Representation of Human Genes in SNOMED CT

Version 0.1 Draft for comment

Introduction

The purpose of this proposal is to introduce a suggested structure for the proper representation of human gene names and their associated chromosomal locations in the context of body structures, specifically as subcellular structures.

It does not intend to replicate all of the genomic relationships that might be represented in a gene ontology such as HUGO, but only those necessary to position clinically relevant genes appropriately. This proposal does not include any other relationships of genes such as the association of genes or gene variants with their associated clinical conditions.

This proposal also does not address the notion of a gene as an informational artifact, but focuses on the structural relationship of a gene to its associated chromosome to allow for queries that may identify all conditions associated with mutations in a specific chromosome.

Background

At the April 2025 Pathology and Laboratory Medicine Working Group meeting in Oslo, a discussion concerning the proper placement of gene names in the SNOMED taxonomy was held. Currently, a small number of gene names are located in the Substance hierarchy under 67271001 |Gene (substance)| that are used to model 14 procedure concepts. Of the 322 subtypes of this grouper, 274 are located under 260071009 |Human leukocyte antigen allele (substance)|. Only very few other named genes are contained in this subhierarchy. Furthermore, there are no genes under 82256003 |Human gene (substance)| and 49046007 |Human structural gene (substance)|. Recent inquiries by members on the proper positioning of genes within the SNOMED taxonomy lead to the conclusion that the proper location was as a subtype of the body structure hierarchy.

The UK currently has a list of approximately 4900 gene names that it would like to add to SNOMED to support their genomics efforts. New Zealand has added 382391000210108 |Gene structure (cell structure)| and 274 gene concepts under the 312237004 |Chromosome structure (cell structure)| in their national extension. The proper structure to support the location of genes within the SNOMED hierarchy is critical to support their and other member countries efforts in the genomics domain.

Issues

In addition to the inclusion of gene names in the body structure hierarchy, the proper relationship to their chromosomes must also be considered. Currently, SNOMED CT only represents chromosomes as pairs under 91272006 |Chromosome (cell structure)|. The following issues with the current structure in SNOMED have been identified. As additional use cases are developed, more issues may arise.

1. Gene names as subtype of 115668003 |Biological substance (substance)| do not allow for their association with a particular chromosome (no existing attribute).
2. The lack of individual chromosomes does not allow for representation of heterozygosity.
3. While genes (usually) reside on both members of a chromosome pair, the precise variant of a gene may differ between the members of a chromosome pair.
4. 312237004 |Chromosome structure (cell structure)| is intended to represent the grouper of all or parts of a chromosome, although these parts do not have any subtypes, due to a lack of conformance to the SEP model for anatomy and a lack of perceived need until now.
    a. E.g. the concept 312242007 |Long arm of chromosome (cell structure)|, a subtype of 312237004 |Chromosome structure (cell structure)| has no subtypes. If this were a logical grouper, then the long arm of each individual chromosome would be represented as subtypes. These would need to be created to support the location of genes on the chromosome.
    b. While the unapproved attribute 263550009 |Arm of chromosome (attribute)| exists, it does not provide additional benefit over the existing body structure attributes under <<733928003 |All or part of (attribute)|.
5. There is a gap in adherence to the established SEP model for cell and subcellular structures. Specifically, the requirement to include an 'entire' subconcept under each 'structure' concept in the cell subhierarchy is not consistently met. The terms 'structure' and 'entire' are often absent, with different semantic tags like `(cell structure)` and `(cell)` used in their place.

Proposal

To address the need for the appropriate representation of gene names as body structures (i.e. cell structure) in SNOMED, the following changes are proposed:

1. Create individual chromosome structure concepts as subtype of 312237004 |Chromosome structure (cell structure)|. Individual “entire” chromosomes would be added (i.e. “Entire chromosome X”. As these will be extant in the Body structure hierarchy, this implies that all subtypes will represent human chromosomes.
2. Move existing Chromosome pair concepts as subtypes of their appropriate Chromosome structure concept.
3. The concept 312237004 |Chromosome structure (cell structure)| will be retained as the grouper for “all-or-part” of a chromosome and will be assigned the PT = Chromosome.

The current subtype 91272006 |Chromosome (cell structure)| will be retained and retermed “Entire chromosome” and the existing 264328003 |Whole chromosome (cell structure)| will be inactivated as duplicate to 91272006 |<Entire> Chromosome (cell structure)|. This is consistent with the SNOMED SEP model.

4. 312242007 |Long arm of chromosome (cell structure)|, 278145009 |Short arm of chromosome (cell structure)|, 70187007 |Centromere (cell structure)|, and Telomere (new concept) would remain as direct subtypes of 312237004 |Chromosome structure (cell structure)| and have “structure” added to the FSN.
5. Each part of identified human chromosomes (e.g. Chromosome 1) will be added as subtypes of the above part groupers. For the long and short arms, the terming may include the “p” and “q” representations, e.g. “Chromosome 1p” for the short arm.
6. Gene names will be added as subtypes of a new concept “Gene structure” (see 6a.), which is a subtype of 312237004 |Chromosome structure (cell structure)|.
    a. New Zealand has added a concept of “Gene structure (cellular structure) which will be requested for promotion and represent the “all-or-part-of” grouper for genes.
    b. E.g. “BRAF gene structure” IS A Gene structure (actually a proto-oncogene) and IS A Chromosome band 7q34 structure (see diagram below).
7. The current concepts 82256003 |Human gene (substance)| and 49046007 |Human structural gene (substance)| will be inactivated.
8. Human gene names will use the current name(s) as represented by the HGNC guidelines (https://www.genenames.org/about/guidelines/#!#genenames) (e.g. FSN= B-Raf proto-oncogene serine/threonine kinase gene, PT = BRAF gene)
9. Non-human genes will remain as substances until a general approach for non-human body structures has been developed.

Prior to the addition of specific gene names, the addition of individual chromosomes and their associated parts would involve 22 + X + Y chromosomes times each of the parts (long arm, short arm, centromere, and telomere). This indicates that 96 (24 x 4) new concepts would be needed to support the proposed top-level structure. If telomeres of both the long and short arm are determined to be needed, this would add an additional 48 concepts. At this point, the use case for adding specific regions of a chromosome (e.g. 11q22.3) is to support the molecular pathology concepts associated with the LOINC extension. However, the plan would only add those relevant to support this use case rather than a comprehensive listing of all karyotypic bands.

Many of the proposed changes can be added to SNOMED algorithmically (e.g. individual chromosomes, chromosome arms, centromeres, and telomeres).

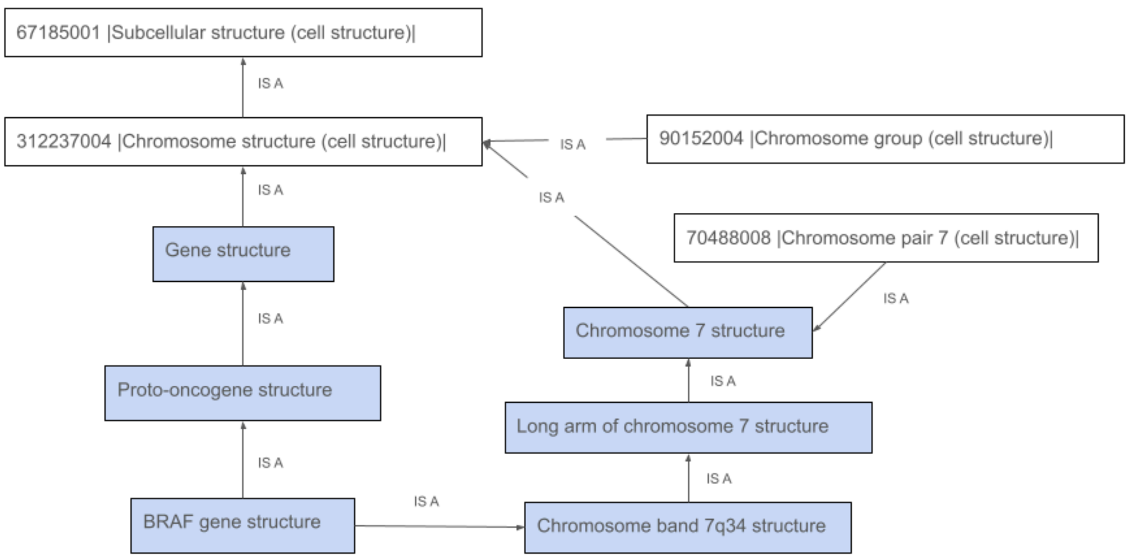
Justification for this proposal

While many references refer to genes as primarily informational units as opposed to structural entities, the addition of genes as structural components of a chromosome is consistent with the HGNC relationship of "chromosomal location" as a defining characteristic of a gene.

Diagrammatic representation of proposal

[Proposal diagrams](#) using the BRAF gene as an example. Cells in blue represent new content.

Taxonomic representation of a single gene (BRAF)



Pros and Cons for adding Gene names as Body Structures

Pros

- Allows for the use of the existing Finding site attribute and role group to represent the location and type of changes to the gene associated with a disorder. The type of change can be described with new morphologic abnormality concepts (e.g. frameshift, translocation, deletion, etc.)
- As conformant with the SNOMED SEP model for anatomy, the IS A relationship between structure concepts supports the classification hierarchy. The "part of" relationships are represented between entire concepts. This approach will utilise description logics for

automatic classification and construction of IS A relationship between structure concepts.

Cons

- Ignores the notion of genes as information units as often represented in the literature and by organizations such as the US National Center for Biotechnology Information (NCBI).
- The relationship between chromosome pairs and individual chromosomes needs to be better clarified due to their different use cases (e.g., karyotyping and genomics).

Known issues with the proposal

1. This proposal does not address the notion of a gene as an informational unit, separate from its physical relationship with its chromosome. In molecular biology, the informational aspect is often the focus rather than the gene's physical nature.
2. As the body structure hierarchy is related to human anatomy, only human genes can be represented. An alternative representation for non-human genes is needed. See Draft Editorial Guideline for creating gene concepts in SNOMED CT - V0.2, for a proposal for non-human gene concepts in SNOMED.
3. While named variants of genes can be represented in this structure, specific sequences that may be of clinical value are not addressed.
4. No specific associations of genes to their known or purported clinical conditions are addressed.
5. This proposal does not address RNA transcripts or protein translations. Proteins associated with coding genes will be addressed in a future proposal.
6. The proposal does not definitively describe the relationship of Chromosomes with other subcellular structures.

Next steps

- Seek input from the community and specifically those Members that are actively involved in representing genes in SNOMED CT.
- Implement consensus proposal in the International release of SNOMED CT.
- Add genes (or promote from extensions) currently needed to support modeling of existing international content.
- Deadline for comments is June 15, 2025

James T. Case 2025-05-08