HL7 FHIR Proposal for Repeat Expansion Variant Model

Version 3

Contents

What's new in this version?

repeat-motif-order is now an extension and replaces component-set

Additions to the Constraints/Guidance section: gene-studied and cytogenomic-nomenclature

1. Background

Repeat expansion, also known as repeat tandem, is characterized by polymorphic nucleotide sequences scattered throughout the human genome (***Paulson, 2018***). 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) (***Depienne 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**).

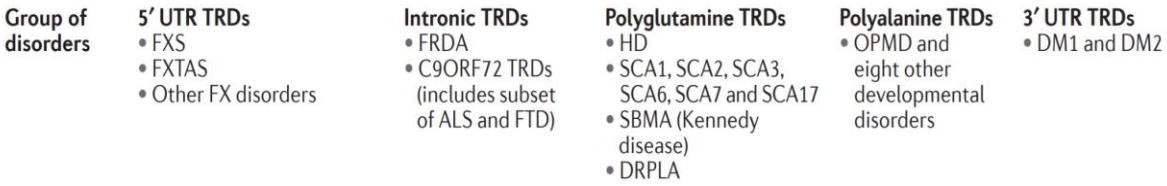


Figure 1 Location of repeat expansion within genes for repeat expansion diseases (**Hannan et al., 2018**).

Currently identified repeat motifs usually range from 3 to 12 base pairs. Trinucleotides (3 base pairs) are the most commonly found repeats (***Paulson, 2018***) (**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, 2018).

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 (including large sequence repeats) is defined in RepeatedSequenceExpression (https://vrs.ga4gh.org/en/stable/terms_and_model.html#repeatedsequenceexpression). The following is an expression of a sequence comprised of a tandem repeating subsequence. Besides type, repeat sequence has two more fields, seq_expr (using sequence ID) and count (using type IndefiniteRange) (**Figures 3, 4**).

Field	Type	Limits	Description
type	string	1..1	MUST be "RepeatedSequenceExpression"
seq_expr	LiteralSequenceExpression | DerivedSequenceExpression	1..1	An expression of the repeating subsequence
count	Number | IndefiniteRange | DefiniteRange	1..1	The count of repeated units, as an integer or inclusive range

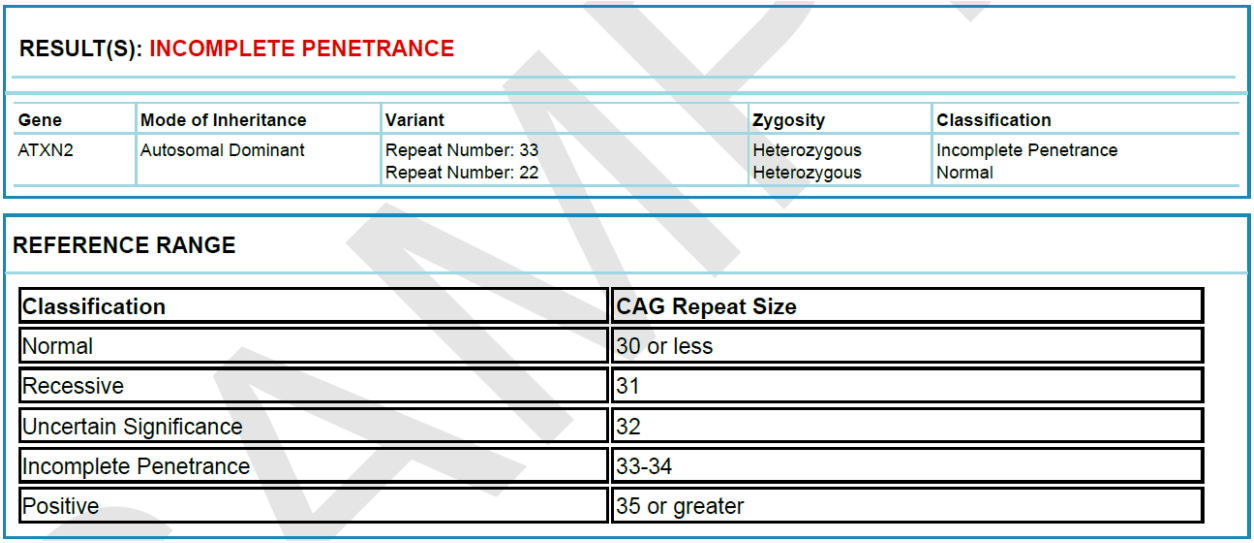
Figure 3 Representation of repeat expansion definition in GA4GH.


```
"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.



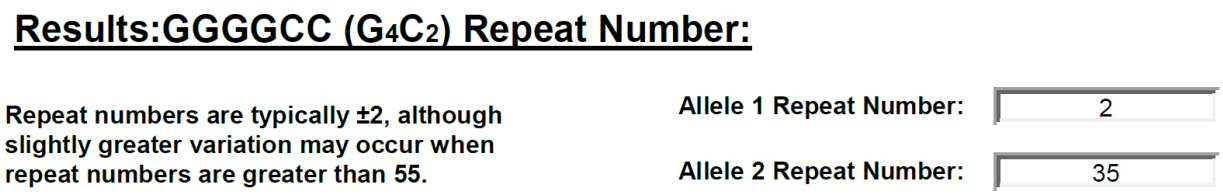
RESULT(S): INCOMPLETE PENETRANCE

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., 2017**). There is the most used format to represent repeat expansion in publication.

However, there are some publications which use different formats, e.g $[CTG]_{\geq n}$ to represent repeat expansions (**Alfadhli et al., 2004**). 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).

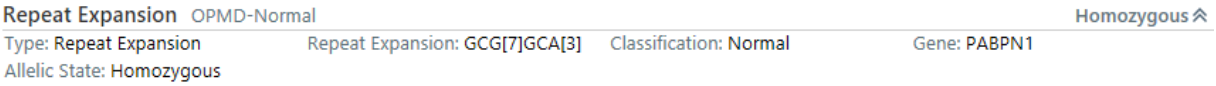


Figure 7 Display of a repeat expansion in Epic. The HGVS recommendation format (e.g., GCG[7]) is currently adopted in Epic for display.

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 R4 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 two new components for representing repeat expansions.

Component	Description	Value type	Example
repeat-motif	Nucleotides of a repeat expansion motif	string	CAG
repeat-number	Number of repeats of a repeat expansion	Quantity	40

We will illustrate these components with examples. Examples with the same number are part of the same representation for a repeat expansion.

Here is an example of how these components are populated if only one motif is present.

```
.component[1].code = repeat-motif
.component[1].value = CAG
.component[2].code = repeat-number
.component[2].value = 40
```

Example 1 Repeat expansion with one motif. This repeat expansion is for Huntington disease (see Table 1).

If there are multiple repeat expansion pairs, an extension, **repeat-motif-order**, must be used both to indicate the sequence position of a given motif-number pair and to group the related components together. Related components with the same **repeat-motif-order** extension describe the same motif. The value must be a natural number starting from 1 and increasing sequentially. GCG[6]GCA[4]GCG[1] is represented as:

```
.component[1].code = repeat-motif
.component[1].value = GCG
.component[1].extension[repeat-motif-order] = 1
.component[2].code = repeat-number
.component[2].value = 6
.component[2].extension[repeat-motif-order] = 1

.component[3].code = repeat-motif
.component[3].value = GCA
.component[3].extension[repeat-motif-order] = 2
.component[4].code = repeat-number
.component[4].value = 4
.component[4].extension[repeat-motif-order] = 2

.component[5].code = repeat-motif
.component[5].value = GCG
.component[5].extension[repeat-motif-order] = 3
.component[6].code = repeat-number
.component[6].value = 1
.component[6].extension[repeat-motif-order] = 3
```

Example 2a Mixed model repeat expansion. This shows OPMD allele-2 in PABPN1 (see Table 2).

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 at least one repeat expansion motif to be represented, both **repeat-motif** and **repeat-number** must be included.
- If there are multiple motifs, extension **repeat-motif-order** must also be included on each of the **repeat-motif** and **repeat-number** components.
    - **repeat-motif-order** must be a natural number starting from (i.e., no less than) 1, be the same for the **repeat-number** corresponding to its **repeat-motif**, and be unique for each **repeat-motif** and **repeat-number** pair. The maximum value of **repeat-motif-order** on a Variant observation's components must be equal to the number of **repeat-motif** and **repeat-number** pairs.

Additionally, we propose that **gene-studied** be required as well, since it is an important piece of information to tell apart different repeat expansions. Continuing from Example 2a:

```
.component[7].code = 48018-6

.component[7].value = PABPN1
```

Example 2b Continuation of mixed model repeat expansion from Example 2a, now with **gene-studied**.

Furthermore, we recommend that labs also send **cytogenomic-nomenclature**. Although this information is generally not sent as of writing, it will help inform the parts of a gene a repeat expansion covers. Continuing from Example 2b:

```
.component[8].code = 81291-7

.component[8].value = 14p11.2-q13
```

Example 2c Continuation of mixed model repeat expansion from Example 2b, now with **cytogenomic-nomenclature**.

While recommending that labs send **exact-start-end** may seem like it provides more precise information about the location of the repeat expansion, it is not easy to determine with current assembly algorithms (e.g., BLAST and k-mer). Repeat expansions often use cosmid contigs and clone short genomic fragments for assembly, and cannot be mapped precisely to the human genome.

We also oppose requiring or recommending that labs send **ref-allele** because:

- The number of repeats can vary in patients with normal alleles; as an example, see Table 1 for Huntington's disease.
- The reference allele can be incorrect. Per **Song et al., 2018**, large repeats can be erroneously represented in reference genomes due to limitations in current sequencing techniques.

- Labs do not send this information, and to our understanding, clinicians and other consumers do not find it useful due to the above reasons.

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.

5 References

Paulson H. Repeat expansion diseases. Handbook of clinical neurology. 2018 Jan 1;147:105-23.

Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges?. The American Journal of Human Genetics. 2021 May 6;108(5):764-85.

Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nature Reviews Genetics. 2018 May;19(5):286-98.

Keum JW, Shin A, Gillis T, Mysore JS, Elneel KA, Lucente D, Hadzi T, Holmans P, Jones L, Orth M, Kwak S. The HTT CAG-expansion mutation determines age at death but not disease duration in Huntington disease. The American Journal of Human Genetics. 2016 Feb 4;98(2):287-98.

de Leeuw RH, Garnier D, Kroon RM, Horlings CG, de Meijer E, Buermans H, van Engelen BG, de Knijff P, Raz V. Diagnostics of short tandem repeat expansion variants using massively parallel sequencing and componential tools. European Journal of Human Genetics. 2019 Mar;27(3):400-7.

Brais B, Bouchard JP, Xie YG, Rochefort DL, Chrétien N, Tomé FM, Lafrentére RG, Rommens JM, Uyama E, Nohira O, Blumen S. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nature genetics. 1998 Feb 1;18(2):164-7.

Grewal RP, Karkera JD, Grewal RK, Detera-Wadleigh SD. Mutation analysis of oculopharyngeal muscular dystrophy in Hispanic American families. Archives of neurology. 1999 Nov 1;56(11):1378-81.

Liu Y, Wilson SH. DNA base excision repair: a mechanism of trinucleotide repeat expansion. Trends in biochemical sciences. 2012 Apr 1;37(4):162-72.

Kim JC, Harris ST, Dinter T, Shah KA, Mirkin SM. The role of break-induced replication in large-scale expansions of (CAG) n/(CTG) n repeats. Nature structural & molecular biology. 2017 Jan;24(1):55-60.

Alfadhli S, Elshafey AE, Bastaki L, Al-Awadi S. CTG repeat number at the myotonic dystrophy locus in healthy Kuwaiti individuals: possible explanation of why myotonic dystrophy is rare in Kuwait. Archives of neurology. 2004 Jun 1;61(6):895-8.

Song JH, Lowe CB, Kingsley DM. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. The American Journal of Human Genetics. 2018 Sep 6;103(3):421-30.