New concepts proposal

Concepts for variant description

Author	Jan Armida	Date last updated	23/02/2022
Project	Genomic Concepts	**Contact person**	Jan Armida
Status	Accepted	**Consulted expert**	WG

1 Rationale

Genetic variants (or genetic variations) define permanent changes in a DNA sequence. Variants in the human genome can be phenotypically neutral (with no detectable health issues), classified as benign in the diagnostic context, or to increase an individual's susceptibility or predisposition to a certain disease (classified as pathogenic or likely pathogenic). At the population scale, genetic variants differ in their frequency inside of a particular population and between different populations, with rare variants generally having stronger effects than common. The understanding of genetic variation and its impact is at the core of the idea of personalized health and medicine. Therefore, it is equally important for population genomics, constitutional diagnostic and predictive genetics and genomics, as well as oncogenomics and pharmacogenomics. Hence, the implementation of concepts describing a genetic variant, its annotation and interpretation (see concept Variant Interpretation) is of utmost importance and is one of the cornerstones of the SPHN Genomic Dataset.

2 Comparison to other standards/data models

2.1 HL7 FHIR

The genetic variation is represented in FHIR within an extension of the resource Observation. The element Genomic Variant is defined by a variant ID and the variant name according to the HGNC nomenclature. Importantly the type of variant is also coded and makes full use of both HGNC and LOINC to fully describe its typology.

2.2 LOINC

LOINC provides a limited selection of codes that can be used to define a variant and its typology. Changes in DNA structure (e.g., single nucleotide variants, structural variants) are defined by the code 48019 - DNA change type, that allows the following answers:

Answer	**Answer ID**
Wild type	LA9658-1
Deletion	LA6692-3
Duplication	LA6686-5
Insertion	LA6687-3
Insertion/Deletion	LA6688-1
Inversion	LA6689-9
Substitution	LA6690-7

2.3 GA4GH Phenopackets

In the GA4GH Phenopackets, the concept of *Genomic Variant* is represented by the *Variation descriptor* building block which is, in turn, an extension of the GA4GH Variation Representation Specification (VRS) framework. *Variation descriptor* functions as a wrapper class which adds human-readable labels and description to the otherwise machine-oriented representation of VRS. A variant descriptor object is represented as follows:

Variation Descriptor

Field	Type	Multiplicity	Description
id	string	1..1	Descriptor ID; MUST be unique within document. REQUIRED.
variation	Variation	0..1	The VRS `Variation` object
label	string	0..1	A primary label for the variation
description	string	0..1	A free-text description of the variation
gene_context	GeneDescriptor	0..1	A specific gene context that applies to this variant
expressions	Expression	0..*	HGVS, SPDI, and gnomAD-style strings should be represented as Expressions
vcf_record	VcfRecord	0..1	A VCF Record of the variant. This SHOULD be a single allele, the VCF genotype (GT) field should be represented in the allelic_state
xrefs	string	0..*	List of CURIEs representing associated concepts. Allele registry, ClinVar, or other related IDs should be included as xrefs
alternate_labels	string	0..*	Common aliases for a variant, e.g. EGFR vIII, are alternate labels
extensions	Extension	0..*	List of resource-specific Extensions needed to describe the variation

Field	Type	Multiplicity	Description
molecule_context	MoleculeContext	1..1	The molecular context of the vrs variation.
structural_type	OntologyClass	0..1	The structural variant type associated with this variant, such as a substitution, deletion, or fusion. We RECOMMEND using a descendent term of SO:0001537.
vrs_ref_allele_seq	string	0..1	A Sequence corresponding to a “ref allele”, describing the sequence expected at a SequenceLocation reference.
allelic_state	OntologyClass	0..1	See allelic_state below. RECOMMENDED.

At the core of Variant Descriptor lies the VRS object *Variation.* This object is the conceptual root of all types of biomolecular variations and delivers the minimal information required to describe a genetic variation and is mainly composed by a description of allele and its position relative to a human reference sequence. In order to enable human-readable representations, the class “Expression” can be instantiated multiple times using different expression nomenclatures (e.g., HGVS, SPDI).

Variation

Field	Type	Limits	Description
_id	CURIE	0..1	Variation Id; MUST be unique within document
type	string	1..1	MUST be “Allele”
location	Location | CURIE	1..1	Where Allele is located
state	Sequence Expression | SequenceState (deprecated)	1..1	An expression of the sequence state

In its simplest implementation, VRS and Phenopackets allows, for example, to represents a single nucleotide variation in a simple, machine-readable way:

Representation of the ClinVar allele 13294

```
variationDescriptor:
  id: "clinvar:13294"
  variation:
    allele:
      sequenceLocation:
        sequenceId: "NC_000010.11"
        sequenceInterval:
          startNumber:
            value: "121496700"
          endNumber:
            value: "121496701"
      literalSequenceExpression:
        sequence: "G"
  moleculeContext: "genomic"
  vrsRefAlleleSeq: "T"
```

```
allelicState:
  id: "GENO:0000135"
  label: "heterozygous"
```

The VRS framework and Phenopackets are used as a reference to define the concepts proposed in this document.

3 Concept information

Contextualized concept name	Contextualized description	Type	Standard	Value set	Meaning binding
Variant Descriptor	human-readable description of the variant				
variant identifier	identifier assigned to a variant from an external repository	Code	ClinVar, RefSNP or other		
genetic variation	variant described by its position and sequence change	Genetic Variation			
notation	description of the variant using a specific nomenclature or syntax	Variant Notation			
variant type	description of the variation by typology of sequence alteration, e.g. substitution, insertion, deletion, etc.	Code	SO	descendant of: SO:0001059 sequence_alteration	
gene context	specific gene context where the variation occurs	Gene			
allele origin	genetic origin of the variant allele, whether it was inherited from a parent or occurred as a spontaneous mutation event in a germline or somatic cell, e.g. de novo variation, germline variation, somatic variation etc.	Code	GENO	descendant of: GENO:0000877 allele origin	GENO: 0000877 allele origin
zygosity	similarity or dissimilarity of allelic state of the variant	Code	GENO	descendant of : GENO:0000133 zygosity	GENO: 0000133 zygosity

Contextualized concept name	Contextualized description	Type	Standard	Value set	Meaning binding
Variant Notation	description of the variant using a specific nomenclature				
notation and version	name and version of the notation used to describe the variation	qualitative		link to value set and specification*	SNOMED CT: 705113004 |Terminology system (qualifier value)|
expression	expression describing the variation according to the chosen nomenclature	string			

*This valueset refers to the SPHN Dataset – Coding System and version tab and needs to be istantiated as for the Code concept (see Examples below).

Contextualized concept name	Contextualized description	Type	Standard	Value set	Meaning binding
Genetic Variation	a genetic variation occurring at a defined position				SO: 0001060 sequence variant; GENO: 0000476 variant
genomic position	location of the variant within a sequence	Genomic Position			
chromosomal location	location of the variant within a chromosome	Chromosomal Location			

Contextualized concept name	Contextualized description	Type	Standard	Value set	Meaning binding
Single Nucleotide Variation	single nucleotide change in a DNA sequence at a specific location	Genetic Variation			SO: 0001483 SNV
reference allele	a base (A;T;C;G) at a specific position that matches with the reference	string			GENO: 0000036 reference allele
alternate allele	any base (A;T;C;G), other than the reference allele at a given position	string			GENO: 0000002 variant allele
genomic position	sequence position where the single nucleotide variant was found	Genomic Position			
chromosomal location	chromosomal location where the single nucleotide variant was found	Chromosomal Location			

Contextual concept name	Contextual description	Type	Standard	Value set	Meaning binding
Genomic Position	genomic position with respect to a reference				GENO: 0000902 genomic feature location
start position	coordinate that indicates the beginning of the contiguous length of the reference sequence	quantitative			GENO: 0000894 start position
end position	coordinate that indicates the end of the contiguous length of the reference sequence	quantitative			GENO: 0000895 end position
coordinate convention	convention used for the interpretation of coordinates	qualitative		Residue coordinates; Inter-residue coordinates	
reference sequence	sequence used as a reference to extrapolate the variant position	Reference			GENO: 0000017 reference sequence

Concept name	Description	Type	Standard	Value set	Meaning binding
Reference	reference construct used for comparison purposes				GENO: 0000152 reference
reference identifier	ID of a reference sequence deposited in a repository	Code	NCBI GenBank or other		

Concept name	Description	Type	Standard	Value set	Meaning binding
Chromosomal Location	chromosome locus defined as cytoband intervals				SO: 0000830 chromosome part; GENO: 0000614 chromosomal region
chromosome	chromosome associated to the cytoband interval	Chromosome			
start cytoband	the start cytoband region defined as region nearer to the p-arm telomere compared to the end cytoband	Code	ISCN		
end cytoband	the end cytoband region defined as region nearer to the q-arm telomere	Code	ISCN		

	compared to the start cytoband

Notes:

- As for VRS, all locations are presumed with respect to the positive/forward/Watson strand.

4 Impact on the SPHN Dataset

No impact on the current SPHN Dataset release.

5 Discussion

The basic structure of GA4GH Phenopackets' *Variation descriptor* and VRS *Variation* was used as the basis for the development of the concepts describing a genetic variation. Within this paradigm, a variant is identified in its minimal and mostly interoperable component by *Genetic variation*, which defines the genomic position and the identified allele (i.e., specific sequence state) that has changed compared to a reference sequence. This information is captured within the concept *Variant Descriptor*, which functions as a wrapper gathering all useful information about the variation, including its representation using different expressions. The *Variant Descriptor* concept can be used to link the variant to a certain disease, impact on protein function and other sources of information.

When needed, the position of the variant within a sequence can be described using different coordinate conventions like residue coordinates or inter-residue coordinates (also known as 0-based coordinate system). It is important to clearly indicate the type of coordinate system used to improve interoperability.

As mentioned above, different conventions exist to describe a variant and different projects may need to communicate their results in a specific format. For that reason, the abstract concept *Variant Notation* allows to create instances that describe the variant following a specific syntax.

6 Examples

In the following example, a simple implementation of the concepts is used to describe a known single nucleotide variation occurring in the epidermal growth factor receptor (EGFR) coding region:

Variant Descriptor:

variant type: substitution (SO:1000002)
notation:
  expression: NM_005228.4(EGFR):c-237A>G
  coding system and version: HGVS-20.05
gene context:
  gene identifier:
    identifier: HGNC:3236
    name: EGFR
    coding system and version: HGNC-2022
allele origin: germline allele origin (GENO:0000888)

This representation is sufficient to describe the variant using a widespread nomenclature as HGVN (note that many other fields can be left empty). In this case, multiple instances of the concept *Variant Notation* would allow to represent the variant at different levels (e.g., variant at the gene, transcript or protein level).

When needed, more information about the variant can be delivered by instantiating the appropriate concepts. Notably, the use of *Genomic Variation* would add an extra layer of information that would greatly benefit potential downstream analysis of the data.

Single nucleotide variation:

position sequence:
  start sequence: 121496700
  end sequence: 121496701
  coordinate convention: Inter-residue coordinate
reference sequence:
    identifier: NC_000010.11
    name: Homo sapiens chromosome 10, GRCh38.p14 Primary assembly
    coding system and version: NCBI-2022

Information about the macro-location of the variant can be given by instantiating the *Chromosomal Location* concept (values unrelated to previous example):

chromosomal location:

chromosome:
  chromosome name:
    identifier: 16312006
    name: Chromosome pair 11 (cell structure)
    coding system and version: SNOMED-CT-2021-01-31
start cytoband:
  identifier: q13.32
  coding system and version: ISCN-2022

end cytoband:

identifier: q13.32

coding system and version: ISCN-2022