New concept proposal

Sequencing Analysis

Author	Eelke van der Horst, Femke Kopmels	**Date last updated**	31/10/2023
Project	General interest	**Contact**	DCC
Dataset release	2024.1	**Consulted expert**	-

1 Rationale

NGS sequencing produces raw sequencing data that should be processed and analysed. There are many options to perform this processing and analysis, such as different bioinformatics pipelines and/or scripts (commonly referred to as Software). Metadata about which pipeline and version was used, as well as the used reference genome, are important to compare and evaluate the sequencing results.

2 Comparison to other standards/data models

2.1 MeSH

MeSH has 'Sequence Analysis' (MESH:D01721) which might be too broad because it includes determination of the sequence and analysis, and includes analysis of all types of macromolecules, not only nucleic acids. Its narrower terms like 'Sequence Analysis, DNA' also include the actual sequencing and preparation, instead of only analysis of sequencer output.

2.2 OBI

OBI defines the class 'sequence analysis data transformation' (OBI:0200187) which has a definition "A sequence analysis data transformation is a data transformation that has objective sequence analysis and has the aim of analysing ordered biological data for sequential patterns". This is equal to the proposed *Sequence Analysis*. Subclasses of this class are specific sequence analysis steps such as genome alignment.

2.3 EDAM

EDAM has a ‘Sequencing analysis’ operation (EDAM:operation_2403). However, this includes analysis of any macromolecule. The ‘Nucleic acid sequence analysis’ (EDAM:operation_2478) is the same as the *Sequence Analysis* concept. *Sequencing Analysis* processes may be typed by operations from EDAMs ‘Nucel acid sequence analysis’ branch. Subsequent parts of *Sequence Analysis*, which are *Data Processing* steps, may be typed by any EDAM operation.

3 Concept information

Concept or concept compositions or inherited	General concept name	General description	Contextualized concept name	Contextualized description	Type	Standard	Value set or subset	Meaning binding	Cardinality for composedOf
concept	Sequencing Analysis	analysis of the output of a nucleic acid sequencing assay	Sequencing Analysis	analysis of the output of a nucleic acid sequencing assay	Data Processing				
inherited	code	coded information specifying the concept	analysis type	code specifying the type of sequencing analysis	Code	EDAM or other	for EDAM: descendant of: EDAM:operation_2945 |Analysis|		1:1
inherited	software	software associated to the concept	software	software used in this sequencing analysis	Software				0:1
inherited	input	input associated to the concept	input	input data file	Data File				1:1
inherited	output	output associated to the concept	output	output data file	Data File				1:n
inherited	start datetime	datetime at which the concept started	start datetime	datetime at which the sequencing analysis started	temporal				0:1
inherited	quality control metric	quality control metric associated to the concept	quality control metric	quality control metric related to the output of the sequencing analysis	Quality Control Metric				0:n

inherited	standard operating procedure	standard operating procedure associated to the concept	standard operating procedure	standard operating procedure that was followed for this sequencing analysis	Standard Operating Procedure				0:1
inherited	predecessor	a preceding process associated to the concept	predecessor	process preceding this sequencing analysis	Data Processing ; Assay			RO:RO:0002087 |immediately preceded by|	0:n
composedOf	reference sequence	reference sequence used to define the coordinates of the concept	reference	sequence used as a reference in the sequencing analysis	Reference Sequence				0:1

General concept name	Cardinality for concept to Administrative Case	Cardinality for concept to Data Provider	Cardinality for concept to Subject Pseudo Identifier	Cardinality for concept to Source System
Sequencing Analysis	0:n	1:1	0:n	1:1

4 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)

5 Discussion

Sequencing Analysis is introduced as a special type of *Data Processing* that always has the aim to analyse data produced by an upstream *Sequencing Assay*, and uses a reference (except in case of *de novo* assembly). Reference is the reference genome in case of single organism sequencing, but can be any reference in case of metagenomics sequencing.

Sequencing Analysis may have many *Data Processing* parts that are executed in a sequence, which can be *Sequencing Analysis* parts themselves, such as alignment to a reference genome, or more general *Data Processing* parts, such as data transformation from SAM to BAM files.

6 Example

Variant calling example 1 using Illumina's DRAGEN pipeline

code: EDAM:operation_3227 |Variant calling|
start datetime: 2023-06-30
reference sequence:
    code:
        name: GCF_000001405.40
        coding system and version: NCBI Annotation Release 105.20220307
        identifier: GRCh37.p13
software:
    name: DRAGEN
    version: v4.0.3

output:
    data file:
        name: snp.vcf
    format code: EDAM:format_3016 |VCF|
predecessor:
    code:
        name: OBI:0002117 |whole genome sequencing assay|
    library preparation: LibraryPreparation
        library preparation kit code: Illumina TruSeq DNA PCR-Free
        intended insert size: 350
        intended read length: 150

input:
identifier: sample_1
collection datetime: 2023-06-26
material type code: 119297000 |Blood specimen|
output:
identifier: sample_2
collection datetime: 2023-06-28
material type code: 258566005 |Deoxyribonucleic acid specimen|
sample:
identifier: sample_2
collection datetime: 2023-06-28
material type code: 258566005 |Deoxyribonucleic acid specimen|
sequencing instrument:
code: OBI:0002630 |Illumina NovaSeq 6000|
sequencing run:
identifier: S0001_A0000001_NGS00001
datetime: 2023-07-04
read count: 500000
average insert size: 351.40
average read length: 156.23
mean read depth: 20.3
data file:
format code: EDAM:format_1930 |FASTQ|
quality control metric:
code: GENEPIO:0000089 |phred quality score|
quantity:
value: 78.33
unit: %
comparator: -