HL7 FHIR Proposal for Repeat Expansion Variant Model

Contents

1. Background

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

The repeat nucleotides (motif) in 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 repeats length correlates with disease onset, which means longer repeats may drive pathology. The unusual CAG repeat expansion encoding 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***), the 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).

The current identified repeat motif usually ranges from 3 to 12 bp and trinucleotide is the most common 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 representation 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). Following is an expression of sequence comprised of tandem repeating subsequence in. 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. The 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 using [ ], using • or using /), 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 System

In Epic System, repeat expansion variant is defined as repeat-nucleotide/repeat-number pair and stored in one related group. We cand display repeat expansion in print group as following (Figure 7).

Type: Repeat Expansion Repeat Expansion: GCG[7]GCA[3] Classification: Normal

Figure 7 Print group display of repeat expansion in Epic. The HGVS recommendation format (e.g GCG[7]) is current adopted in Epic System for displaying in print group.

2. Challenges

However, repeat expansion variant is out of the scope of HL7 Version 2 Implementation Guide. In Lab Results Interface (LRI), it clearly 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.

So we need to design a model to represent repeat expansion variant.

3. Repeat Expansion of Representation in FHIR

3.1 FHIR Resource

As one of distinct variant type, Repeat Expansion Variant should be represented as Variant, which is derived from Finding. This will be in consistent with other existing variant type in HL7, such as copy number variant. It should also have path starting at Observation.

```
http://hl7.org/fhir/uv/genomics-reporting/StructureDefinition/variant
```

3.2 FHIR Element

According to the GA4GH (**1.3.2**) and lab report (**1.3.3**), we identify two main terms that are required for representing repeat expansion, repeat nucleotides (or repeat motif) and repeat number (or repeat count). Another challenge for representing Repeat Expansion is how to represent the ordered structure of repeat-expansion pair (mixed model has multiple lines of pairs in sorted order). The order of nucleotide-number pair plays an essential role in representing its biological/genomics meaning (e.g. repeat expansion ATG[30]CTG[20] is completely different from CTG[20]ATG[30] in biological sense) and should be guaranteed to be represented as well. So we propose to add one more element (sequence-order) to address this order issue of repeat pair.

So there are three main elements used for represent repeat expansion with the consideration of mixed model:

(1) **repeat-nucleotides**, defined as the repeat motif (e.g., CTG)
(2) **repeat-number,** defined as the number of motif repeated (e.g. 35)

(3) **sequence-order,** defined as the line number of given nucleotide-number-pair (e.g. line 1 for CTG[30]CAG[50] is the line for CTG[30] pair).

We intend to use Observation.extension to represent the repeat-nucleotides, repeat-number and sequence-order for Repeat Expansion Variant.

3.3 FHIR Representation Strategy

We proposed **Represent Repeat Expansion in Nested Extension Structure** to represent repeat expansion extension in FHIR.

3.3.1 Interface Design and Interface Message for Repeat Expansion Filing

In **1.3-Repeat Expansion Name Convention**, we discussed the different name convention of repeat expansion from different scenario and found there is no standard to represent it across all scenarios. Thus the first challenge we have is how to input and save repeat expansion variant information in an efficient way. Compared to using concatenated string, such as CTG[10]CAG[20] (HGVS format), $(CTG)_{10}$ • $(CAG)_{20}$, $[CTG]_{10}$ • $[CAG]_{20}$ or $(CTG)_{10}$ / $(CAG)_{20}$ (academic convention), or TYPE:CTG**,** Length:540 (sample interface message for lab reports) .

```
OBX|13|ST|069551-0^Genomic Alt Allele^BRLI^^^LOINC||TYPE:CTG,LENGTH:540||||||F||||MD2
```

There are different choice of delimiter (• / **,**) and symbols ([ ] ( )), which makes string concatenating error prone and hard to build a well-accepted standard for all scenario.

Therefore, to address this issue, we choose to use separate pieces of repeat-nucleotide and repeat-number, to avoid any misuse and confusing parts in repeat expansion variant (Figure 8).

```
OBX||ST|564^Repeat Nucleotides^BNH_LRR|2a.a|AAA||||||||||||||||||||||RSLT|VAR
OBX||ST|564^Repeat Nucleotides^BNH_LRR|2a.c|CCC||||||||||||||||||||||RSLT|VAR
OBX||ST|564^Repeat Nucleotides^BNH_LRR|2a.b|TTT||||||||||||||||||||||RSLT|VAR
OBX||NM|565^Repeat Number^BNH_LRR|2a.a|100||||||||||||||||||||||RSLT|VAR
OBX||NM|565^Repeat Number^BNH_LRR|2a.b|200||||||||||||||||||||||RSLT|VAR
OBX||NM|565^Repeat Number^BNH_LRR|2a.c|300||||||||||||||||||||||RSLT|VAR
```

Figure 8 Sample interface message OBX in Epic. This message is to file data representing repeat expansion AAA[100]TTT[200]CCC[300], which are ordered by reptation ID (2a.a, 2a.b and 2a.c, respectively).

3.3.2 Rational for Representing Repeat Expansion using nested Extension in FHIR

The reason for choosing nested extension, instead of concatenated string, is similar to the reason mentioned in **3.3.1**. Since the large discrepancy of symbols and formats to represent repeat expansion, one certain format, such as $(CTG)_{10}$ may not be recognized and well-accepted across different labs/facilities since they argue CTG[10] should be their convention to represent it.

Splitting them as two separate parts (repeat-nucleotide and repeat-number) will largely reduce the confusing and successful address above issue. For instance, repeat-nucleotide: CTG and repeat-number: 10 would be more acceptable than above two concatenated strings.

3.3.3 Proposal of Representing Repeat Expansion using nested Extension in FHIR (Option 1)

We are going to use following nested extension structure to build each repeat-pair in each line for one repeat expansion variant (**Figure 9**).

(1) **repeat-nucleotides**, defined as the repeat motif (e.g., CTG)
(2) **repeat-number,** defined as the number of times the motif is repeated (e.g. 35)
(3) **sequence-order,** defined as the line number of given nucleotide-number-pair (e.g. line 1 for CTG[30]CAG[50] is the line for CTG[30] pair).

Figure 9 Sample representation of repeat expansion in FHIR. This extension represents repeat expansion GCG[7]GCA[1] (To be note, this extension is different from GCA[1]GCG[7]).

There are several advantages of applying nested Extension for repeat expansion variant.
(1) It is well-structured, and match data stored in database
(2) Avoid the misuse and confusing of delimiter/symbol usage to represent repeat expansion variant in different scenarios.
(3) Represent repeat nucleotide and repeat number separately, easy for validation

3.3.4 Alternative Strategy (Option 2): Represent Repeat Expansion in Concatenated String in FHIR

We may still consider this approach since it is more human-readable and easy to transmit. And it is easy to fit into component if we choose repeat expansion as component instead of extension in future.

4. Conclusion

Currently, we would still choose option 1 (3.3.3), which is using nested extension structure to represent repeat expansion variant, in order to minimize the conflict and confusing parts by using concatenated string.