Genome Variation Ontology for annotation of complex structural variations*

Shuichi Kawashima[1,*,†], Takatomo Fujisawa[2,†] and Toshiaki Katayama[1,†]

[1]*Database Center for Life Science (DBCLS), 178-4-4 Wakashiba, Kashiwa-shi, Chiba 277-0871, Japan*
[2]*DDBJ Center, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan*

Abstract

The Genome Variation Ontology (GVO) is an ontology for the systematic description of various genomic variations, including complex structural variations in genomes. With advances in the discovery and genotyping methods for genomic variations, it is expected that new types of complex structural variations (SVs) will be discovered. We have developed the Ge-nome Variation Ontology to annotate all types of genomic variations, which includes complex SVs. Terms on genomic variations from dbSNP, dbVar, gnomAD, SO, VariO, and HGVS were collected and clustered manu-ally to generate 47 concepts. GVO is available at http://genome-variation.org/resource/gvo

Keywords

ontology, genome variation, structural variation, RDF

1. Introduction

In recent years, large-scale sequencing projects have been carried out, such as the 100,000 Genomes Project by Genomics England. And based on massive amounts of individual genome sequences, a number of genomic structural variations (SVs) have been reported. For example, gnomAD-SV [1] contains a large number of novel SVs discovered using an improved multi-algorithm ensemble method against high-coverage WGS. In addition it is envisaged that the widespread use of long-read sequence technologies will generate an ever-increasing amount of information on SVs in the near future.

As ontologies that are able to be used for annotation of genomic variations, the Sequence Ontology (SO) [2] and the Variation Ontology (VariO) [3] are available: SO is an ontology for annotation of sequence features, while VariO is an ontology for describing the effects, consequences and mechanisms of mutations in DNA, RNA and proteins. These include terms for canonical SVs, such as deletion and translocation. However, besides canonical SVs, gnomAD-SV, for example, also contains complex SVs classified into 11 subtypes, some of which are not available in the existing ontologies.

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

[*] Corresponding author.

[†] These authors contributed equally.

kwsm@dbcls.rois.ac.jp (S. Kawashima); tf@nig.ac.jp (T. Fujisawa); ktym@dbcls.jp (T. Katayama)
0000-0001-7883-3756 (S. Kawashima); 0000-0001-8978-3344 (T. Fujisawa); 0000-0003-2391-0384 (T. Katayama)

```
:Ins
    a owl:Class ;
    rdfs:label ”Ins”@en ;
    rdfs:seeAlso <http://purl.obolibrary.org/obo/SO_0000667>,
        <http://purl.obolibrary.org/obo/VariO_0142>,
        <http://varnomen.hgvs.org/recommendations/DNA/variant/
            insertion/> ;
    rdfs:subClassOf :Variation ;
    skos:definition ”Insertion”@en .
```

Figure 1: The gvo:Ins class for genomic insertion.

We have developed TogoVar database that integrates allele frequencies from Japanese populations and providing annotations for variant interpretation [4]. One of the notable feature of TogoVar is that all data is described in RDF. While TogoVar has targeted single nucleotide variants (SNVs) and some canonical SVs, we have a plan expand the target to complex SVs. In order to store complex SVs in TogoVar, an ontology that can be used to annotate them is needed. To prepare for the situation, we have developed Genome Variation Ontology (GVO), which is an ontology for de-scribing all types of genomic variation.

2. GVO: Genome Variation Ontology

We collected terms on genomic variation from gnomAD, dbSNP, dbVar, SO, VariO and HGVS and performed manual clustering on equivalent concepts among them. As a result, we obtained the 47 ontology classes corresponding to genomic variation types. Then we have organized these classes hierarchically under gvo:Variation to construct GVO. These concepts include several complex SVs, such as paired-duplication inversion and paired-deletion inversion, which are not available in the existing ontologies. GVO is available in the web site (http://genome-variation.org/resource/gvo) and BioPortal (https://bioportal.bioontology.org/ontologies/GVO). Figure 1 shows an example of the GVO classes.

We plan to use GVO with the FALDO ontology to convert genomic variations distributed in VCF format into RDF. Therefore we also introduce some properties needed for this in GVO.

References

[1] R. L. Collins, et al., A structural variation reference for medical and population genetics, Nature 581 (2020) 444–451. doi:10.1038/s41586-020-2287-8.

[2] K. Eilbeck, et al., The sequence ontology: a tool for the unification of genome annotations, Genome Biology 6 (2005) R44. doi:10.1186/gb-2005-6-5-r44.

[3] M. Vihinen, Variation ontology for annotation of variation effects and mechanisms, Genome Research 24 (2013) 356–364. doi:10.1101/gr.157495.113.

[4] N. Mitsuhashi, et al., Togovar: A comprehensive japanese genetic variation database, Human Genome Variation 9 (2022) 44. doi:10.1038/s41439-022-00222-9.