

New concept proposal

Sequencing Run

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

A discrete execution of a Sequencing assay, is a Sequencing run. Re-executing a specific assay using the same settings and on the same sample, results in a new run of the same assay, and the results of individual runs may differ. Metadata about sequencing runs provides the essential context for understanding results and to assess the quality of the produced sequencing data. To record metadata at the run level of the same assay, the *Sequencing Run* concept is proposed.

2 Comparison to other standards/data models

2.1 ENA



ENA's 'Run' concept is connected to an 'Experiment' and can hold information about the number of reads, sequencing date, and sequencing location. Similar to the design of the ENA model, we propose to separate information that is equal for all runs, such as experimental setup, from information related to a specific run.

2.2 EGA

In EGA, 'Samples' are not directly linked to a 'Dataset' and 'Runs'. However, 'Runs', 'Studies' and 'Samples' can be linked to an 'Experiment'. Similar to EGA, our proposed *Sequencing Run* is not directly connected to a *Sample* but via an *Assay* (which is structurally similar to 'Experiment' in EGA).

2.3 FAIR Genomes

The FAIR Genomes data model does not have a specific 'Run' concept but stores properties that you would expect in such a concept under 'Sequencing':

A project of	 <p>Schweizerische Akademie der Medizinischen Wissenschaften Académie Suisse des Sciences Médicales Accademia Svizzera delle Scienze Mediche Swiss Academy of Medical Sciences</p>	 <p>Swiss Institute of Bioinformatics</p>	<p>SIB Swiss Institute of Bioinformatics PHI Personalized Health Informatics Group www.sphn.ch dcc@sib.swiss</p>
--------------	---	--	--

- SequencingIdentifier
- SequencingData
- MedianReadDepth
- ObservedReadLength
- ObservedInsertSize
- PercentageQ30
- PercentageTr20
- OtherQualityMetrics

We introduced properties similar to ‘MedianReadDepth’, ‘ObservedReadLength’ and ‘ObservedInsertSize’ in our *Sequencing Run* concept. For the quality metric related to a specific sequencing run, we proposed a separate *Quality Control Metric* concept.

2.4 Genomics England

Genomics England reuses the GA4GH data model. Their ‘Reads’ concept is comparable to ‘Run’ in other data models. A ‘ReadGroup’ is defined as all the data that is processed the same way by the sequencer. This concept holds, for example, the predicted insert size and the ID of the reference set that the reads in the group are aligned to.

Similar to Genomics England, we propose to allow grouping of *Sequencing Runs*. In our proposal, all the runs that need to be grouped together will be connected to the same *Sequencing Assay*.

2.5 Minimal Information for Reporting a Genomics Experiment Standard

The “[Minimal information for reporting a genomics experiment](#)” paper from Kostiantyn Dreval et al. (2022) presents recommended information for reporting sample sequencing details. The listed properties were considered for inclusion in the proposed *Sequencing Run*, *Assay* and *Sample Processing* concepts. One notable difference between this standard and other models is that ‘average insert size’ and ‘average read length’ here are attributed to Quality Metric. Generally they fall under “*Run-like*” concepts in other models.

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 composed Of
concept	Sequencing Run	the valid and completed operation of a high-throughput sequencing instrument associated with a sequencing assay	Sequencing Run	the valid and completed operation of a high-throughput sequencing instrument associated with a sequencing assay	SPHNConcept			NCIT:C148088 [Sequencing Run]	
composedOf	identifier	unique identifier identifying the concept	identifier	unique identifier identifying the sequencing run	string				0:1
composedOf	datetime	datetime of the concept	datetime	datetime the sequencing run was performed	temporal				0:1
composedOf	read count	ready count associated to the concept	read count	the number of sequencing reaction results that were pooled to assemble a sequence for a genomic region of interest in a sequencing run	Quantity				0:1
composedOf	average insert size	average insert size associated to the concept	average insert size	the average insert size found during the nucleic acid sequencing run	Quantity				0:1
composedOf	average read length	average read length associated to the concept	average read length	the average length for nucleic acid sequencing reads generated in a sequencing run	Quantity				0:1

composedOf	mean read depth	mean read depth associated to the concept	mean read depth	the number of times a particular locus (site, nucleotide, amplicon, region) was sequenced in a sequencing run	Quantity				0:1
composedOf	data file	data file associated to the concept	data file	data file associated to the sequencing run	Data File				1:n
composedOf	quality control metric	quality control metric associated to the concept	quality control metric	quality control metric associated with the sequencing run	Quality Control Metric				1:n

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 Run	-	1:1	-	-

4 Impact on the SPHN Dataset

-

5 Discussion

During the development of this concept proposal, there was a thorough comparison between the existing data models and how they store information related to (sequencing) runs. Models often call the concept 'Run', but this naming might not necessarily be applicable to all our use cases. Also, a 'Run' concept in the context of sequencing needs different attributes than one in the context of (bioinformatics) data analysis for example. Therefore, we aimed for an intermediate that captures information about the action.

The proposed *Sequencing Run* concept holds information about results of sequencing runs. In the proposed Omics extension to the SPHN model, a *Sequencing Assay*, of type *Assay*, can be connected to a *Sequencing Run*. Assays, in turn, are connected to a *Sample*. Information related to library preparation will be connected to this *Sample* and not be included in the *Sequencing Run* because it will be the same for all *Sequencing Runs* of the same *Sequencing Assay* that are all performed on the same *Sample*. Attributes of the *Library Preparation* concept includes not only the library preparation kit and target enrichment kit that were used, but also the intended insert size and intended read length. The average insert size and read length are attached to the *Sequencing Run* concept because this can differ on a per-run basis. Information about the instrument that was used for a sequencing run is connected to the (*Sequencing*) *Assay* because it is assumed that this will be the same for all sequencing runs of a particular *Assay*.

Multiple libraries could be sequenced in the same run (on the same machine). In theory, these could be one from WES and one Amplicon etc. For these situations, multiple *Sequencing Assays* can be linked to the same *Sequencing Run*.

6 Example

Example of a NGS sequencing run

identifier: S0001_A0000001_NGS00001

datetime: 2023-07-04

read count:

value: 500000

unit: {#}

comparator: -


average insert size:

value: 351.40

unit: {#}

comparator: -

average read length:

A project of	 <p>Schweizerische Akademie der Medizinischen Wissenschaften Académie Suisse des Sciences Médicales Accademia Svizzera delle Scienze Mediche Swiss Academy of Medical Sciences</p>	 <p>Swiss Institute of Bioinformatics</p>	<p>SIB Swiss Institute of Bioinformatics PHI Personalized Health Informatics Group www.sphn.ch dcc@sib.swiss</p>
--------------	---	--	--

value: 156.23
 unit: {#}
 comparator: -
 mean read depth:
 value: 20.3
 unit: {#}
 comparator: -
 data file:
 format code: EDAM:format_1930 |FASTQ|
 quality control metric:
 code: GENEPIO:0000089 |phred quality score|
 quantity
 value: 78.33
 unit: %
 comparator: -