HL7 FHIR Proposal for Repeat Expansion Variant Model

Version 2

Contents

1. Background

Repeat expansion, also known as repeat tandem, is characterized by polymorphic nucleotide sequences scattered throughout the human genome (***Paulson, 2019***). Repeat expansion is a well-known process that results in at least 50 known disorders, including Huntington's Disease (HD), Myotonic Dystrophy Type 1 (DM1), Myotonic Dystrophy Type 2 (DM2), and Oculopharyngeal muscular dystrophy (OPMD) (***Dpienne et al., 2021***).

The repeat nucleotides, or motifs, in a repeat expansion variant are relatively conservative and short, with a length ranging from 3 to 15 bp (microsatellites with 1–9 bp repeats; minisatellites with 10–99 bp repeats). The category of normal or pathological repeats strongly depends on the length of the repeat motif within genes (***Hannan et al., 2018***).

1.1 Repeat Expansion Disease Model

There are two different models of repeat expansion variants: simple model (with unique repeat) and mixed model (with mixed and complicated repeats).

1.1.1 Simple Repeat Expansion Disease Model

As the most well-known repeat expansion disease, Huntington's disease (HD) is caused by a CAG repeat expansion in the HTT gene. Repeat length can change over time, both in individual cells and between generations, and repeat length correlates with disease onset, which means longer repeats may drive pathology. The unusual CAG repeat expansion encodes a toxic **polyglutamine** tract which leads to pathogenic phenotype (***Keum et al., 2016***).

According to ACMG standard (***American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease, 2014***), laboratory reports are recommended to use the following definition of normal and mutation category for HTT repeat expansion variant (**Table 1**).

Table 1 CAG($_n$) repeat expansion category and descriptors of HTT

Allele Category	Repeats Range	Allele Example
Normal allele	<=26	CAG[25]
Mutable normal allele	27-35	CAG[35]
HD allele with reduced penetrance	36-39	CAG[39]
HD allele	>=40	CAG[40]

Each report must include **the CAG repeat numbers of both alleles** with the precision of sizing fulfilling the criteria recommended by the ACMG Biochemical and Molecular Genetics Resource Committee.

1.1.2 Mixed Repeat Expansion Disease Model

OPMD (OMIM #164300) is a rare disorder and it is caused by a short TRE (trinucleotide repeat expansion) in the first exon of the gene encoding for the polyadenylate-binding protein nuclear 1 (PABPN1) located on chromosome 14q11.1. In the wild-type PABPN1, the first methionine (ATG) is followed by a 10 alanine repeat (NM_004643: GCG[6]GCA[3]GCG[1], with both GCG and GCA encode alanine) (***Leeuw et al., 2019***), Thus OPMD is also known as **polyalanine** disease.

Pathogenic PABPN1 mutation was reported to either have an 11 to 18 total alanine length (compared to normal length 10) (***Brais et al., 1998***) or have abnormal GCG length (8-13) only (compared to normal GCG length 6) (***Grewal et al., 1999***) (**Table 2**). Different from HTT, the expansion length of repeats of PABPN1 did not correlate with clinical features based on above research.

Table 2 GCN($_n$) repeat expansion category of PABPN1

Allele Category	Repeats Example
Normal allele	GCG[6] GCA[3] GCG[1]
OPMD allele -1	GCG[8-13] GCA[3] GCG[1]
OPMD allele -2	GCG[6] GCA[4] GCG[1]

**Sample category from reported pathogenic cases, no standard for OPMD category released from ACMG. OPMD allele-1 is based on Brais's report and OPMD allele-2 is based on Grewal's report.*

1.2 Other Repeat Expansion Types

The location and length of repeat expansion may vary from gene to gene. Besides polyglutamine (e.g., HD) and polyalanine (e.g., OPMD) repeats, the location of repeat expansion can also be in non-coding region, including 5'UTR, 3'UTR or intronic loci (***Hannan et al., 2018***) (**Figure 1**).

Group of disorders	5′ UTR TRDs	Intronic TRDs	Polyglutamine TRDs	Polyalanine TRDs	3′ UTR TRDs
	• FXS • FXTAS • Other FX disorders	• FRDA • C9ORF72 TRDs (includes subset of ALS and FTD)	• HD • SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17 • SBMA (Kennedy disease) • DRPLA	• OPMD and eight other developmental disorders	• DM1 and DM2

Figure 1 Location of repeat expansion within genes for repeat expansion diseases (Hannan et al., 2018).

Currently identified repeat motifs usually range from 3 to 12 base pairs. Trinucleotides (3 base pairs) are the most commonly found repeats (***Paulson, 2019***) (**Figure 2**).

• CAG – at least 10 diseases (Huntington disease, spinal and bulbar muscular atrophy, dentatorubral-pallidoluysian atrophy and seven SCAs)

• CGG – fragile X, fragile X tremor ataxia syndrome, other fragile sites (GCC, CCG)

• CTG – myotonic dystrophy type 1, Huntington disease-like 2, spinocerebellar ataxia type 8, Fuchs corneal dystrophy

• GAA – Friedreich ataxia

• GCC – *FRAXE* mental retardation

• GCG – oculopharyngeal muscular dystrophy

• CCTG – myotonic dystrophy type 1

• ATTCT – spinocerebellar ataxia type 10

• TGGAA – spinocerebellar ataxia type 31

• GGCCTG – spinocerebellar ataxia type 36

• GGGGCC – *C9ORF72* frontotemporal dementia/amyotrophic lateral sclerosis

• CCCCGCCCCGCG – EPM1 (myoclonic epilepsy)

Figure 2 Repeat nucleotides within genes for its associated repeat expansion diseases (Paulson, 2019).

1.3 Repeat Expansion Name Convention

1.3.1 HGVS

According to the repeated sequence variant nomenclature recommendation released from HGVS (https://varnomen.hgvs.org/recommendations/DNA/variant/repeated), repeat expansion representations should use following format:

Simple model with unique repeat

"prefix""position_first_nucleotide_first_repeat_unit""repeat_sequence"["copy_number"]

g.123CAG[23]

Mixed model with complicated repeats

"prefix""range_repeated_sequence""repeat_sequence_1"["copy_number"]"repeat_sequence_2"["copy_number"]

g.123_191CAG[19]CAA[4]

1.3.2 GA4GH

In GA4GH, Repeat Expansion (include large sequence repeats) is defined in RepeatedSequenceExpression (https://vrs.ga4gh.org/en/stable/terms_and_model.html?highlight=repeated#repeatedsequenceexpression). The following is an expression of a sequence comprised of a tandem repeating subsequence. Besides type, repeat sequence has two more field, seq_expr (using sequence ID) and count (using type IndefiniteRange) (**Figure 3, Figure 4**).

Information Model

Field	Type	Limits	Description
type	string	1..1	MUST be "Repeated SequenceExpression"
seq_expr	Sequence Expression and NOT Repeated SequenceExpression	1..1	an expression of the repeating subsequence
count	Number | DefiniteRange | IndefiniteRange	1..1	the inclusive range count of repeated units

Figure 3 Representation of repeat expansion definition in GA4GH.


```
"count": {
  "comparator": ">=",
  "type": "IndefiniteRange",
  "value": 6
},
"seq_expr": {
  "location": {
    "interval": {
      "end": {
        "type": "Number",
        "value": 44908822
      },
      "start": {
        "type": "Number",
        "value": 44908821
      },
      "type": "SequenceInterval"
    },
    "sequence_id": "ga4gh:SQ.IIB53T8CNeJJdUqzn9V_JnRtQadwWCbl",
    "type": "SequenceLocation"
  },
```


Figure 4 Representation of repeat expansion example in GA4GH.

1.3.3 Laboratory Reports

In repeat expansion genetic lab reports (**Figure 5 and Figure 6**), the repeat expansion variant is usually reported within a table in allele-specific level. Repeated nucleotides (e.g., CAG in Figure 5, GGGGCC in Figure 6) and repeat number (33,22 in Figure 5 and 2,35 in Figure 6) are the required fields to be reported. The reference range is also attached as part of the report.

RESULT(S): INCOMPLETE PENETRANCE

Gene	Mode of Inheritance	Variant	Zygosity	Classification
ATXN2	Autosomal Dominant	Repeat Number: 33 Repeat Number: 22	Heterozygous Heterozygous	Incomplete Penetrance Normal

REFERENCE RANGE

Classification	CAG Repeat Size
Normal	30 or less
Recessive	31
Uncertain Significance	32
Incomplete Penetrance	33-34
Positive	35 or greater

Figure 5 Lab reports of ATXN2 gene with CAG Repeats.

Results:GGGGCC (G_4C_2) Repeat Number:

Repeat numbers are typically ±2, although slightly greater variation may occur when repeat numbers are greater than 55.

Allele 1 Repeat Number: 2

Allele 2 Repeat Number: 35

Figure 6 Lab reports of C9orf72 gene with GGGGCC Repeats.

1.3.4 Academic Research Publications

There is no standard about how to represent repeat expansion to report related academic/clinical findings. In some publications, expansion is reported in the format of $(RepeatMotif)_{repeatNumber}$, e.g $(CTG)_n \bullet (CAG)_n$ or $(CAG)_n/(CTG)_n$ (**Liu et al., 2012; Kim et al.,2016**). There is the most used format to represent repeat expansion in publication.

However, there are some publications which use different format, e.g $[CTG]_{\geq n}$ to represent repeat expansion (**Alfahli et al., 2004; Watkins et al., 1995**). Though there is no significant difference between different format (using ( ), [ ], •, or /), the lack of standard makes it hard to find a way to represent repeat expansion variant across all scenario.

1.4 Repeat Expansion in Epic

In Epic, a repeat expansion variant is defined as repeat-nucleotide/repeat-number pairs and stored as a list of these pairs. We can display repeat expansions like so (Figure 7).

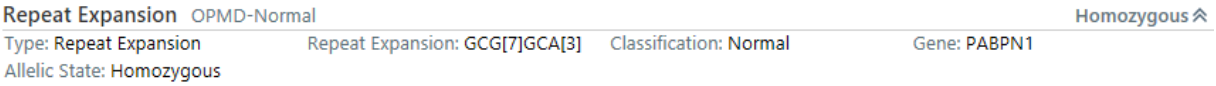

Figure 7 Display of a repeat expansion in Epic. The HGVS recommendation format (e.g., GCG[7]) is currently adopted in Epic for display.

2. Challenges

The Lab Results Interface (LRI) specifies that trinucleotide repeats, as well as the number of trinucleotide repeats, are out of scope of the HL7 Implementation Guide (**for details, see part 5.2.2**).

- Gene/chromosome fusions (and trinucleotide repeats), and similar studies that are also reported as simple lab tests whose quantitative results may be the number of blood cells containing a specified anomaly, the ratio of a marker gene, or the number of trinucleotide repeats, and are accommodated by existing LOINC codes.

Therefore, we need to design a model to represent repeat expansion variants.

3. Representing Repeat Expansions in FHIR

3.1 FHIR Resource

As a genomic variant, repeat expansion variants should be represented using an Observation resource of Variant profile (`http://hl7.org/fhir/uv/genomics-reporting/StructureDefinition/variant`). This will be in consistent with other existing variant type in HL7, such as copy number variants.

3.2 FHIR Element

According to GA4GH (**1.3.2**) and lab reports (**1.3.3**), we identify two terms required for representing a repeat expansion: repeat motif (also known as repeat nucleotides) and repeat number (a.k.a. repeat count).

Another challenge for representing repeat expansions is how to represent the ordered structure of repeat expansion pairs, as the mixed model has multiple lines of pairs in sorted order. The order of nucleotide-number pair plays an essential role in representing its biological/genomic meaning (e.g., repeat expansion ATG[30]CTG[20] is completely different from CTG[20]ATG[30] in the biological sense) and should be represented as well. Therefore, we will add one more element to address the ordering of repeat pairs.

3.3 FHIR Representation Strategy

Please note that prior versions of our proposed FHIR representation strategy can be found in the JIRA ticket (`https://jira.hl7.org/browse/FHIR-34418`) and will not be repeated here.

We propose three new components for representing repeat expansions.

Component	Description	Value type	Example
repeat-motif	Nucleotides of a repeat expansion motif	string	CTG
repeat-number	Number of repeats of a repeat expansion	Quantity	30
repeat-motif-order	Sequence position of a given motif-number pair	Quantity	1

For instance, the second motif of CTG[30]CAG[50] has component values of CAG, 50, and 2, respectively.

Here is an example of how these components are populated if only one motif is present. We can optionally omit **repeat-motif-order** like so:

```
.component[1].code = repeat-motif
.component[1].value = CTG
.component[2].code = repeat-number
.component[2].value = 30
```

If there are multiple repeat expansion pairs, all three components are used. Furthermore, we will use an extension, **component-set**, to group the related components together. Related components with the same **component-set** extension describe the same motif. The **component-set** value is determined by the sender/creator of the resource. CTG[30]CAG[50] is represented as:

```
.component[1].code = repeat-motif
.component[1].value = CTG
.component[1].extension[component-set] = "repeat-expansion-set1"
.component[2].code = repeat-number
.component[2].value = 30
.component[2].extension[component-set] = "repeat-expansion-set1"
.component[3].code = repeat-motif-order
.component[3].value = 1
.component[3].extension[component-set] = "repeat-expansion-set1"

.component[4].code = repeat-motif
.component[4].value = CAG
.component[4].extension[component-set] = "repeat-expansion-set2"
.component[5].code = repeat-number
.component[5].value = 50
.component[5].extension[component-set] = "repeat-expansion-set2"
.component[6].code = repeat-motif-order
.component[6].value = 2
.component[6].extension[component-set] = "repeat-expansion-set2"
```

This representation strategy avoids using confusing delimiters/symbols as well as nested extensions. It is not entirely favorable to use extensions to group components together, but resulting multiple repeat expansion motifs (i.e., mixed model) is currently not a common scenario, so we do not expect this extension to be used to represent most repeat expansions. As of writing and to our knowledge, there is one test in the United States (`https://www.preventiongenetics.com/testInfo?val=Oculopharyngeal-Muscular-Dystrophy-via-the-PABPN1-%28GCN%29-Repeat-Expansion`) that can result mixed model repeat expansions.

3.4 Constraints/Guidance

Constraints or guidance should be added to ensure the necessary components are included. Briefly:

- If there is one repeat expansion motif, both **repeat-motif** and **repeat-number** must be included.
- If there are multiple motifs, **repeat-motif**, **repeat-number**, and **repeat-motif-order** must be included, and each component must have an extension **component-set**.

4 Conclusion

We discussed a discrete representation for repeat expansions that concisely and unambiguously describes the repeat motif, number, and motif order, and is compatible with the format used in lab reports today.