Evaluation of Nanopore Sequencing for Measuring Repeat Expansions in Genetic Disease

David W. Mohr and Alan F. Scott

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Introduction

Approximately two dozen human disorders are known to be caused by repeat expansions. Most of these cause neurological (e.g., Huntington's disease, Fragile X syndrome, or ALS) or neuromuscular disease (e.g., the spinal muscular dystrophies) and the severity of the phenotype or age of onset can be related to the expansion size. Further, many repeat expansion diseases show "anticipation," the progressively earlier onset of disease in later generations due to an expansion of a "premutation" repeat in an asymptomatic parent. In some cases the expanded regions can be hundreds of nucleotides in length which makes them difficult to assay using Sanger or short-read technologies. In many cases specialized tests have had to be developed for each disorder. Further, simple repeats (e.g., CAG_n) can be difficult for polymerases to process. An ideal assay would be one that captures repeat expansions from large-insert unamplified genomic DNA and measures repeat numbers with reasonable accuracy. As part of the Oxford Nanopore MinIon Access Program we have been exploring the use of nanopore sequencing for this purpose.

Materials and Methods

Library Preparation:

We obtained a pCDNA3.1 cDNA clone with ~9Kb insert spanning the Huntingtin (HTT) gene (NM_002111.7) and fragmented to ~10kb using Covaris g-tube. Oxford library was made with 1ug input DNA using SEQ-MAP004 Genomic Library protocol. DNA was subjected to NEB PreCR treatment (cat# M0309), end repaired using NEB End Repair Module (cat# E6050), cleaned with Agencourt AMPureXP beads, and dA-tailed with NEBNext A-tailing module (cat# E6053). Both a leader molecule and hairpin adapter were ligated to the A-tailed DNA. This results in fragments with leader at both ends, or a hairpin and leader at either end. The library is then enriched using Life Technologies His-tagged beads (cat# 10103D), which capture hairpin fragments with bound enzyme only, thus increasing the amount of molecules that have tether and leader.

Oxford Nanopore sequencing:

12ul of library was loaded onto a MinION R7.3 flowcell and run for 24 hours using MinKNOW 0.48.2.12. Basecalling was performed with Metrichor v2.24.46566, r7.X 2D Basecalling rev 1.10. Oxford Nanopore sequencing is performed by recording changes in ionic current as DNA 5-mers pass through protein pores at the single molecule level. A high quality library produces template and complement sequences (referred to as '1D' reads) by reading the leader end of a molecule, progressing through the hairpin, and reading the complement strand. If a molecule has both template and complement sequences, a consensus is generated, producing a higher quality sequence (referred to as '2D' reads).

Data Analysis:

Metrichor fast5 files containing 2D reads were converted to FASTA format using poretools (1). Since our clone had a known length of 82 CAG repeats, we constructed a reference based on the pCDNA3.1 vector and the HTT gene with the full repeat sequence. Reads were aligned to reference using LAST(2), with the following parameters: -s 2 -T 0 -Q 0 -a 1. Unique alignments were kept with last-map-probs.py and converted to BAM format using maf-convert.py and samtools. We then extracted reads that spanned the repeat region and ran blastn against the constructed reference to evaluate sequence identity and repeat expansion length.

Sanger Validation:

HTT clone was sequenced with T7 primer on an ABI3730 using standard reaction conditions. Since the repeat expansion is close to the priming site we were able to confirm the CAG repeat length of 82.

References:

1. Loman NJ, Quinlan AR: Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics (2014); 30 (23): 3399-3401.
2. Kiełbasa SM, Wan R, Sato K, *et al.*: Adaptive seeds tame genomic sequence comparison. *Genome Res.* 2011; **21**(3): 487–493

Results

Gene	MIM #	Disorder	Repeat	Normal	Premutation	Pathogenic
HTT	613004	Huntington's disease	CAG	10 to 35	27 to 35	36 to 120
FMR1	309550	fragile X	CGG	5 to 40	~55 to 200+	>200
ALS/ c9orf72	614260	amyotrophic lateral sclerosis	GGGGCC	<=23	>30	up to 250
DMPK	605377	myotonic dystrophy	CTG	5 to 37	50 to 150	up to 1000
FXN	606829	Friedreich ataxia	GAA	5 to 30		70 to 1000
ATXN10	611150	spinocerebellar ataxia 10	ATTCT	10 to 29		400 to 4500
ATXN7	607640	spinocerebellar ataxia 7	CAG	7-17		38-130
ZNF9	116955	myotonic dystrophy 2	CCTG	~7	22-33	75-11,000
ATN1	607462	dentatorubro-pallidoluysian atrophy	CAG	7-23		49-75
ATXN2	601517	spinocerebellar ataxia 2	CAG	14-31	27-33	34-57
ATXN8 ATXN8OS	613289 603680	spinocerebellar ataxia 8	CAG	~11-75		75-116
NOP56	614154	spinocerebellar ataxia 36	GGCCTG	5-14		~650-2500
CSTB	601145	progressive myoclonic epilepsy 1A	CCCCGCCCCGCG	2-3		50-75
AR	313700	spinal and bulbar muscular atrophy, X-linked 1	CAG	10-36		38-72

Table 1. Selected repeat-expansion disorders.

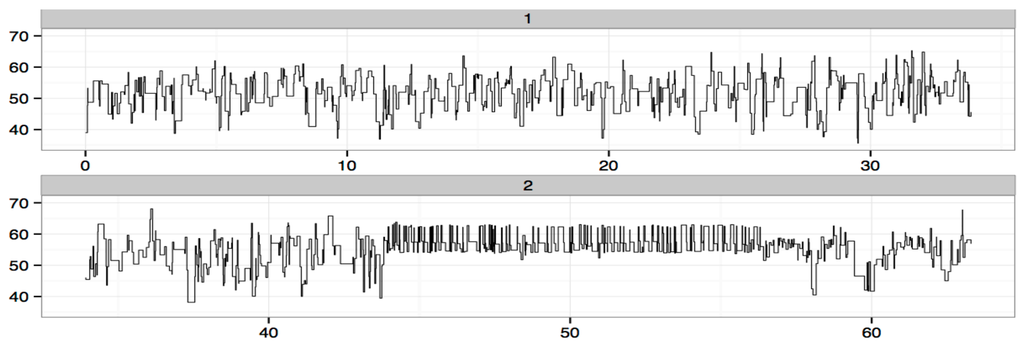

Figure 1. Squiggle Plot for read spanning CAG repeat region. The X-axis is time in seconds, the Y-axis is mean signal in picoamps. The CAG repeat expansion is clearly visible in the raw data at approximately 44 seconds to 56 seconds. Plot generated with poretools(1).

>channel_208_read_6_twodirections ./clone_3000/9020Dell7_150115_clone_3000_2_0701_1_ch208_file6_strand.fast5
CTAATTAATGGCTAATATAGGAGATTATACCAATAGTACGACGGGTTGCTATCCAGCTTGGTACGAAGGCTCGATTGGGCCTCAGTCAAAT
GCATTCTGCAGATATCCAGCACAGTGGCGGGCTAGTTGATCACATGGCGATCCTGGAAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCT
TC**CAGCAGCAGCAGGCAGCAGCAGCAGCAGCGAGCAGCAAGGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCG**
CAGCAGCAGCAGCAGCAACCGGCATAGCAGCAGCAGCAGCAGCGCAGCAGCAGCAGCAGCAAGGCAGCAGCAGCAGCAGCAGC
AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCGC
CACCGCCGCCGCCGCCGCCGCCTCCGCATTCCTCAGCTTCCTCAGCCGCCGCAGGCACAGCCGCTTGCTTCTTCATCATCGCACTCCCCCGCC
GCCGCCCCGCCACCCGGCCCCGGCTTGGCTGGGGGAGCGCTGCACTACCAATAAAGAAAGAATTTCAGCTACCAAGAAGACCGTGTGAAT
CATTGTCTGACAATATGTGAAAACAGCTTGGCACAGTCAGAAATTCTCCAGAAATTTCAGAAACTTCTGGGCATCGCTATGGAACTTTTCTG
CTGTGCAGTGATGACGCAGAGTCAGATGTCAGGATGGTGAGTACGAATGCCTCAACAAAGTTATCAAAGCTTTGATGGATTCTAATCTTCC
AAGTACAGCTCGAGCTCTACAAGGAGATTAAGAAAAACGGCGCACCGCGGCCAAGCCTTCGCGCTGCTCTGTGGCGTTTTGCAGACTCGC
ACCTGGTGCTATCCGTGGCCAGAAATGGTCGCCCGTACCTGGCTAACCTGCTGCCGCTGCCTCACCCGTACCAGCAAGCGTCCGGAAGAGT
GATCCGTTCAGGAGACTCTGGCAGCGAGGCTGTGCCAAAGTGATCATGCTCCCATTCGGTAATTGCGAATGTGACAACGGAGATTAAGGT
GCTGCTGAAAGCATTCATTGCTAACCTGAAGTCATCTTCTCCAAACCATCCGTCGCACTGCGGCGGGTTCCGTCCCCGGTCAGCATCTGTC
AGCACAGCCGTACTCAGTACTTTTACTCATGGCTGCTGAATGTCCTGCTGGGTCTGCTGGTCCCAGTGGAGGATGAACACTCCACCCTGCTG
ATTCTCGGCGTTCTGCTGACGCTGCGTTATCTGGTTCCGCTCCTGCAACAGCAGGTGAAGAGACATCGCACTGAACTCTTTTGGCGTTACTG
GCGCAAAGAAATGGAGGTTTCCCCGTCCGCTGAGCAGCTGGTGCCATTATGAACTGACCCTGCACCACTCAACATCAGTACCACATACGAA
CGTCGTCACCGGCGCACTGAAGCTTATCCAGCATATGTTCGCACGCCTCCTCCAGAGCTCCTGCATACTCTGACTGCCGTGGTGGATTGGCC
AACTCACCGCTGCCAAAGAAGAATCTGAGGTCGCTCCCGTTCTGCTCTAATCCGGGTGATTGCGGGCGGGCAGCAGGTGCAGCCCAGTCCT
CTCCCGGTAAGCAGAAAGGTAAAGTCTCTGCTGTGTGAGGAAGAAGCGCTGGAGACTAATTCCCGAGTTACCATTCGTTCCATACGTTTCC
ATAGACTAGGTAACTAGGTCTAGTTAAGGATACCATTTGCGGTATAGTAGGGCGGCTTCTTCTGGTGTCACCGCTGGTTCGATAGAGTAAG
TCATCATCACCGAACAGCCACGTTCACAGCACACCCTGCAGGCCGATAGCGTGATGGGCTTCGTGAAATCTGACGTTAGGCAACCAAAGGG
TTAACCAGAAGACATCGCGTATAGAGATAGCCTAGGTATTAGATTCGCATACCTGCAGGGTGGGTACCGTGGCAGTACCTTAACAACGGGT
AGCCTATTTAGCGATTAGCATTCTTTGGGAGAATACTACAGGGCCCAGACATAGTGGTGGGAGCGATTTACTCTGAAATGTGTGTACGGTA
CTACAGATAACCATGTACTGGGTCTGCAGATTGGTCAATATAAGATGAAGACGAAGAGGCAACCGGTATTCTGCCAAGTAGGCAGCGTCT
GAGGCTTTCGATTAGCTCTGGCCTGCGGAGCAGCACTAATACTTAGCAAAACATGTCTACATTGCGCTTTACCATCCATAACAGCTCTGTAA
GCGTGCCTACGTTATACGATTATGGGGATAACAGAATAGTGAGGAAGGAGTAAACCTTGTCACATGTAGATATAATACGGTCAGAGCGCC
CATAGATGAGCCCACTCGATATGCATTGCGTGCGTCTTAGTTGAGCTTCCCTTTCTATAAACTTAGGTGAGATCAAAACCTTGATGGTCCCG
ACTTAGTCGTGCGTGTATCCGATCAAACTATAGTGCTAGCTACTGCGTAGTGGATAGTCCTCCGCCTTGAGGCCTGGGATCTTAAAGTCTAC
ATCTCTGTGTAAATGGTCAACAGATGACATAGGAGTATCTGAAACTAATACCGTCTGGGTTCTATAAACCCACTTAACACGGTATCTACTAC
CCGGTGTCGTGACAGCCGCAATCCTGTGCGCGGGATGATCTGTAGCATTCTGTCACCGCTCGTTTCCACGTCACCGATTGGGGATGGGTCT
ATCCGTGTAACATGGATCGAAGCCTTTTCGGGCGACTTATCCCACTGCTGCCACAAAACCCTTGAAAGATGAGTCCAGCGTTACGTCCAAAA
TACCCGCCTGTATTATGTGCCCCGTCCGTAAATATTGTGTCATGAGCCTGTGCCGACTTCTATAGCGAACTGGGCCTGCGACTACCATCGTA
GAGTGTTCTGACTCTGCGTAACCACTCTATTGGCTCGTTCGGAGCATGAGGTGGGCGAATAGGGTCGGACAAGA

Figure 2. FASTA sequence for 2D read that spans CAG repeat in HTT gene. Matching 5' and 3' flanking sequence is underlined.

Discussion

Our ultimate goal was to obtain long reads that span the repeated regions and flanking sequence for all known repeat expansion disorders and to explore the genome for new repeats that may be functionally relevant. In these experiments we initially attempted streptavidin pull down of the HTT repeat using ~90bp biotinylated oligomers in flanking genomic sequence. As we did not obtain sufficient DNA to generate a library, we opted to test a plasmid containing the full length HTT cDNA made from a patient with 82-CAG repeats, which resulted in useable data (figure 1, 2). With rapid improvements in both library prep protocols and nanopore sequencing chemistry, we believe these DNA input limitations are surmountable without PCR, particularly if sample tethering methods that concentrate template at the membrane surface are developed and improved.

Read Name	Blast Identity (read spanning repeat)	Gaps (repeat only)	Mismatches (repeat only)	Repeat Count	matching flanking Sequence
ch262_file12_strand.fast5	71%	11	8	82	5',3'
ch284_file2_strand.fast5	85%	7	8	82	5',3'
ch321_file5_strand.fast5	87%	10	18	82	5',3'
ch366_file5_strand.fast5	70%	0	0	Read begins in repeat region	3'
ch372_file2_strand.fast5	87%	1	5	Read ends prior to repeat end	5'
ch40_file3_strand.fast5	84%	7	16	82	5', 3'
ch422_file24_strand.fast5	85%	2	5	32	5'
ch456_file2_strand.fast5	81%	25	9	74	5'
ch483_file0_strand.fast5	87%	6	5	doesn't span entire repeat	3'
ch79_file3_strand.fast5	70%	27	3	74	5',3'
ch8_file0_strand.fast5	70%	18	3	77	5',3'
ch120_file5_strand.fast5	71%	18	14	77	3'
ch208_file6_strand.fast5	77%	9	3	84	5',3'

Table 2. blastn results for reads that mapped to HTT gene with 82 CAG repeat expansion. Reads with high identity and matching flanking sequence correctly size the repeat expansion. Reads with lower identity and matching flanking sequence can size repeat expansion using pattern matching.

Conclusion

We believe that repeat expansion identification and characterization is an ideal application for nanopore technologies, and have demonstrated the technical ability to span long trinucleotide repeats using the MinION. Provided there is matching sequence flanking the repeat, the HTT repeat expansion can be correctly resolved without mapping to reference(table 2). Extremely high base quality is not necessary since it is the total number of repeats that are to be counted and small sequence errors are tolerable. Second, standard methods for repeat counting are difficult and not easily applied at a genome-scale level. Short read technologies, including synthetic long reads, cannot be used for repeat targets exceeding ~200-300 bp. At present, the variability in library preparation and/or flow cell quality appears to be the biggest hurdle in our hands. Further efforts include resolving repeats of different size k-mers and larger disease repeats. We look forward to a time when the technology is sufficiently mature to handle whole genomes with individual molecules of 50+ kb.