Clinical Validation of Short Tandem Repeat Expansion Variant Calling Using Short-Read WGS Data

Andrea Oza[1], Diana M. Toledo[1], Marina DiStefano[1], Areesha Salman[1], Katherine Lafferty[1], Mark Fleharty[1], Edyta Malolepsza[1], Ning Li[1], Alyssa Friedman[1], Samuel DeLuca[1], Teni Dowdell[1], Justin Abreu[1], Fontina Kelley[1], Grace VanNoy[1], Ben Weisburd[1], Heidi Rehm[2,1], Niall Lennon[1]

1 The Broad Institute of MIT and Harvard, Cambridge, MA, USA
2 Massachusetts General Hospital, Boston, MA, USA

Andrea Oza
oza@broadinstitute.org

Broad Clinical Labs
genomics@broadinstitute.org

Introduction

Expansions of short tandem repeats (STRs) are a type of genetic variant that are responsible for many neurological disorders including Fragile X syndrome, Friedreich ataxia, Huntington disease, and myotonic dystrophy. Testing for short tandem repeat (STR) expansions previously required specific assays such as triplet-repeat primed PCR or Southern blot.

Recently developed computational tools, such as ExpansionHunter, can now identify STR expansions using PCR-free short-read whole genome sequencing (WGS) data to genotype STR loci.[1] This allows for a single test to be run to identify several different variant types including small variants (i.e. - single nucleotide variants, small indels), copy number variants, and repeat expansions, and can potentially avoid a diagnostic odyssey for patients with rare disease.

The Rare Genomes Project (RGP) at the Broad Institute has tested and implemented ExpansionHunter to identify variants responsible for rare and undiagnosed conditions. To date, 10 individuals from the RGP with features of ataxia, myopathy, or muscular dystrophy and onset ranging from birth to 81 years have been diagnosed with an STR expansion in 6 different genes (*NOTCH2NLC*, *ATXN1, ATXN2, RFC1, CNBP, PABPN1*).

Patient	Age of onset	Phenotype	STR Locus	Disease	EH Genotype	Inheritance	Repeat Unit
1	47	Muscle weakness	NOTCH2NLC	NIID	32/88	AD	GGC
2	60	Ataxia	ATXN1	SCA1	29/41	AD	TGC
3	51	Myopathy	ATXN2	SCA2	22/33	AD	GCT
4	60	Ataxia	RFC1	CANVAS	78/78	AR	AAGGG
5	78	Myopathy	CNBP	DM2	16/469	AD	CAGG
6	61	Myopathy	PABPN1	OPMD	6/9	AD	GCG
7	40	Myopathy	CNBP	DM2	16/402	AD	CAGG
8	Birth	Hypotonia and chorea	ATXN2	SCA2	27/96	AD	GCT
9	25	Muscular dystrophy	RFC1	CANVAS	15/416	AR	AAGGG
10	60	Muscular dystrophy	CNBP	DM2	15/464	AD	CAGG

Table 1. Rare Genomes Project and short tandem repeats identified using ExpansionHunter

Success of ExpansionHunter in RGP has led the Broad Clinical Laboratories (BCL) to validate Illumina DRAGEN ExpansionHunter (EH) for clinical WGS testing. The product offerings include technical deliverables (.bam, .vcf file, and index files) and/or an interpretive clinical report.

Materials & Methods

The validation cohort includes 22 DNA samples sourced from the Coriell Institute with known STR expansions in six genes (FMR1, ATXN1, FXN, HTT, C9ORF72, and DMPK) that vary by repeat size, motif, inheritance pattern, and patient sex. Coriell sizing was performed by Southern blot and/or PCR analysis. The normal, premutation, and full expansion repeat ranges were determined for each gene, along with a cutoff flag that would be used to flag potentially expanded alleles (Table 2).

Gene	Region	Repeat Motif	Normal Range	Pre-mutation range	Full Mutation Range	Cutoff flag
FMR1	5′ UTR	CGG	5-44	55-200	>200	45
ATXN1	Exon	CAG	≤35	36-38	≥39	36
FXN	Intron	GAA	5-33	34-65	>66	34
HTT	Exon	CAG	≤26	27-35	>36	27
C9ORF72	Intron	GGGGCC	2-24	Unknown	61 to >4000	25
DMPK	3′ UTR	CTG	5-34	35-49	>49	35

Table 2. Genes included in the validation cohort, with the repeat motif, and the normal, premutation, and full mutation repeat ranges. Grey zones/Intermediate alleles are not shown. A cutoff flag was set to identify samples with potential expansions.

Samples were run on Illumina NovaSeq 6000 system and called using DRAGEN v3.10.4 ExpansionHunter. Performance was analyzed for each sample by comparing the EH genotyping call in the repeat VCF to the truth data from Coriell. Repeats smaller than the sequencing read length (~150bp) were expected to be accurately sized (+/-1 repeat) using EH; however, the size of STRs that are ~150bp or longer are likely underestimated by EH.[2] This was accounted for by allowing STRs >150bp to pass validation if the predicted repeat size is within +/-1 repeat **or** if the size of the repeat correctly correlates to the gene's normal, premutation, or pathogenic repeat range. To assess intra-run repeatability, three replicates of the same sample run in the same batch (same technologist, same reagents, same sequencing flow cells) will be reviewed for concordance of the EH calls.

Results

- Accuracy (+/-1 repeat) (Fig 1).
  - 31.6% of samples were within 1 repeat of the actual repeat size (Fig. 1)
  - All alleles smaller than 44 repeats (<132bp) were called accurately, EH underestimated the size of alleles >56 repeats in length (>168bp)
- Sizing by Classification (Fig. 3)
  - EH had 92% concordance. 3 discordant calls (two full mutations, one premutation)
- All expansions would be identified using the proposed Cutoff flag (Table 3)

Fig. 1: EH Repeat Number Accuracy with Truth Data (+/-1 Repeat)

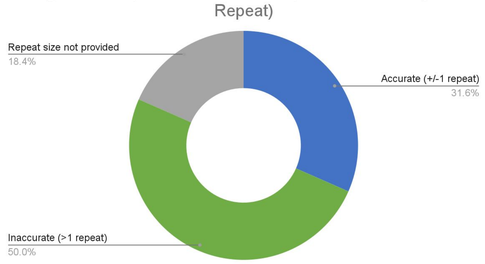

Figure 1. Repeats <150 basepairs were called accurately within +/1 repeat. Repeats >150 basepairs in size were all inaccurately called, and were typically undercalled by EH. N/A indicates that a specific repeat size was not provided by Coriell, as the expansions in C9ORF72 were only described as "expanded", and two of the normal alleles in DMPK were not described (see details in Table 3).

Fig. 2a: EH Repeat Size vs. Coriell Repeats (<150bp) Size

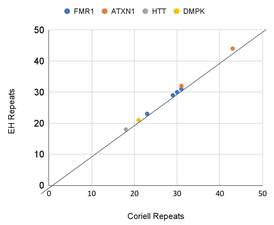

Fig. 2b: EH Repeat Size vs. Coriell Repeats (>150bp) Size

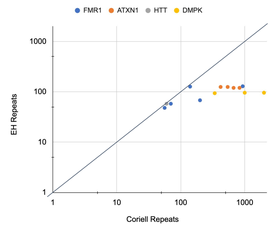

Figure 2. EH repeat number compared to the number reported by Coriell. 2a) Samples with Coriell repeats <50 (<150bp) in length. 2b) Samples with Coriell repeats >50 (>150bp) in length, using a log scale for x and y. Repeats >150bp are consistently undercalled by EH.

Sample #	Sex	Coriell Repeat Information: Gene	Repeat Motif	Allele 1 Repeat #	Allele 1 Repeat Class.	Allele 2 Repeat #	Allele 2 Repeat Class.	ExpansionHunter Repeat Calls: Allele 1 Repeat #	Allele 1 Repeat Class.	Allele 2 Repeat #	Allele 2 Repeat Class.	CI (Allele 1/Allele 2)
1	F	FMR1	CGG	23	Normal	30	Normal	23	Normal	30	Normal	23-23/30-30
2	F	FMR1	CGG	23	Normal	70	Premutation	23	Normal	**58**	Premutation	23-23/52-78
3	F	FMR1	CGG	23	Normal	95-12-140	Premutation	23	Normal	**127**	Premutation	23-23/92-154
4	F	FMR1	CGG	28-29	Normal	>200	Full Mutation	29	Normal	**68**	**Premutation**	29-29/59-104
5	M	FMR1	CGG	N/A	N/A	23	Normal	n/a	n/a	23	Normal	23-23
6	M	FMR1	CGG	N/A	N/A	56	Premutation	n/a	n/a	**48**	**Normal**	**48-48**
7	M	FMR1	CGG	N/A	N/A	117	Premutation	n/a	n/a	**80**	Premutation	**62-113**
8	M	FMR1	CGG	N/A	N/A	931-940	Full Mutation	n/a	n/a	**129**	**Premutation**	**93-187**
9	F	FMR1	CGG	29	Normal	31	Normal	29	Normal	31	Normal	29-29/31-31
10	F	ATXN1	CAG	31	Normal	43	Full Mutation	32	Normal	44	Full Mutation	32-32/44-44
11	M	FXN	GAA	420	Full Mutation	541	Full Mutation	**125**	Full Mutation	**125**	Full Mutation	66-135/85-169
12	F	FXN	GAA	830	Full Mutation	670	Full Mutation	**120**	Full Mutation	**120**	Full Mutation	61-125/79-159
13	F	HTT	CAG	18	Normal	60	Full Mutation	18	Normal	**58**	Full Mutation	18-18/52-68
14	M	C9ORF72	GGGGCC	ND	Normal*	ND	Full Mutation	2	Normal	167	Full Mutation	2-2/167-294
15	M	C9ORF72	GGGGCC	ND	Normal*	ND	Full Mutation	2	Normal	211	Full Mutation	2-2/193-320
16	M	C9ORF72	GGGGCC	ND	Normal*	ND	Full Mutation	8	Normal	187	Full Mutation	8-8/153-258
17	F	C9ORF72	GGGGCC	ND	Normal*	ND	Full Mutation	13	Normal	182	Full Mutation	13-13/181-343
18	F	C9ORF72	GGGGCC	ND	Normal*	ND	Full Mutation	2	Normal	294	Full Mutation	2-2/277-450
19	F	DMPK	CTG	21	Normal	340	Full Mutation	21	Normal	**94**	Full Mutation	**21-21/74-111**
20	M	DMPK	CTG	ND	Normal*	≤1000	Full Mutation	12	Normal	**96**	Full Mutation	**12-12/80-130**
21	M	DMPK	CTG	ND	Normal*	≤2000	Full Mutation	13	Normal	**98**	Full Mutation	**12-12/80-130**
Intra-run repeatability												
22_1	F	FXN	GAA	ND	Normal*	500	Full Mutation	**96**	Full Mutation	8	Normal	**8-8/80-137**
22_2	F	FXN	GAA	ND	Normal*	500	Full Mutation	**103**	Full Mutation	8	Normal	**8-8/87-157**
22_3	F	FXN	GAA	ND	Normal*	500	Full Mutation	**74**	Full Mutation	8	Normal	**8-8/64-106**

Table 3. Sample Details and EH Results. Each sample's allele repeat sizes and their corresponding classification from Coriell and EH are shown. C9ORF72 samples did not have a specific repeat size available and were described only as "expanded". Similarly, the normal allele sizes for samples 20 and 21 were not described. All alleles called by EH were identified by the proposed cutoff flag shown for each gene in Table 2. Class.: Classification, ND: Not described,*: Presumed Normal, N/A: Not Applicable, Y- Yes, N- No, CI: Confidence Interval. **BOLD** indicates inaccurate (>1 repeat difference) from Coriell data, or the actual repeat size was not contained within the EH confidence interval.

Fig. 3: EH Concordance with Actual Allele Classification

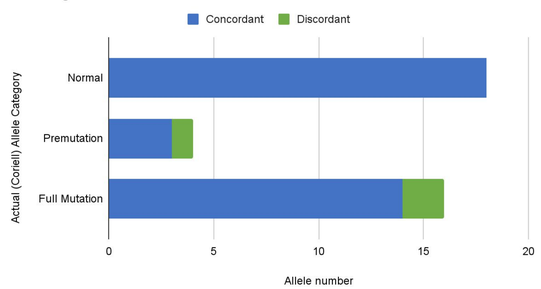

Figure 3. Three repeats fell within the incorrect allele classification range. All were in FMR1. One premutation was sized as a normal allele, and two full mutations were sized as premutations. However, all expanded alleles would be caught for clinical reporting using the flagging cutoff (Fig. 3)

Discussion

- As expected, sequencing read length (~150 bp) limited the ability for EH to call a repeat size accurately. All repeats below 150bp in length were called accurately (+/-1 repeat), whereas none of the repeats >150bp were accurately called.
- Determining the class of expansion (normal, premutation, full mutation) is also limited for expansions beyond ~150bp. Based on these data, clinical interpretive reporting using EH will need to rely on the flagging cutoffs to distinguish between normal and potentially expanded alleles. However, all of the loci in this validation had flagging cutoffs that were under the 150bp limit. However, this may not be the case for all loci included in the EH caller.
- When potentially expanded alleles are identified, further review either by visualization[3], and/or orthogonal confirmation will be required to provide accurate repeat lengths. This information is highly important to provide an accurate diagnosis with anticipatory guidance for patients with rare disease.

Conclusion

- EH is a powerful tool that allows for repeat expansions to be identified using PCR-free WGS data; however, its ability to accurately size repeats is limited by sequencing read length.
- EH can successfully and accurately identify expanded STRs when an appropriate flagging cutoff is set below the sequencing read length.
- Adding EH to clinical WGS will provide critical information to patients with phenotypes that could be due to repeat expansions. This could reduce the number of tests required for genetic diagnosis and shorten the diagnostic odyssey.
- Future improvements include transitioning to NovaSeqX, updating to the newest version of the DRAGEN pipeline, and expanding product offerings to include interpretive reporting for EH reporting are planned at BCL.
- Orthogonal confirmation is recommended for accurate sizing when an expanded STR is flagged above a set cutoff.

References

1. Dolzhenko E, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017 Nov;27(11):1895-1903. doi: 10.1101/gr.225672.117. Epub 2017 Sep 8. PMID: 28887402; PMCID: PMC5668946.
2. Ibañez K, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022 Mar;21(3):234-245. doi: 10.1016/S1474-4422(21)00462-2. PMID: 35182509; PMCID: PMC8850201.
3. Dolzhenko E, et al.. REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. Genome Med. 2022 Aug 11;14(1):84. doi: 10.1186/s13073-022-01085-z. PMID: 35948990; PMCID: PMC9367089.

Data used in this poster was generated at the Broad Institute. For more information please visit: https://broadclinicallabs.org/