

New concept proposal

Structural Variations

Author	Jan Armida	Date last updated	13/10/2023
Project	General Interest	Contact person	DCC
Status	2024.1	Consulted expert	Thomas Müller

1 Rationale

A structural variation is a genetic alteration involving changes in DNA's structure, encompassing a broad spectrum of genomic variations, such as insertions, deletions, duplications, inversions and more complex alterations, which can vary in size and clinical significance. These concepts are designed to address a number of common structural variations and need to be paired with the SPHN Concept of Variant Descriptor for high-level descriptions. More complex structural variations can be described using HGVS annotations enabled by Variant Descriptor.

Concepts included:

- Genomic Insertion
- Genomic Deletion
- Copy Number Variation







2 Comparison to other standards/data models

In GA4GH Variation Representation Specification (VRS) represents small insertions and deletions as part of the Molecular Variation class "Allele". An allele is identified by its location and sequence expression as illustrated below:

Field	Туре	Limits	Description
_id	CURIE	01	Variation Id. MUST be unique within document.
type	string	11	MUST be "Allele"
location	CURIE Location	11	Where Allele is located
state	Sequence Expression	11	An expression of the sequence state

More complex variations such as copy number variants are subclasses of the Systemic Variation class, which describes a variation of multiple molecules in the context of a system (e.g. a genome, sample, or homologous chromosomes).

Copy Number Count:

Field	Туре	Limits	Description
_id	CURIE	01	Variation Id. MUST be unique within document.
type	string	11	MUST be "CopyNumberCount"
subject	Location CURIE Feature	11	A location for which the number of systemic copies is described.
copies	Number Indefini- teRange DefiniteRange	11	The integral number of copies of the subject in a system



2 Concept information

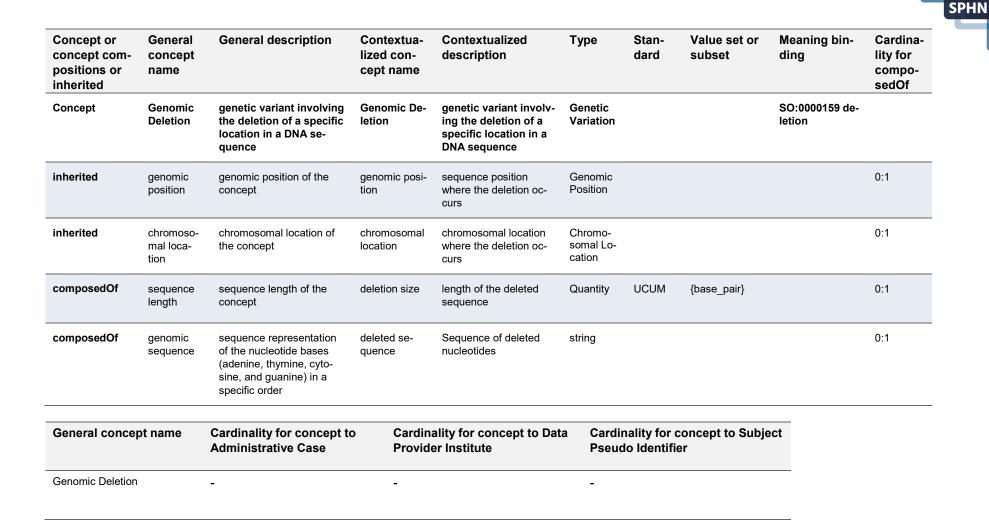
Concept or concept compositions or inherited	General concept name	General description	Contextua- lized con- cept name	Contextualized description	Туре	Stan- dard	Value set or subset	Meaning bin- ding	Cardina- lity for compo- sedOf
Concept	Genomic Insertion	genetic variant involving the addition of a DNA sequence at a specific location	Genomic Insertion	genetic variant involv- ing the addition of a DNA sequence at a specific location	Genetic Variation			SO:0000667 insertion	
inherited	genomic position	genomic position of the concept	genomic posi- tion	sequence position where the insertion oc- curs	Genomic Position				0:1
inherited	chromoso- mal loca- tion	chromosomal location of the concept	chromosomal location	chromosomal location where the insertion oc- curs	Chromo- somal Lo- cation				0:1
composedOf	sequence length	sequence length of the concept	insertion size	length of the inserted sequence	Quantity	UCUM	Unit: {base_pair}		0:1
composedOf	genomic sequence	sequence representation of the nucleotide bases (adenine, thymine, cyto- sine, and guanine) in a specific order	inserted sequence	Sequence of nucleo- tides added to an exist- ing DNA sequence	string				0:1





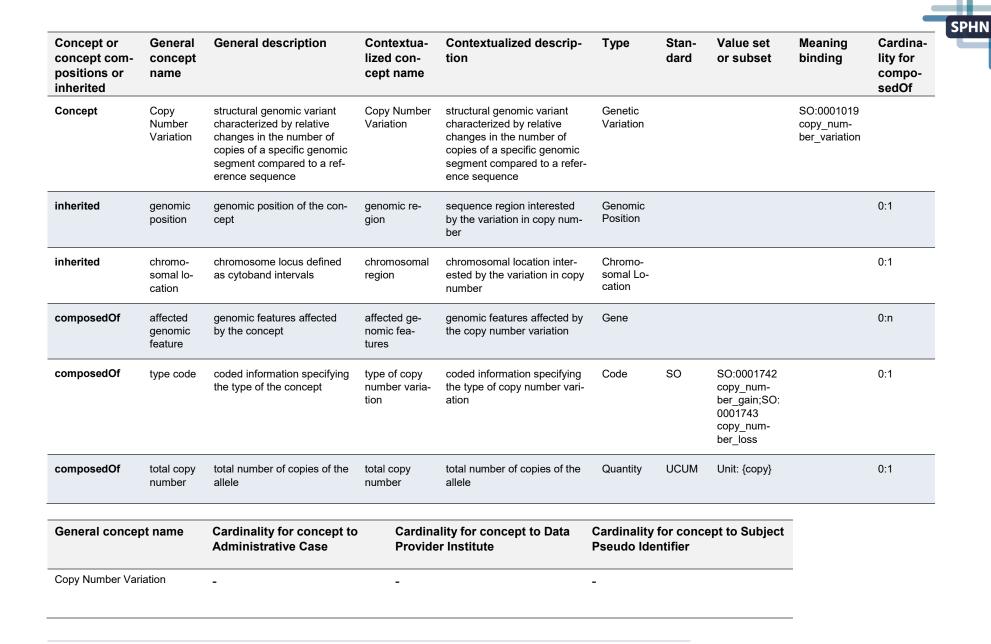
General concept name	Cardinality for concept to Administrative Case	Cardinality for concept to Data Provider Institute	Cardinality for concept to Subject Pseudo Identifier
Genomic Insertion	-	-	-





Personalized

Network



Personalized

Network



3 Impact on the SPHN Dataset

Optional (if existing concepts need to be adapted because of this new concept, state here the currently released version of the existing concept and the proposed adapted version)

4 Discussion

This series of concepts represents an expansion of the SPHN Genomic Variant framework, aimed at addressing the gaps in the current set of variant concepts. All the concepts introduced here inherit from the SPHN Genetic Variation superclass and are intended to be instantiated alongside the Variant Descriptor. This approach enables a high-level, human-readable representation of the variant through the Variant Descriptor, as well as a machine-readable and easily queryable representation of the variants through the various subclasses.

In the context of these concepts, the term 'structural variants' is used to encompass all genomic changes, such as insertions, deletions, and others, that involve alterations spanning more than one nucleotide. For single nucleotide variations like SNPs, we recommend using the Single Nucleotide Variation concept.







5 Examples

The following examples include an instantiation of Variant Descriptor as an integral part of the SPHN variant representation.

Example 1: Insertion of 5bp in a specific sequence location

```
Variant Descriptor
variant type: insertion (SO:1000034)
allele origin: somatic allele origin (GENO:0000882)
notation [Variant Notation]:
        expression [string]: NC_000023.10:g.32862923_32862924insATGCC
        coding system and version: HGVS-20.05
genetic variation [Genomic Insertion]:
        position sequence [Genomic Position]:
                start sequence [quantitative]: 32862923
                end sequence [quantitative]: 32862924
                coordinate convention [qualitative]: Residue coordinate
                reference sequence [Code]:
                        identifier: NC_000023.10
                        name: Homo sapiens chromosome X, GRCh37.p13 Primary Assembly
                        coding system and version: NCBI-2023
        chromosomal location [Chromosomal Location]: -
        sequence length [Quantity]:
                value: 5
                unit: {base pair}
        genomic sequence [string]: ATGCC
```



Example 2: Complex insertion not fully supported by Genomic Insertion can still be described by Variant Descriptor and Variant Notation.

Variant Descriptor

variant type: insertion (SO:1000034)

allele origin: somatic allele origin (GENO:0000882)

notation [Variant Notation]:

expression [string]: LRG_199t1:c.419_420ins[T;450_470;AGGG]

coding system and version: HGVS-20.05

In this case, the insertion of T is also followed by an extra copy of the sequence (450 to c.470) and another four nucleotides (AGGG). Such insertion cannot be fully explained by Genomic Insertion.



Example 3: Deletion of 3bp in a specific sequence location

```
Variant Descriptor
variant type: deletion (SO:0000159)
allele origin: somatic allele origin (GENO:0000882)
notation [Variant Notation]:
        expression [string]: NC_000023.11:g.33344590_33344592del
        coding system and version: HGVS-20.05
genetic variation [Genomic Deletion]:
        position sequence [Genomic Position]:
                start sequence [quantitative]: 33344590
                end sequence [quantitative]: 33344592
                coordinate convention [qualitative]: Residue coordinate
                reference sequence [Code]:
                        identifier: NC_000023.11
                        name: Homo sapiens chromosome X, GRCh38.p14 Primary Assembly
                        coding system and version: NCBI-2023
        chromosomal location [Chromosomal Location]: -
        sequence length [Quantity]:
                value: 3
                unit: {base pair}
        genomic sequence [string]: GAT
```



Example 4: Deletion of chromosome 7q34 (broad deletion)

```
Variant Descriptor
variant type: deletion (SO:0000159)
allele origin: somatic allele origin (GENO:0000882)
notation [Variant Notation]: -
genetic variation [Genomic Deletion]:
        position sequence [Genomic Position]: -
        chromosomal location [Chromosomal Location]:
                chromosome [Chromosome]:
                        chromosome name: Chromosome pair 7 (SNOMED CT:70488008)
                        start cytoband:
                                identifier: q34
                                coding system and version: ISCN-2022
                        end cytoband:
                                identifier: q34
                                coding system and version: ISCN-2022
        sequence length [Quantity]: -
        genomic sequence [string]:-
```



Example 5: Copy number gain issue from a microduplication of the 15q11.2 region of Chromosome 15 observed in a rare genetic condition involving the Burnside-Butler region. Here the genes involved in the duplication are NIPA1, NIPA2, CYFIP1 and TUBGCP5.

```
Variant Descriptor
variant type: copy number variation (SO:0001019)
allele origin: germline allele origin (GENO: 0000888)
genetic variation [Copy Number Variation]:
        chromosomal location [Chromosomal Location]:
               chromosome [Chromosome]:
                       chromosome name: Chromosome pair 15 (SNOMED CT: 71678009)
                       start cytoband:
                               identifier: q11.2
                               coding system and version: ISCN-2022
                       end cytoband:
                               identifier: q11.2
                               coding system and version: ISCN-2022
        affected feature [Gene]:
               code: NIPA1 (HGNC:17043)
        affected feature [Gene]:
               code: NIPA2 (HGNC:17043)
        affected feature [Gene]:
               code: CYFIP1 (HGNC:13759)
        affected feature [Gene]:
               code: TUBGCP5 (HGNC:18600)
        type code [Code]: copy number gain (SO:17043)
        total copy number [Quantity]:
               value: 2
```

unit: {copy}