Accurate and simple FXN-GAA repeats (Friedreich ataxia loci) estimation by long read targeted sequencing

Pooja Sharma^{# MSc, 1,2}, *BharathramUppili*^{# Btech, 1,2}, *Istaq Ahmad*^{PhD, 3}, *Shweta Sahni*^{Msc. 3},
Vivekananda A^{MSc, 1,2}, *Mohammed Faruq*^{MBBS, PhD 1,2*}

¹Genomics and Molecular Medicine, CSIR-Institute of Genomics and Integrative Biology (CSIR -IGIB), Mall Road, Delhi 110007, India

²Academy for Scientific and Innovative Research, Ghaziabad, 201002, India

³Neurology Department, Neuroscience Centre, All India Institute of Medical Sciences, New Delhi 110029, India.

Equal contribution

E-mail: faruq.mohd@igib.res.in/faruq.mohd@igib.in

Keywords : Friedreich ataxia, Long read sequencing, Trinucleotide repeats, GAA repeats, FRDA, NGS

Abstract

Friedreich ataxia, an autosomal recessive disorder is caused by tandem GAA nucleotide repeats expansion in intron 1 of the FXN (frataxin gene). The GAA repeats above 66 in length are considered as pathogenic and commonly occurring repeats are 600-1200. Clinically the spectrum of the features is confined mainly to the neurological tissue, however, cardiomyopathy and diabetes mellitus has been observed in 60% and 30% of the subjects. The accurate detection of GAA repeats count is of utmost importance for clinical genetic correlation as no study has attempted an approach which is of high throughput nature and defines the sequence of GAA repeats. Largely, the method for detection of GAA repeats so far is either conventional PCR based screening and southern blot which is the gold standard method. We describe for the first time a method of long read sequencing wherein we utilized approach of long range targeted amplification of FXN-GAA repeats and sequencing on oxford MinION platform. We were able to achieve the successful amplification of GAA repeats ranging from 180-1200 at 250x coverage. The total throughput achievable for 96 samples can be less than 24 hours on one flow cell as per our protocol and is scalable and deployable at clinical day to sequencing.

INTRODUCTION

Friedreich ataxia (FRDA) is an autosomal recessive disorder that affects neurological (spinal cord, sensory nerves and cerebellum) and extra-neuronal systems (heart and pancreas). Disease develops more often at below 25 years of age. The causal mutation is a hyperexpansion of GAA repeats within intron 1 of the FXN gene on chr9q21.11. The GAA repeat length is variable varies from in healthy subjects and is observed between 5-33 in healthy subjects, while alleles greater than 66 to 1300 repeats represent full length pathogenic alleles (1,2). The commonest GAA repeat length observed in patients is ~600-1200. The cardinal clinical symptoms and signs of FRDA include are gait ataxia, dysarthria, peripheral neuropathy, kyphoscoliosis, pes cavus, however, only ~60% of the patients show hypertrophic cardiomyopathy and nearly one third of patients manifest diabetes mellitus. Nearly 25% of all FRDA cases do not subscribe to the essential diagnostic criterion for age of onset before 25 years as defined by Harding (3), and are referred to as the Late Onset FRDA (LOFA) (4). Atypical features in LOFA are defined by clinically milder disease phenotypes with a much slower disease progression, attributed to the size of the expanded repeat on GAA1 (5, 6).

The method of detection of FXN-GAA repeats is through conventional PCR, coupled with Triplet repeat-primed PCR which is helpful in screening; however, Southern blot is the gold standard method to detect the GAA length. The determination of precise and accurate GAA-length holds considerable great importance because the observed clinical variability in FRDA is largely unexplained at large. Some of the pertinent clinical and technical issues which warrants alternate and advanced tools to detect not only the GAA repeat length but also the repeat the interruption are, i) the determination of sequence configuration of GAA track would allow its correct interpretation if its purity of it can be ascertained vis-a-vis against the interrupted sequences, ii) the GAA sequence determination would further allow in understanding of the genotype-phenotype correlation, iii) it would allow us to decipher the relation between late onset Friedreich ataxia (LOFA) and interrupted GAA sequence, iv) southern blot is a low throughput technique for rapid and determination of GAA length. Sequence variation within the

GAA repeat region has been reported to affect the phenotypic variability in FRDA, especially in cases with late onset of disease (7). The aforementioned issues underscore the need for development and deployment of a novel approach to determine GAA length via next generation sequencing.

While short read sequencing for detection of repeat mediated disorders has already progressed much, long read sequencing as offered by the Oxford Nanopore technology (ONT) offers many unprecedented advantages over the former. Besides their rapidly shrinking costs and improved read quality, long read sequencing platforms provide additional maneuvering with respect to the bioinformatics analysis pipelines. This is of special significance in diseases attributable to complex expanded alleles with multiple repeat interruptions. Long read sequencing or 3rd generation sequencing technologies are a promising tool for querying the human genome for the assessment of long tracts of tandem nucleotide repeat expansion and over the last few years, several novel genes have been identified through long read sequencing technologies, i.e SMRT and oxford nanopore based platforms. LRSeq has led to the discovery of various new tandem nucleotide repeats (TNR) mutations which would otherwise not be detectable through short read sequencing methods. The discovery of these TNRs has opened up new avenues for solving the pathology of challenging neurological disorders, such as e.g. BAFME, NOTCH2NLC etc (8, 9).

The shortcomings of the existing approaches along with the unparalleled advantages offered by the ONT based long read sequencing makes it imperative to deploy this. It is very important now to use the same technology for accurate detection of TNR sequences for known disease loci. and Efforts made in this direction in the past have been successful for FMR1-CGG loci, HD-CAG etc.

To the best of our knowledge, no study has so far reported the accurate GAA length determination using a long read sequencing platform and the present study reports a comprehensive and simple protocol used for accurate determination of GAA repeats using nanopore technology. We were successful in estimating the length estimation of GAA repeat regions within the range of 200-1000 repeats.

METHODOLOGY:

The subjects included in the present study were recruited from our repository of DNA samples at the CSIR-Institute of Genomics and Integrative Biology. Genomic DNA was extracted using salting out method⁽¹⁾. For the initial phase, we screened the samples for the prevalent known ataxia subtypes (SCA1, SCA2, SCA3, SCA6, SCA7, SCA12, SCA17 and FRDA) using fluorescently labeled primers.^[2] The PCR product size (fragment length of each of the loci) were analyzed by capillary electrophoresis on ABI 3730XL DNA Analyzer (Thermo Fishers Scientific) and size estimation through Genescan software v 4.0. For FRDA screening GAA expansion detection was performed through triplet repeat primed PCR.^[3] *FXN*-GAA numbers were estimated by Long Range PCR using primers (FP-5'GGAGGGATCCGTCTGGGCAAAGG^{3'}; RP-5'CAATCCAGGACAGTCAGGGCTTT^{3'}) with KOD plus enzyme (Kit No 201 200U, 200 Reaction TOYOBO Com. Ltd Japan).^[4] Long Range PCR was performed with initial denaturation at 94°C for 3 minutes. Two step PCR of 20 seconds denaturation at 94°C and 2.30 minutes annealing and extension at 68°C in addition with 15 seconds increase per cycle for total 22 cycles. The final extension was carried out at 68°C for 10 minutes. PCR amplicons were checked on 1% agarose gel with 1kb DNA ladder (Promega) and stored at 4°C. The amplified products were purified using ampure beads (beckman coulter) in 1.6x ratio by volume of the PCR product.

Library preparation and sequencing:

The purified amplicons with a length of 1.4kb to 4.5kb were sequenced by using Oxford Nanopore Sequencing platform. Around 400ng of purified long range amplicons were taken as input for sequencing. The library was prepared following Nanopore Native barcoding genomic DNA with EXP-NBD196 and SQK-LSK109 protocol according to the latest sequencing version: NBE_9121_v109_revF. The DNA amplicons were end repaired and dA-tailed using NEBNext FFPE DNA Repair Mix (New England Biolabs (NEB), Ipswich, USA and NEBNext End repair/dA-tailing (NEB). Native barcodes were ligated to the end repaired DNA with NEB Blunt/TA Ligase Master Mix (SQK-LSK109). Adaptor ligation to the pooled barcoded library was performed using NEBNext® Quick

Ligation module ((NEB E6057). The final library was purified with ampure beads in 0.4x ratio of the reaction volume and Oxford Nanopore short fragment buffer (SFB- SQK- LSK109). The prepared library of around 70ng was sequenced using R9.4.1 flow cell on MinION Mk1C for around 16 hours.

Sequence Analysis

Using Guppy basecaller (version 3.5.2) the raw fast5 files were demultiplexed into different fastq files. The alignment was performed with minimap2 3 (v 2.1) with FXN gene sequence (GRCh37 - chr9:71650479-71693993) as the reference (13, 14). The bam file is sorted and indexed using samtools 2 (v 1.9). For finding the repeat expansion inside the sample we have used STRique (14). Initially, Indexing of the fast5 was performed to tabulate all the fast5 reads from all the samples, later using the HMM model as given by the STRique tool (r9_4_450bps.model) the repeat count per each read for all the samples had been calculated. Using ggplot2, dplyr packages from R the data has been visualized.

RESULTS and DISCUSSION

GAA repeat amplification through long range-PCR

The long range amplification using flanking primer sequence has amplified sample specific PCR products of varying length ranging from 1,500 base pair - 4000 base pairs (Fig.1). The PCR products of length ~1.5 kbp are expected to carry small normal GAA repeats (7-20) while fragments of 2.2 kb were expected to carry ~200 GAA, likewise, fragments of size ~4kb corresponds to GAA ~1000 repeats. The long range amplification of the GAA flanking region was efficient in capturing all the expected length of abnormal GAA repeats.

Targeted GAA flanking region sequencing on MinION flow cell

We sequenced four representative PCR products from four subjects (patient1, patient2, patient3 and a healthy subject) with expected GAA (allele1/allele2) length of 200/1000,

300/1000, 800/800 and 6/6 respectively. We obtained a total of 140k reads for four samples (average 23k reads/sample that mapped to the target region) with mean coverage ~260x (165-350X). The data QC showed the max sequenced reads of 4kb length with some non-specificity off target longer products, the percent GC content of 42-45 and read mapping percentage ranging from 59% to 66%. (Fig.2)

GAA repeat sequence determination

Using STRique, the length estimation of GAA showed for patient-1, biallelic expansions mutations with highest peak at 250 bp (allele1) and 1070 (allele2); for patient-2, allele1/allele2 (250/1100) and patient 3 (525/830). (Fig.3). the squiggle (series of current pattern in GAA repeat region was detected along with flanking non-repetitive region (fig.4).

Conclusion

Despite their important role in determining the clinical variability in FRDA, GAA repeat structure investigation has so far precluded large-scale screening. Long-read sequencing with its promising ability for identification of causative repeat expansions, discernment of repeat configuration, repeat length estimation and detection of base modifications, is aptly positioned to overcome these challenges. For the first time we demonstrate the ability of long read ONT sequencing in accurately determining the size of an expanded intronic region of upto 1000 GAA repeats in FRDA patients.

Acknowledgement

We acknowledge the funding from CSIR-IGIB, OLP1120 and ICMR funded project GAP240. We thank sincerely to the patients and the families for their participation.

Reference

1. Campuzano V, Montermini L, Moltò MD, et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*.1996;271(5254):1423-1427.
doi:10.1126/science.271.5254.1423
2. Dürr A, Cossee M, Agid Y, et al. Clinical and genetic abnormalities in patients with Friedreich's ataxia. *N Engl J Med*. 1996;335(16):1169-1175.
doi:10.1056/NEJM199610173351601
3. Harding AE. Friedreich's ataxia: a clinical and genetic study of 90 families with an analysis of early diagnostic criteria and intrafamilial clustering of clinical features. *Brain*. 1981 Sep;104(3):589-620
4. Dürr A, Cossee M, Agid Y, Campuzano V, Mignard C, Penet C, Mandel JL, Brice A, Koenig M. Clinical and genetic abnormalities in patients with Friedreich's ataxia. *N Engl J Med*. 1996 Oct 17;335(16):1169-75
5. De Michele G, Filla A, Cavalcanti F, Di Maio L, Pianese L, Castaldo I, Calabrese O, Monticelli A, Varrone S, Campanella G, et al. Late onset Friedreich's disease: clinical features and mapping of mutation to the FRDA locus. *J Neurol Neurosurg Psychiatry*. 1994 Aug;57(8):977-9.
6. Ragno M, De Michele G, Cavalcanti F, Pianese L, Monticelli A, Curatola L, Bollettini F, Coccozza S, Caruso G, Santoro L, Filla A. Broadened Friedreich's ataxia phenotype after gene cloning. Minimal GAA expansion causes late-onset spastic ataxia. *Neurology*. 1997 Dec;49(6):1617-20
7. McDaniel DO, Keats B, Vedanarayanan VV, Subramony SH. Sequence variation in GAA repeat expansions may cause differential phenotype display in Friedreich's ataxia. *Mov Disord*. 2001 Nov;16(6):1153-8.
8. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, Sugano S, Qu W, Ichikawa K, Yurino H,

- Higasa K, Shibata S, Mitsue A, Tanaka M, Ichikawa Y, Takahashi Y, Date H, Matsukawa T, Kanda J, Nakamoto FK, Higashihara M, Abe K, Koike R, Sasagawa M, Kuroha Y, Hasegawa N, Kanesawa N, Kondo T, Hitomi T, Tada M, Takano H, Saito Y, Sanpei K, Onodera O, Nishizawa M, Nakamura M, Yasuda T, Sakiyama Y, Otsuka M, Ueki A, Kaida KI, Shimizu J, Hanajima R, Hayashi T, Terao Y, Inomata-Terada S, Hamada M, Shiota Y, Kubota A, Ugawa Y, Koh K, Takiyama Y, Ohsawa-Yoshida N, Ishiura S, Yamasaki R, Tamaoka A, Akiyama H, Otsuki T, Sano A, Ikeda A, Goto J, Morishita S, Tsuji S. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet.* 2018 Apr;50(4):581-590.
9. Sone J, Mitsunashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, Kohno Y, Takiyama Y, Maeda K, Doi H, Koyano S, Takeuchi H, Kawamoto M, Kohara N, Ando T, Ieda T, Kita Y, Kokubun N, Tsuboi Y, Katoh K, Kino Y, Katsuno M, Iwasaki Y, Yoshida M, Tanaka F, Suzuki IK, Frith MC, Matsumoto N, Sobue G. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet.* 2019 Aug;51(8):1215-1221
 10. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 1988;16(3):1215. doi:10.1093/nar/16.3.1215
 11. Warner JP, Barron LH, Goudie D, et al. A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet.* 1996;33(12):1022-1026. doi:10.1136/jmg.33.12.1022
 12. Ciotti P, Di Maria E, Bellone E, Ajmar F, Mandich P. Triplet repeat primed PCR (TP PCR) in molecular diagnostic testing for Friedreich ataxia. *J Mol Diagn.* 2004;6(4):285-289. doi:10.1016/S1525-1578(10)60523-5 1.
 13. Li, Heng et al. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics (Oxford, England)* vol. 25,16 (2009): 2078-9. doi:10.1093/bioinformatics/btp352

14. Li, Heng. "Minimap2: pairwise alignment for nucleotide sequences." *Bioinformatics (Oxford, England)* vol. 34,18 (2018): 3094-3100.
doi:10.1093/bioinformatics/bty191
15. Giesselmann, P., Brändl, B., Raimondeau, E. et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* (2019) doi:10.1038/s41587-019-0293-x

Figure legend

Figure1: Long range-PCR base detection of GAA repeats and flanking region

Figure2: Quality parameters of the reads obtained through nanopore sequencing showing read length and % GC content.

Figure3: Histogram showing length wise GAA repeat distribution across four samples. The variability in length is indicative of somatic mosaicism

Figure4: Current series pattern of representative reads in the form of squiggle for each of the samples.

1. **patient1:- 270/1004**
2. **patient2:- 287/1037**
3. carrier.1:- 6/ 120
4. carrier.2:- 6/287
5. carrier.3:- 6/370
6. carrier.4:- 6/370
7. carrier.5:- 6/204
8. carrier.6:- 6/204
9. carrier.7:- 6/287
10. carrier.8:- 6/ 1370
11. carrier.9:- 6/ 1070
12. carrier.10:- 6/ 454
13. **Healthy:- 6/ 6**
14. Healthy:- 6/ 6
15. **patient3:- 870/870**
16. patient4:- 537/ 870
17. L- 1000bp ladder

FastQC: Sequence Length Distribution

FastQC: Per Sequence GC Content

GAA-264

GAA-1073

GAA-289

GAA-1111

GAA-542

GAA-828

GAA-6

GAA-8