New concept proposal

Sequencing Assay

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

Sequencing assay metadata is essential for providing context to sequencing output, ensuring data quality, enabling data integration, and facilitating collaboration and reproducibility in genomics research. For different types of omics research, different types of assays will be relevant, and NGS sequencing has its own set of unique attributes. Therefore, we propose this *Sequencing Assay* concept, as a specialisation of *Assay* to describe sequencing experiments.

2 Comparison to other standards/data models

2.1 FAIR Genomes

FAIR Genomes defines a *Sequencing* module with elements for capturing essential metadata. It is defined as "The determination of complete sequences (typically nucleotide), including those of genomes (full genome sequencing, *de novo* sequencing and resequencing), amplicons and transcriptomes.", which is imported from the used ontology term EDAM:topic_3168 to which it is aligned. FAIR Genomes' application domain, clinical NGS data, as well as its clustering of sequencing metadata fits the SPHN genomics cases well. We therefore mirror the design of FAIR Genomes and bind the properties of Sequencing Assay and Sequencing Run to the corresponding elements in FAIR Genomes Sequencing module, such as 'read length' and 'read depth'. However, some properties that are elements of the FAIR Genomes Sequencing module, have been composed into a dedicated concept, for instance, those related to quality control metrics or the sequencing platform/model.

The 'Sequencing method' attribute defines the type of sequencing that was performed, e.g. "Next generation sequencing", "Whole genome sequencing", etc. This is a more generic qualifier for the experimental setup that is not reused here. Instead, the type of sequencing is implied by the type of library protocol of the library preparation step.

2.2 OBI and EFO

OBI includes a ‘Sequencing Assay’ class (OBI:0600047), defined as: “An assay that uses chemical or biochemical means to infer the sequence of a biomaterial”. The ‘Sequencing Assay’ itself is a subclass of ‘Assay’ and has different subclasses including ‘DNA sequencing assay’ and ‘RNA sequencing assay’. EFO defines the ‘assay by sequencer’ class (EFO:0003740) which is a specialisation of ‘assay’ (OBI:0000070), and has the exact synonym ‘sequencing assay’. It is defined as “an assay that exploits a sequencer as the instrument to find results”. This class has the equivalent meaning of the proposed *Sequencing Assay* concept.

2.3 ENA

In ENA, there is no specific concept to indicate a sequencing assay. Metadata about the sequencing performed, such as information on the instrument (platform and instrument model), as well as library preparation details, are part of the ‘Experiment’ object. ENA’s ‘Run’ object holds information about the individual runs that were performed for a specific sequencing experiment. ENA’s metadata model is too generic to reuse.

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 Assay	an assay that exploits a sequencer as the instrument to generate results	Sequencing Assay	an assay that exploits a sequencer as the instrument to generate results	Assay			EFO:0003740 |assay by sequencer|	
inherited	code	coded information specifying the concept	code	code specifying the type of sequencing assay	Code	OBI, EFO or other	for OBI: descendant of OBI:0600047 |sequencing assay|; for EFO: EFO:0003740 |assay by sequencer|		1:1
inherited	identifier	unique identifier identifying the concept	identifier	unique identifier identifying the sequencing assay	string				0:1
inherited	start datetime	datetime at which the concept started	start datetime	datetime at which the sequencing assay was first executed	temporal				0:1
inherited	standard operating procedure	standard operating procedure associated to the concept	standard operating procedure	standard operating procedure that was followed for this sequencing assay	Standard Operating Procedure				0:1
inherited	data file	data file associated to the concept	data file	data file associated to the sequencing assay	Data File				0:n

inherited	sample	any material sample for testing, diagnostic, propagation, treatment or research purposes associated to the concept	sample	material that is being sequenced by this sequencing assay	Sample				0:n
inherited	predecessor	a preceding process associated to the concept	predecessor	sample processing preceding the sequencing assay	Sample Processing			RO:0002087 | immediately preceded by |	0:n
composedOf	library preparation	library preparation associated to the concept	library preparation	the library preparation that is part of the sequencing assay	Library Preparation				0:1
composedOf	sequencing instrument	device associated to the concept	sequencing instrument	the device which is used to perform the sequencing assay	Sequencing Instrument				0:1
composedOf	sequencing run	sequencing run associated to the concept	sequencing run	sequencing run performed as part of the sequencing assay	Sequencing Run				0:n
composedOf	intended read length	intended read length associated to the concept	intended read length	the number of nucleotides intended to be ordered from each side of a nucleic acid fragment obtained after the completion of a sequencing assay	Quantity				0:1
composedOf	intended read depth	intended read depth associated to the concept	intended read depth	the number of times a particular locus (site, nucleotide, amplicon, region) was intended to be sequenced as part of the sequencing assay	Quantity				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 Assay	0:n	1:1	0:n	1:1

4 Impact on the SPHN Dataset

-

5 Discussion

For different types of omics research, different types of assays will be relevant, each with their unique set of attributes. Therefore, we propose this *Sequencing Assay* concept, as a type of *Assay* that is specific to sequencing.

Library preparation is an essential part of a sequencing assay. We therefore add a dedicated property. If a Library Preparation is recorded, it also implies that it is a 'part' of the *Sequencing Assay*.

A sequencing assay may produce multiple data files, either different files for a single run or multiple runs. It is possible to define Run-specific information using *Sequencing Run*, or leave this information out. If a datafile is produced by a sequencing run, it follows that it is also related to the parent *Sequencing Assay*.

When multiple runs are executed for the same *Sequencing Assay*, the onset datetime for this concept will be equal to the run datetime of the run that was first executed.

6 Example

```
code:
        name: OBI:0002117 |whole genome sequencing assay|
identifier: example_assay123
start datetime: 2023-07-04
library preparation:
        library preparation kit: Illumina TruSeq DNA PCR-Free
sample:
        identifier: sample_1
        collection datetime: 2023-06-28
        material type code: 258566005 |Deoxyribonucleic acid specimen|
sequencing instrument:
        code: OBI:0002630 |Illumina NovaSeq 6000|
intended read length:
        value: 150
        unit: {#}
intended read depth:
        value: 20
        unit: {#}
sequencing run:
        identifier: S0001_A0000001_NGS00001
        datetime: 2023-07-04
```

```
read count:
        value: 500000
        unit: {#}
average insert size:
        value: 351.40
        unit: {#}
average read length:
        value: 156.23
        unit: {#}
mean read depth:
        value: 20.3
        unit: {#}
data file:
        format code: EDAM: format:1930 |FASTQ|
quality control metric:
        code: GENEPIO:0000089  |phred quality score|
        quantitative value:
                value: 78.33
                unit: %
```