library preparation. In this workflow, DNA is first digested with Cas9 and guide RNAs to enrich 20 target genes, followed by ligation of SMRTbell® adapters on the cut end. After removal of non-ligated templates, libraries are pooled and sequenced on a PacBio sequencer.

Here, we describe the accuracy testing results of this novel approach. Accuracy of calling the number of trinucleotide repeats was established by processing previously characterized samples from Coriell with known expansions in DMPK, ATX1, ATXN3, FMR1, and FXN. Additionally, DNA extracted from blood of healthy donors without any known expansions, and from brain tissue of Huntington disease individuals was performed and compared to whole genome PacBio data.

The CRISPR-Cas9 approach combined with long-read PacBio sequencing offers a powerful method for characterizing trinucleotide repeat expansions, overcoming the limitations of traditional techniques. This novel approach paves the way for improved understanding of the genotype-phenotype and molecular basis of neurological and developmental disorders caused by repeat expansions.



C.

DISEASE	TARGET GENES
Spinocerebellar ataxia	ATN1, ATXN1, ATXN2, ATXN3, ATXN7, ATXN8, ATXN10, CACNA1A, PPP2R2B, TBP
Friedrich’s ataxia	FXN
CANVAS disease	RFC1
Fragile-X disease (FXS)	FMR1
Huntington’s disease	HTT
Myotonic dystrophy	DMPK, CNBP
ALS, FTD	C9orf72
Fuchs endothelial corneal dystrophy	TCF4
Spinal bulbar muscular atrophy / Kennedy’s disease	AR
Oculopharyngeal muscular dystrophy	PABPN1

Figure 1. (A) Schematic of repeat expansion analysis using CRISPR-Cas9. (B) High molecular weight (HMW) DNA is dephosphorylated followed by CRISPR-Cas9 targeting with gRNAs against either end of the target regions. Subsequently, PacBio SMRTbell adapters are ligated to generate a final library which is sequenced on a PacBio Revio™ sequencer. Bioinformatic analysis and visualization is performed to identify repeat expansions. (C) Target genes contained within the PacBio PureTarget™ panel.

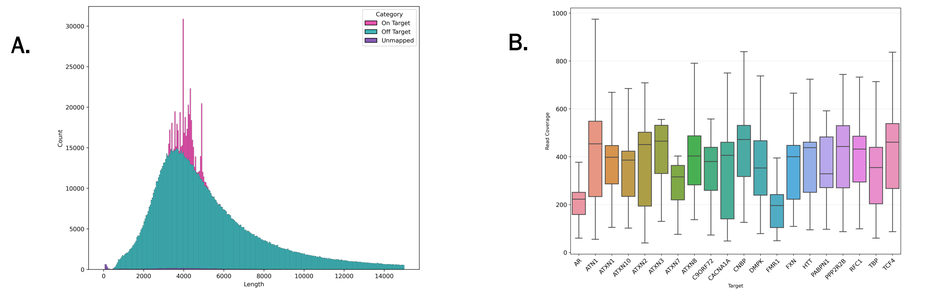
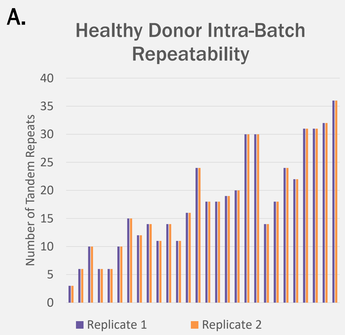


Figure 2. Read alignment characteristics. (A) Plot showing read mapping properties. A large proportion of reads are off-target. Nevertheless, a significant number of reads align to the 20 on-target regions included within the PureTarget panel. (B) Boxplot illustrating on-target coverage on a per-gene basis. Samples included a mixture of HMW and partially degraded samples which likely contributed to greater variability. A total of 22 samples were analyzed across 2 independent processing batches, with 99.8% of regions having at least 50x coverage.



B.

GENE	REPEAT UNIT	SAMPLE ID	BATCH	ALLELE 1		ALLELE 2	
				Expected Repeats	Observed Repeats	Expected Repeats	Observed Repeats
DMPK	CAG	NA03697	1	N/A	12	< 500	383
			2		12		384
					12		383
					12		383
ATXN1	CAG	NA06926	1	29	30	52	54
			2		30		54
					30		54
					30		54

Figure 3. Assay repeatability. (A) HMW DNA from a healthy donor was analyzed twice within the same processing batch. Identical number of tandem repeats was observed for each replicate in every gene. (B) Two Coriell samples were analyzed in triplicate in one batch and independently in another batch with largely identical results consistent with the expected number of repeats.

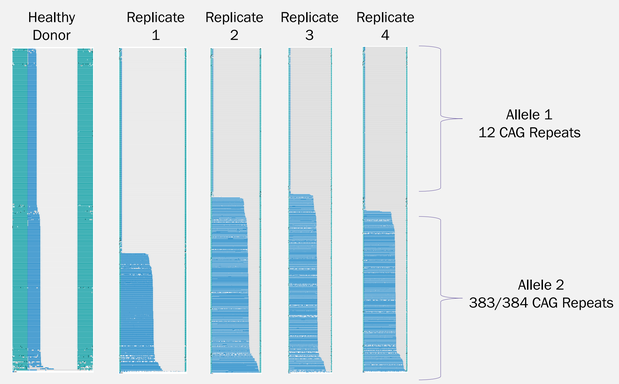


Figure 4. Example illustrating DMPK repeat expansion. Results are shown in a waterfall plot for each of four independent technical replicates of Coriell sample NA03697 alongside a healthy donor. Heterogeneity is observed with regards to repeat number of allele 2 of NA03697. Teal: flanking sequence; Blue: CAG repeat.

A.

GENE	REPEAT UNIT	SAMPLE ID	ALLELE 1		ALLELE 2	
			Expected Repeats	Observed Repeats	Expected Repeats	Observed Repeats
ATXN3	GCT	NA06153	23	20	71	69
FMR1	CGG	NA06905	23	23	70	79
FXN	GAA	NA15850	650	634	1030	705

B.

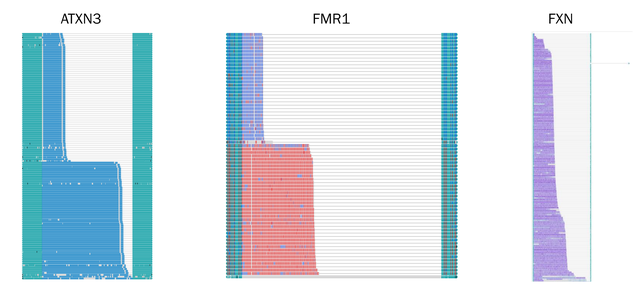


Figure 5. Concordance of PureTarget in Coriell reference samples. (A) Summary of tandem repeats in three Coriell samples. (B) Waterfall plots of ATXN3, FMR1, and FXN repeats identified in part A, including CpG methylation status in FMR1. Teal: flanking sequence; Blue/purple: repeat unit; Red/blue (FMR1): methylated/unmethylated CpG.

A.

GENE	REPEAT UNIT	DISEASE STATUS	ALLELE 1		ALLELE 2	
			HiFi WGS Repeats	PureTarget Repeats	HiFi WGS Repeats	PureTarget Repeats
HTT	CAG : CCG	HD	19 : 10	19 : 10	44 : 9	44 : 9
		HD	22 : 9	22 : 9	49 : 9	50 : 9
		Control	17 : 9	17 : 9	20 : 12	20 : 12
		Control	17 : 12	17 : 12	21 : 12	21 : 12
		Control	19 : 12	19 : 12	19 : 12	19 : 12

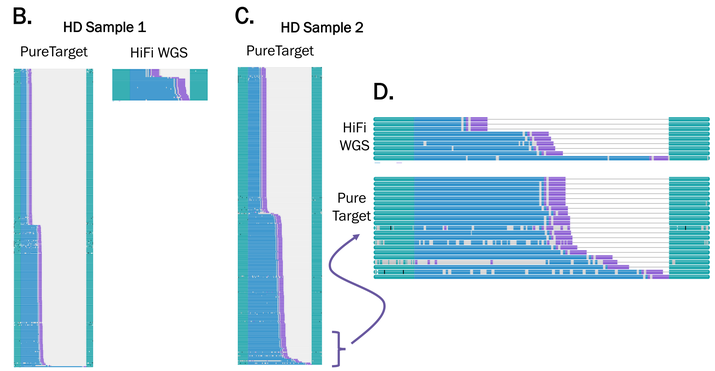


Figure 6. Concordance of PureTarget compared to HiFi WGS. HMW DNA obtained from five brain tissue samples was analyzed by PureTarget and HiFi WGS. (A) Summary table of HTT genotype. (B,C) HTT TR plots within two Huntington's Disease brain tissues identified by PureTarget and HiFi WGS. (D) Magnified view of panel C highlighting rare reads with > 100 CAG repeat units. Teal: flanking sequence; Blue: CAG repeat; Purple: CCG repeat.

Conclusions

- The PureTarget workflow enables detection of tandem repeat (TRs) in 20 genes with high inter- and intra-batch consistency.
- Expected TRs are identified in all genes tested, including DMPK, ATXN1, FMR1, and HTT.
- PureTarget is cost-effective to identify HTT expansions with greater sensitivity and scalability compared to HiFi WGS.

Acknowledgements: We are grateful to Ping Xu, Mia Sutton, and Ye Wang (Azenta) for sequencing assistance, and Sarah Kingan (PacBio) for discussions and technical support on the PureTarget workflow.