Targeted sequencing of 50+ pathogenic repeat expansions using Twist target enrichment and PacBio HiFi sequencing

Tina Han[1], Holly Corbitt[1], Eirini M. Lampraki[2], Fer Tornos[1], Sam Holt[3], David Stucki[2], Esteban Toro[1], Sarah Kingan[3], Chad Locklear[1]

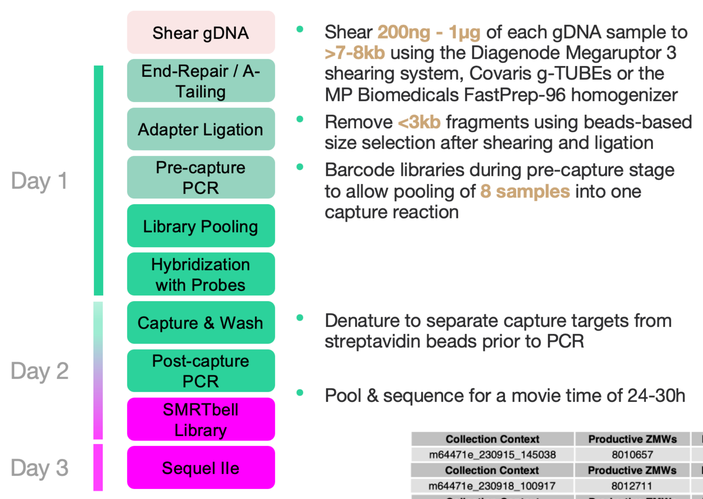
[1]Twist Bioscience, South San Francisco, California, USA; [2]Pacific Biosciences UK, Ltd., London, UK; [3]Pacific Biosciences, Menlo Park, California, USA

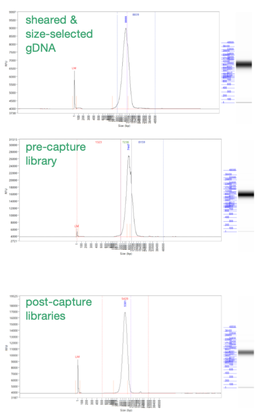
1. Introduction

The expansion of unstable genomic short tandem repeats (STRs) has been identified as the causal DNA mutation in more than 30 Mendelian diseases. PCR-based DNA fragment analysis, short-read sequencing, or legacy molecular genotyping methods like Southern blot are frequently used to analyze STR expansions, but are not capable of determining the exact length and sequence composition. Targeted sequencing allows for high-resolution characterization of dozens to thousands of gene regions at a scale and cost that is more accessible than whole genome sequencing.

2. Methods

Here we describe the performance of a Twist target enrichment panel sequenced with PacBio long HiFi reads to measure the size of STR expansions with base-pair resolution. Using a proprietary algorithm, we designed a gene panel targeting 50+ regions with known pathogenic repeat expansion alleles in human samples. We sheared 200-1000 nanograms (ng) of each gDNA sample to >7-8 kilobases (kb) using the Diagenode Megaruptor 3 shearing system, Covaris g-TUBEs or the MP Biomedicals FastPrep-96 homogenizer. After end-repair, A-tailing, and adapter ligation, 10-bp unique dual indices were added during PCR with KOD Xtreme Hot Start DNA polymerase. Every 8 samples were pooled for an overnight hybridization. After capture and wash, the post-capture libraries were then amplified and converted into SMRTbell library and sequenced on PacBio Sequel IIe instrument (also compatible with the Revio system) with resulting HiFi read length of 5-10kb. We genotyped repeat copy numbers using Tandem Repeat Genotyping Tool (TRGT) and visualized reads spanning repeats using TRVZ.





Collection Context	Productive ZMWs	Productivity 0	(%)	Productivity 1	(%)	Productivity 2	(%)
m64471e_230915_145038	8010657	1113895	13.90	6579553	82.09	321223	4.01
Collection Context	Productive ZMWs	Productivity 0	(%)	Productivity 1	(%)	Productivity 2	(%)
m64471e_230918_100917	8012711	1291128	16.11	6495682	81.05	227861	2.84
Collection Context	Productive ZMWs	Productivity 0	(%)	Productivity 1	(%)	Productivity 2	(%)
m64471e_230929_154904	8012245	1224127	15.27	6602000	82.37	188544	2.35

SMRT Link v11.0 was used to generate HiFi reads, mark PCR duplicates, and demultiplex. Variants were called using a PacBio targeted sequencing pipeline, including DeepVariant, phasing with WhatsHap, and targeted metric calculation with Picard. More information is available on Github: https://github.com/PacificBiosciences/HiFiTargetEnrichment

3. Targeted Repeat Expansions (selected list)

Gene	Motif	Phenotype
AFF2	GCC	Fragile XE syndrome (FRAXE)
AR	GCA	Spinal-bulbar muscular atrophy (SBMA)
ATN1	CAG	Dentatorubral-pallidoluysian atrophy (DRPLA)
ATXN1	TGC	Spinocerebellar ataxia type 1 (SCA1)
ATXN2	GCT	Spinocerebellar ataxia 2 (SCA2)
ATXN3	GCT	Spinocerebellar ataxia type 3 (SCA3)
ATXN7	GCA, GCC	Spinocerebellar ataxia type 7 (SCA7)
ATXN8	CTA	Spinocerebellar ataxia type 8 (SCA8)
ATXN10	ATTCT	Spinocerebellar ataxia type 10 (SCA10)
C9ORF72	GGCCCC	C9orf72-linked frontotemporal dementia or amyotrophic lateral sclerosis (C9FTD/ALS)
CACNA1A	CTG	Spinocerebellar ataxia type 6 (SCA6)
CNBP	CA, CAGA, CAGG	Myotonic dystrophy type 2 (DM2)
DMPK	CAG	Myotonic dystrophy 1 (DM1)
EIF4A3	CCTCGCTGTGCCGCTGCCGA	Richieri-Costa–Pereira syndrome (RCPS)
FMR1	CGG	Fragile X syndrome (FXS)
FXN	A, GAA	Friedreich's ataxia (FRDA)
HTT	CAG, CCG	Huntington's disease (HD)
NUTM2B	GGC	Oculopharyngeal Myopathy with Leukoencephalopathy 1 (OPML1)
PABPN1	GCG	Oculopharyngodistal myopathy (OPDM)
PHOX2B	GCN	Congenital central hypoventilation syndrome (CCHS)
RFC1	AAAAG	Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS)

5. Conclusions

From this pilot study, we provide design and workflow guidance to researchers interested in targeted long-read sequencing for scalable and cost-efficient hybrid capture of 50+ different repeat expansions with long read lengths, minimizing coverage bias, and maximizing accuracy to fully capture all variant types. Currently, the optimization of repeat length of *C9ORF72* and *CNBP* is still ongoing. Having access to clinical samples would greatly assist with the validation effort.

6. Reference

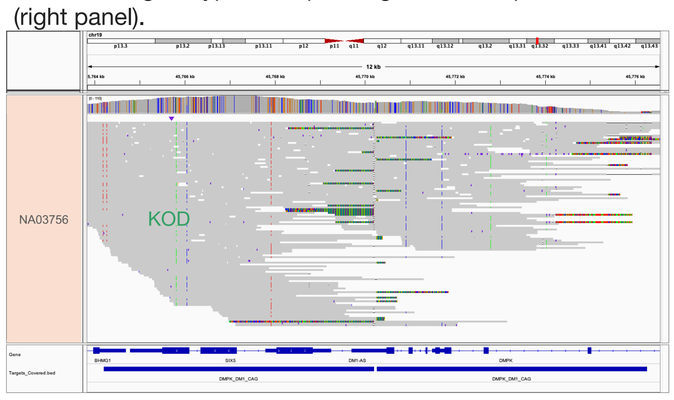
1. Twist Long Read Library Preparation and Standard Hyb v2 Enrichment https://www.twistbioscience.com/resources/protocol/long-read-library-preparation-and-standard-hyb-v2-enrichment
2. TRGT: Tandem Repeat Genotyper https://github.com/pacificBiosciences/trgt/

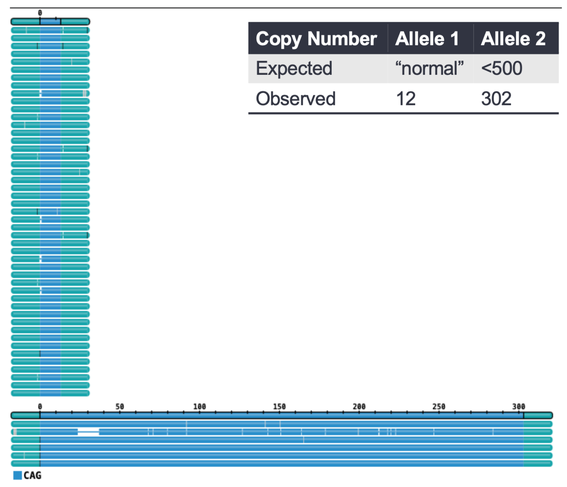
4. Results

We benchmarked performance of the long-read capture workflow in reference samples, of which 9 have characterized pathogenic repeat expansions alleles with diverse repeat base composition (namely NA15850, NA03696, NA16212, NA23709, NA13509, NA03756, NA06751, NA06968, NA23629) and 1 normal control (HG001=NA12878). Preliminary analysis showed two CAG repeats in the *HTT* and *AR* genes, a CTG repeat expansion in *DMPK*, a GAA repeat expansion in *FXN*, and GC-rich expansions in several other genes could be captured and sequenced. We also tested four different commercially available PCR polymerases and KOD Xtreme Hot Start DNA polymerase consistently provided the highest sequence coverage across the samples and targets for each condition. We found concordance between observed and expected repeat number for the genes *HTT*, *DMPK*, *AR*, and *PABPN1*. Optimizing the coverage of a few genes and testing on clinical samples with expanded alleles are ongoing.

CAG repeats in *DMPK* (Myotonic dystrophy 1)

Spanning reads with expanded allele could be observed in IGV (left panel). Output of TRGT as TRVZ plots showing concordant genotype and spanning read for expanded allele (right panel).

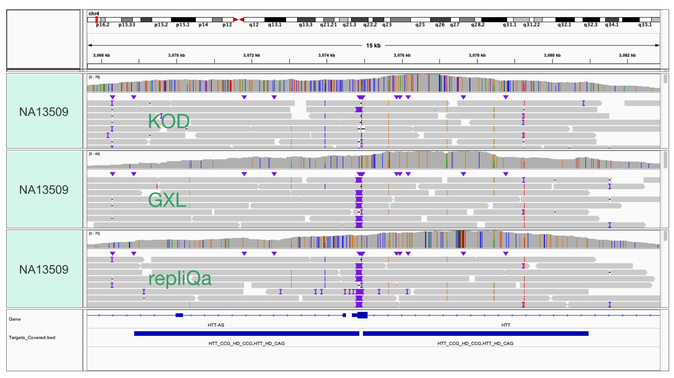


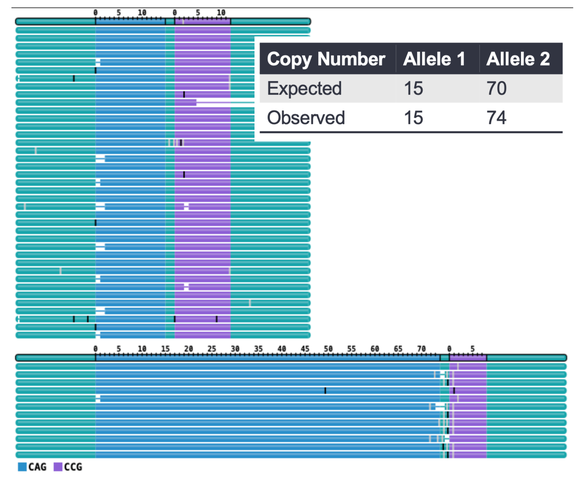


Copy Number	Allele 1	Allele 2
Expected	“normal”	<500
Observed	12	302

CAG and CCG repeats in *HTT* (Huntington’s disease)

Good coverage could be obtained from 3 different polymerases tested (left panel). Output of TRGT as TRVZ plots showing concordant genotype (right panel).





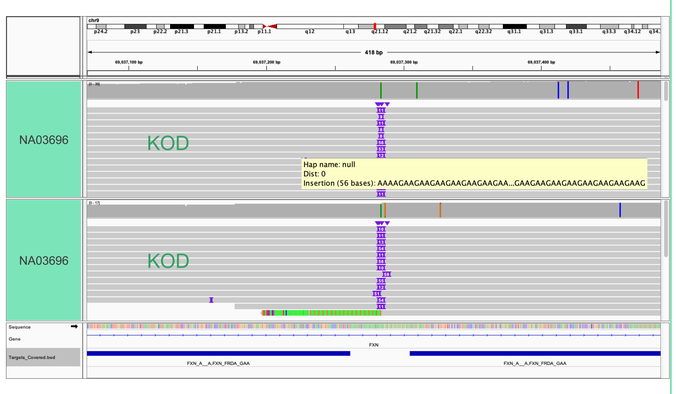
Copy Number	Allele 1	Allele 2
Expected	15	70
Observed	15	74

AAAAG repeats in *RFC1* (CANVAS)

Size and STR motif are in concordance with published literature [Stevanovski *et al., Sci. Adv.* 8, eabm5386 (2022)]

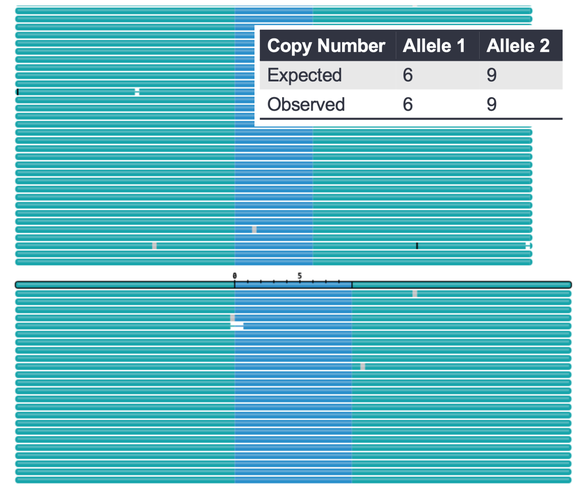
GAA repeats in *FXN* (Friedreich's ataxia)

A long stretch of A and GAA repeats could be detected



GCG repeats in *PABPN1* (Oculopharygnodistal myopathy)

The expanded allele is with 9 bp insertion, which equals to 3 more copies than the normal allele with 6 reference count (left panel). Output of TRGT as TRVZ plots showing concordant genotype (right panel).



Copy Number	Allele 1	Allele 2
Expected	6	9
Observed	6	9