New concept proposal

Gene

Author	**Jan Armida**	**Date last updated**	**01/10/2022**
Project	Genomic Concepts	**Contact person**	Jan Armida
Status	Accepted	**Consulted expert**	WG

1 Rationale

The Gene concept is at the core of most datasets representing analysed and annotated genomic information. This is particularly the case for projects (within SPHN such as SVIP, SPO and SwissGenVar) that need to share clinically relevant genetic variants. After annotation and interpretation, these variants are associated with a specific gene(s) or other elements in the genome (e.g. regulatory regions). In addition, querying and filtering the data by a specific gene of interest is a simple and common way for secondary use of the data. The addition of definition Gene allows to provide all desired variants, annotation and interpretation information linked to a specific gene. Examples include: calculation of allele frequencies for genetic variants affecting a specific gene, stratification of patients by variants in a specific gene, causative role of a gene in a disease phenotype etc. This is currently impossible using the existing concepts featured in the last SPHN Dataset release.

2 Comparison to other standards/data models

2.1 HL7 FHIR

Gene is represented in FHIR within an extension of the resource *Observation.* The element *Gene* is described as “A region (or regions) that includes the sequence elements necessary to encode a functional transcript. A gene may include regulatory regions, transcribed regions and/or other functional sequence regions”. FHIR recommends the use of the HUGO Gene Nomenclature Committee (HGNC) nomenclature to define human gene names.

2.2 LOINC

LOINC provides a code to identify the gene, Gene studied [ID] – 48018. This is used to list the gene(s) examined in full or in part by the study.

2.3 GA4GH Phenopackets

Born as a use case of the GA4GH initiative SchemaBlocks, Phenopackets aims at providing an open standard for sharing disease and phenotype information. Phenopackets are divided in building blocks that together forms a schema that allows the description of a "patient/sample in the context of rare disease, common/complex disease, or cancer". The concepts of *Gene* are here represented by the *Gene descriptor* building block described as such: "Gene Descriptors are used to transmit the information that the gene is thought to play a causative role in the disease phenotypes being described in cases where the exact variant cannot be provided, either for privacy reasons or because it is unknown".

3 Concept information

Contextualized concept name	Contextualized description	Type	Standard	Value set	Meaning binding
Gene	fundamental unit of heredity that contains necessary elements to encore for a transcript.				SNOMED CT: 67271001 | Gene (substance) |
gene identifier	unique gene id according to a specific nomenclature, e.g. HGNC	Code	HGNC, Ensembl or other		
organism	organism associated to the gene	Organism			
transcript	RNA product(s) of the gene	Transcript			
protein	protein product(s) of the gene	Protein			

4 Impact on the SPHN Dataset

The addition of *Gene* does not require any further change in the current SPHN Dataset release.

5 Discussion

During the development of the gene concept, we saw no reason to stray from what has been established in current standards/data models. Following a similar structure of what is already seen in both GA4GH Phenopackets and FHIR, we included the use of the HGNC (recommended) or any other formal nomenclature available from http://identifiers.org to define a unique gene identifier. This allows us to couple our simple Gene concept to any external terminology. For example, the HGNC nomenclature for human genes, scheduled to be added to the RDF schema, will allow to access a number of additional fields (e.g., approved symbol and aliases, gene name and locus) by simply using the HGNC identifier. Of note, gene symbols (e.g. *CFTR*) and names (e.g. *CF transmembrane conductance regulator*) are volatile and unreliable keys. These shall not be used to define a *Gene*. In order to increase the flexibility of the concept and pave the road for future extensions, the possibility to specify the organism of origin, as well as a number of direct products of the gene (transcripts and proteins) has been added to the concept.

6 Example

In the following example, a simple implementation of the concepts is used to describe the epidermal growth factor receptor (EGFR) gene:

Gene:

identifier:
- **identifier:** HGNC:3236
- **name:** EGFR
- **coding system and version:** HGNC-2022-11

organism:
- **identifier:** NCBITaxon:9606
- **name:** Homo sapiens
- **coding system and version:** NCBITaxon-2022

In this case, no specific transcript or protein associated to the gene was reported.