Sequencing the Previously Unsequenceable Using Amplification-Free Targeted Enrichment Powered by CRISPR/Cas9

Jenny Ekholm, Yu-Chih Tsai, Ian McLaughlin, Brendan Galvin and Janet Ziegle
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Genomic regions with extreme base composition bias and repetitive sequences have long proven challenging for targeted enrichment methods, as they rely upon some form of amplification. Similarly, most DNA sequencing technologies struggle to faithfully sequence regions of low complexity. This has especially been true for repeat expansion disorders such as Fragile X syndrome, Huntington's disease and various Ataxias, where the repetitive elements range from several hundreds of bases to tens of kilobases.

We have developed a robust, amplification-free targeted enrichment technique, called No-Amp Targeted Sequencing, that employs the CRISPR/Cas9 system. In conjunction with Single Molecule, Real-Time (SMRT) Sequencing, which delivers long reads spanning the entire repeat expansion, high consensus accuracy, and uniform coverage, these previously inaccessible regions are now accessible. This method is completely amplification-free, therefore removing any PCR errors and biases from the experiment. Furthermore, this technique also preserves native DNA molecules, allowing for direct detection and characterization of epigenetic signatures.

The No-Amp method is a two-day protocol, compatible with multiplexing of multiple targets and samples in a single reaction, using as little as 1 µg of genomic DNA input per sample. We have successfully targeted a number of repeat expansion disorder loci (*HTT*, *FMR1*, *ATXN10*, *TCF4*, *C9orf72*) with alleles as long as >2700 repeat unites (>13 kb). Using the No-Amp method we have isolated hundreds of individual on-target molecules, allowing for reliable repeat size estimation, mosaicism detection and identification of interruption sequences – all aspects of repeat expansion disorders which are important for better understanding the underlying disease mechanisms.

No-Amp Targeted Sequencing Overview

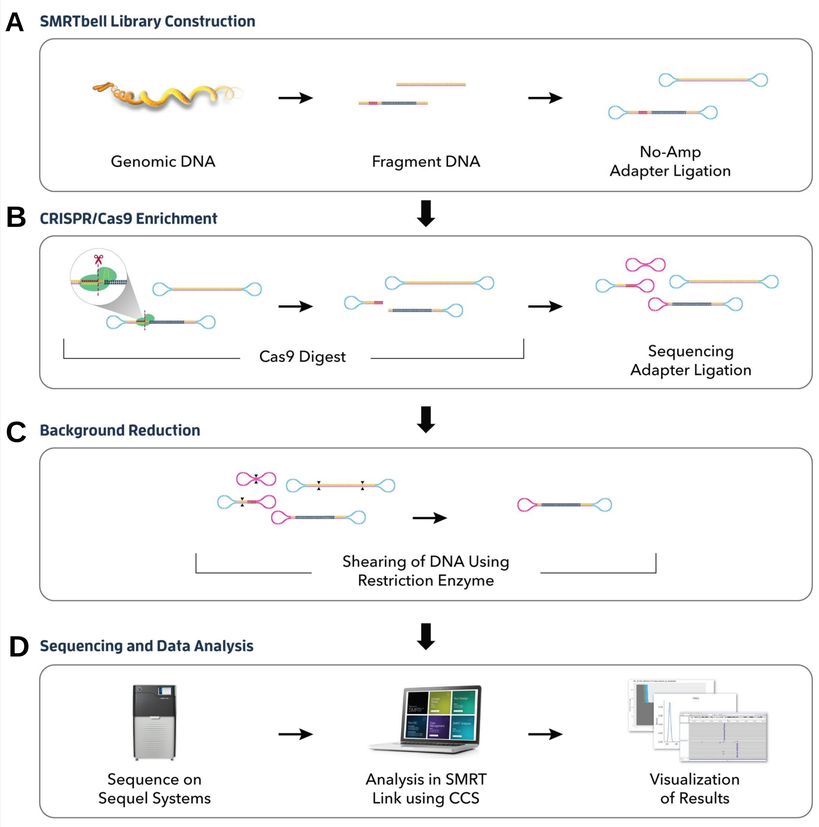

Figure 1. Overview of the No-Amp workflow. **(A)** The first step of No-Amp targeted sequencing is to construct a SMRTbell library by shearing the DNA using a restriction enzyme that does not cut within the target. This is followed by No-Amp adapter ligation. These adapters lack the sequencing polymerase binding sites and can therefore not be used for sequencing. **(B)** A guide RNA is designed to target the 5' end of the region of interest where the Cas9 will do a double-stranded digest. The fragments containing the target will now have a free end where the sequencing adapter can be added. **(C)** In order to remove background, two to four restriction enzymes and exonucleases that do not cut the region of interest will be used. **(D)** The SMRTbell libraries will then be sequenced on a Sequel System and analyzed in SMRT Link using circular consensus sequencing (CCS) analysis. The results can then be viewed in IGV, or using command-line tools that show repeat lengths for both alleles, characterization of mosaicism and interruption sequences.

Sequencing of Repeat Expansion Loci

Target Gene	Associated Diseases	Chr	Repeat
HTT	Huntington's Disease	4	CAG
ATXN10	Spinocerebellar Ataxia Type 10	22	Variable
FMR1	Fragile X and Fragile X-associated Tremor/Ataxia Syndrome	X	CGG

Table 1. Targeted disease loci. Three repeat expansion disease loci were targeted using the No-Amp method.

Targeted sequencing using the No-Amp method was carried out on Coriell cell lines from patients with known repeat expansion disease loci for *HTT* and *FMR1*.

We used 2 µg of input gDNA per sample and multiplexed 5 samples using barcodes as well the 2 disease loci on one SMRT Cell.

Results

We generated 140-718 on-target reads per sample for both *HTT* and *FMR1*, with equal representation of the normal and expanded allele. This represents enrichment factors of 18.879-197.508.

With CCS we are able to sequence each molecule multiple times, allowing for ≥QV20 for allele calling. The results are summarized in Table 2.

A. *HTT* Results

Sample ID	# On-Target Reads		Results Comparisons	
	Total	Per normal and expanded allele	Coriell Reported Repeat alleles sizes	PacBio repeat allele sizes
GenScript	565	565, N/A	N/A	17/18
NA13505 (M)	683	331, 352	22/50	23/52
NA13509	485	242, 243	15/70	15/75
NA20253 (M)	429	219, 210	22/108	23/113
NA14044 (M)	148	122, 26	19/205	20/185-1708

B. *FMR1* Results

Sample ID	# On-target reads		Results comparisons	
	Total	Per normal and expanded allele	Coriell Reported Repeat alleles sizes	PacBio repeat allele sizes
GenScript	718	718, N/A	N/A	30
NA20236	238	125, 110	31/53	31/55
NA20241	140	77, 63	29/101+	29/92-141
NA06896	236	125, 111	23/95-150	23/110-257
NA07537	249	219, 40	28-29/300	29/306-395

Table 2. **Comparative results for the Coriell samples.** The tables above show the number of on-target reads for the Control (GenScript) and Coriell Samples for targets **(A)** *HTT* and **(B)** *FMR1* for PCR genotyping and No-Amp, respectively.

Impact of gDNA Quality on Results

Sample quality has the biggest effect on target recovery. The figures below show the Femto Pulse traces for samples NA13509 and NA14044. The lower quality (degraded) DNA of sample NA14044 results in fewer on-target reads and an imbalance in the ratio of normal to expanded allele counts. While the higher quality DNA for sample NA13509 yielded twice as many on-target reads and an even balance between the normal and expanded allele.

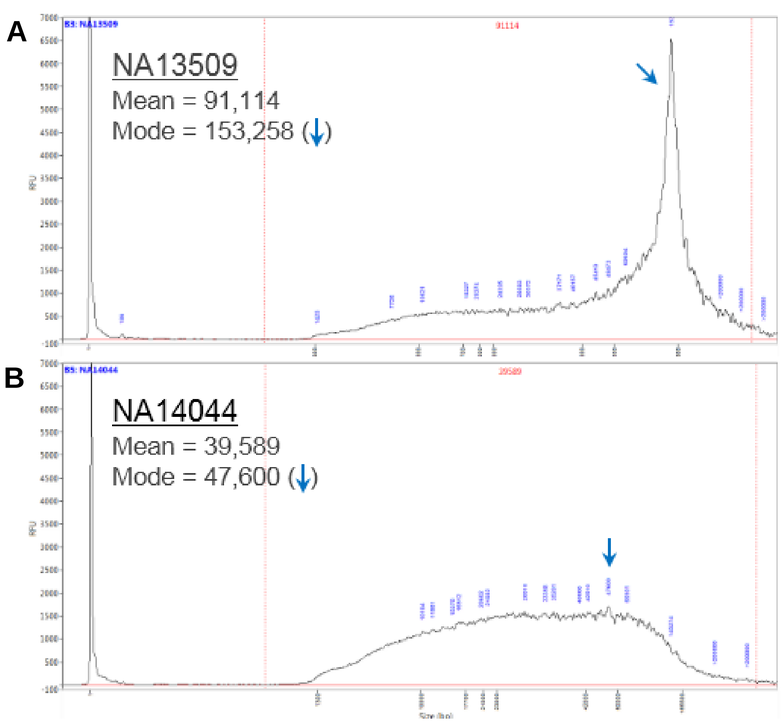

Figure 2. **Effect of DNA quality on yield.** Femto Pulse QC traces for DNAs **(A)** NA13509 and **(B)** NA14044.

Visualization of Results

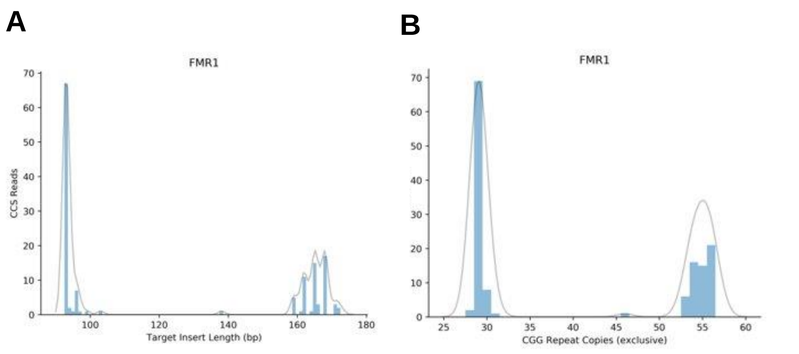

Figure 3. Repeat size and characterization mosaicism. Using command-line tools the repeat size distribution for both alleles can seen for sample NA20236 for *FMR1*. **(A)** Shows the target repeat length in base pairs. **(B)** Shows the target repeat length in terms of repeat copies.

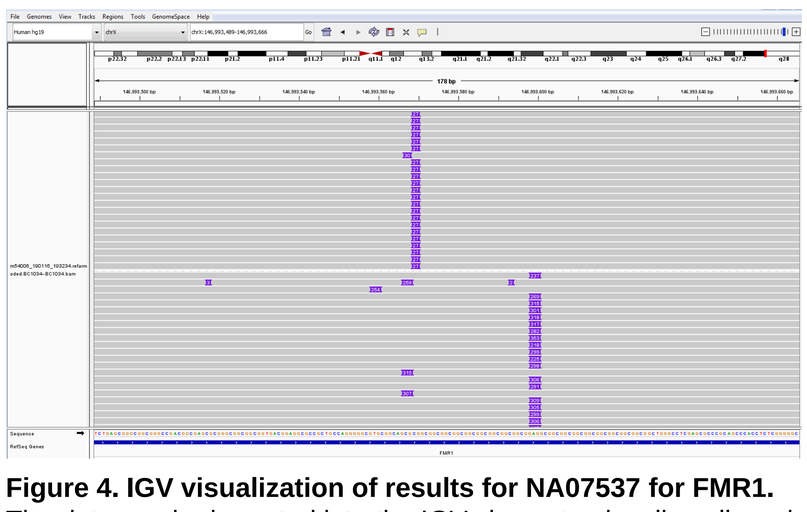

Figure 4. IGV visualization of results for NA07537 for FMR1. The data can be imported into the IGV viewer to visualize all reads for both alleles (29/306-395) of the target region.

Visualization of Results

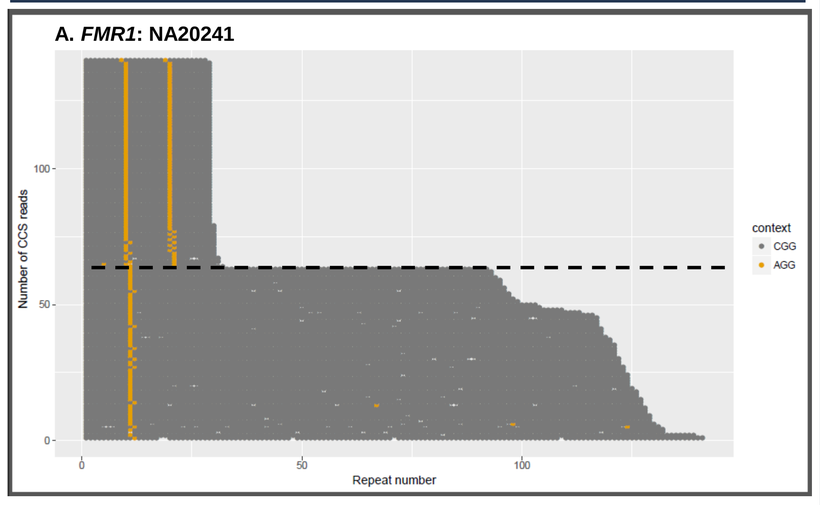

Figure 5. **Waterfall plots identifies interruption sequences. (A)** For NA20241 two interruption sequences (AGG) in the *FMR1* gene can be seen for the normal allele, whereas the expanded allele only has one interruption sequence. **(B)** For NA13509 the normal allele and the expanded allele both have varying numbers of repeat elements in the *HTT* gene, CAG being the most dominant one.

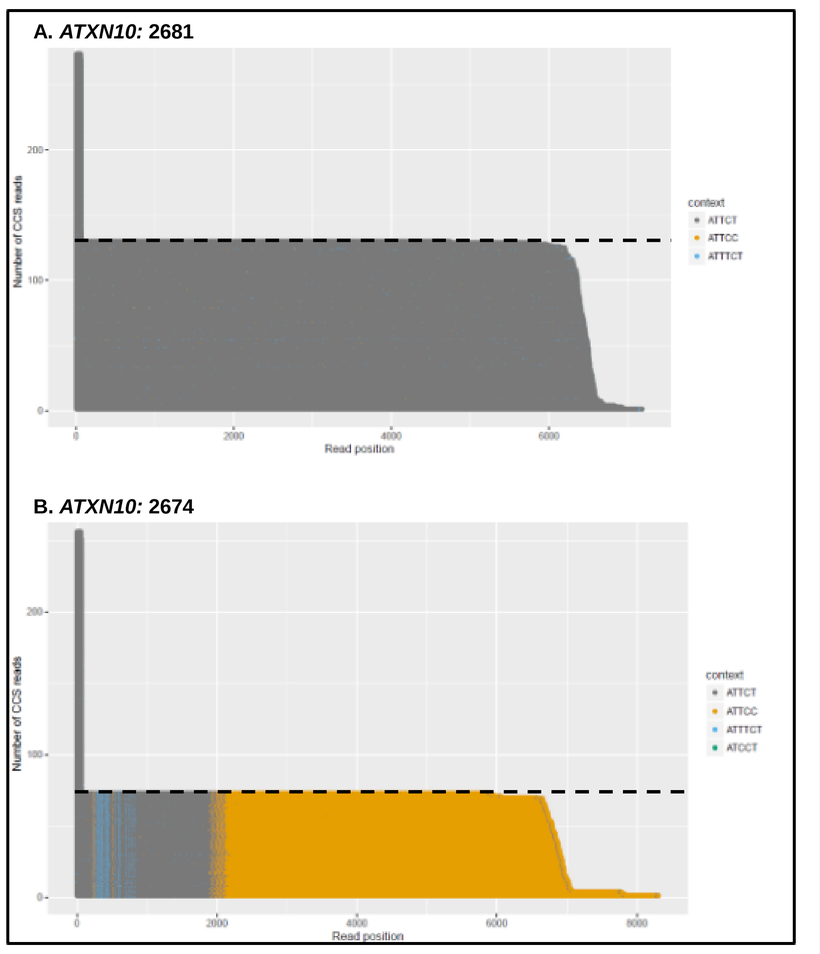

Figure 6. **Interruption sequences can be clinically relevant.** Schüle, et al.[1] has previously described a large pedigree segregating both SCA10 and Parkinson's disease. Here the samples were re-sequenced using the Sequel System **(A)** Patient 2681 that has been diagnosed with Parkinson's disease reveals a pure ATTCT repeat motif with no interruptions. **(B)** Patient 2674 from the same family reveled aside from the ATTCT expansion also interruption sequences (ATTCC, ATTTCT) resulting in the differential diagnosis of ataxia accompanied by seizures.

Conclusion

The No-Amp method allows elimination of PCR bias and errors, and captures holistically in one experiment:

- Base-level resolution into the expanded region
- Repeat count for both normal and mutated expanded allele
- Medically relevant interruption sequences
- Characterization of somatic mosaicism

The No-Amp method overview:

- Sequel System compatibility
- 2-day protocol
- Multiplexing of regions of interest
- Multiplexing up to 10 samples
- Input gDNA requirement:
  - 1-2 µg/sample (when multiplexing)
  - 10-20 µg/SMRT Cell
- Enriched target size: 1-5+ kb
- Validated targets: *HTT, FMR1, ATXN10*
- Data analysis:
  - CCS analysis
  - Visual reporting tools for repeat count, mosaicism, sequence interruptions

References

1. Schüle, B., et al. (2017). Parkinson's disease associated with pure ATXN10 repeat expansion. *njp Parkinson's Disease*, 3:27.