# New concept proposal
# Library Preparation

| 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

*Library Preparation* is a type of *Sample Processing* that is essential to prepare a DNA or RNA sample for sequencing using a next generation sequencing assay. It usually consists of isolating nucleic acids from the input sample, fragmenting the targeted sequences, attaching adapter sequences, and optional amplification, resulting in a sample of DNA fragments i.e. the NGS library. This library is then sequenced using the sequencing instrument of the sequencing assay. Participants that characterise this process are the Library preparation kit and Target enrichment kit, i.e. ready-to-use kits consisting of reagents and a fixed protocol to perform a step in library preparation. The choice of these kits depend on the intended type of NGS experiment, i.e. whole genome sequencing, whole exome sequencing, or targeted sequencing, and the sequencing platform that will perform the subsequent sequencing assay. Other essential attributes are the intended insert size (for paired-end sequencing) and whether PCR amplification was used.

## 2    Comparison to other standards/data models

### 2.1   FAIR Genomes

FAIR Genomes has a 'Sample preparation' module which describes the same experimental step as *Library Preparation*; a meaning binding between the 'fg:Sample_preparation' class and the *Library Preparation* concept is therefore proposed. While the label 'Sample preparation' in the context of FAIR Genomes can only mean preparation of a sample for nucleic acid sequencing, this label is ambiguous in the context of the SPHN, since it can also mean preparation for other (omics) assays. We therefore denote this concept with a more specific label 'Library Preparation'. The elements for the library preparation kit and target enrichment kit from FAIR Genomes are mirrored in this concept, reusing the value sets as defined by FAIR Genomes. The 'intended insert size' is also mirrored from FAIR Genomes.

*Note: FAIR Genomes is referred to in the SPHN meaning bindings with the prefix 'fg'.*

## 2.2 ENA

The ENA 'Experiment' object has fields that describe the library preparation, including insert size, library name, library source, library selection, library strategy, and an optional free text library description. However, information on library source is part of the upstream sample processing and specimen, while the library strategy is reflected in the type of sequencing assay that this library preparation is part of. Any special library selection is either implicit to the NGS kits used or should be indicated with a preceding sample processing step.

## 2.3 OBI and EFO

The Ontology of Biomedical Investigations (OBI) has a 'sample preparation for sequencing assay' (OBI:0001902) concept, defined as "A sample preparation for assay that preparation of nucleic acids for a sequencing assay". This corresponds to *Library Preparation* in case of NGS sequencing, however, the OBI class might also be used for non-NGS-based sequencing. It also has the 'library preparation' class (OBI:0000711). EFO reuses both OBI classes, however, defines more specific subclasses for library preparation than OBI, for instance 'cDNA library construction' and 'single cell library construction' (and more specific 'CITE-seq').

# 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 | Library Preparation | process which results in the creation of a library from fragments of DNA | Library Preparation | process which results in the creation of a library from fragments of DNA | Sample Processing | | | OBI:0000711 \| library preparation \| | |
| composedOf | kit code | coded information specifying the kit associated to the concept | library preparation kit code | pre-filled, ready-to-use reagent cartridges intended to improve chemistry, cluster density and read length as well as improve quality (Q) scores for this sample. Reagent components are encoded to interact with the sequencing system to validate compatibility with user-defined applications. | Code | EFO, GENEPIO, FAIRGenomes or other | | fg:Sample_preparation_Library_preparation_kit | 0:1 |
| composedOf | target enrichment kit code | coded information specifying the target enrichment kit | target enrichment kit code | indicates which target enrichment kit was used to prepare this sample. Target enrichment is a pre-sequencing DNA preparation step where DNA sequences are either directly amplified (amplicon or multiplex PCR-based) or captured (hybrid capture-based) in | Code | EFO, GENEPIO, FAIRGenomes or other | | fg:Sample_preparation_Target_enrichment_kit | 0:1 |

| | | associated to the concept | | order to only focus on specific regions of a genome or DNA sample. | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| composed Of | intended insert size | intended insert size associated to the concept | intended insert size | in paired-end sequencing, the DNA between the adapter sequences is the insert. The length of this sequence is known as the insert size, not to be confused with the inner distance between reads. So, fragment length equals read adapter length (2x) plus insert size, and insert size equals read length (2x) plus inner distance. | Quantity | | | | 0:1 |
| composed Of | gene panel | gene panel associated to the concept | gene panel | collection of genes that are the focus of sequencing | Gene Panel | | | | 0:1 |
| inherited | code | coded information specifying the concept | code | code specifying the type of library preparation | Code | OBI, EFO or other | for OBI: descendant of: OBI:0000711 \|library preparation\| | | 0:1 |
| inherited | standard operating procedure | standard operating procedure associated to the concept | standard operating procedure | standard operating procedure that was followed for this library preparation | Standard Operating Procedure | | | | 0:1 |
| inherited | input | input associated to the concept | input | the sample for which a library is created | Sample | | | | 0:n |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| inherited | output | output associated to the concept | output | the NGS library that is produced | Sample | | | | 0:1 |
| inherited | start datetime | datetime at which the concept started | start datetime | start of library preparation | temporal | | | | 0:1 |
| inherited | predecessor | a preceding process associated to the concept | predecessor | process preceding this library preparation | Sample Processing | | | RO:0002087 \|immediately preceded by\| | 0:n |
| inherited | quality control metric | quality control metric associated to the concept | quality control metric | quality control metric related to the output of the library preparation | Quality Control Metric | | | EDAM:data_3914 \|quality control report\| | 0: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 |
|---|---|---|---|---|
| Library Preparation | 0:n | 1:1 | 0:n | 1:n |

# 4 Impact on the SPHN Dataset

-

# 5 Discussion

Library preparation can be considered an integral part of performing an NGS sequencing experiment. However, since the configuration/parameters of these experimental processes can vary separately, and also because they can be executed at different facilities at different times by different people, while also having their independent quality control metrics, we represent these as separate concepts that are part of the *Sequencing Assay*. Note that the parent *Sequencing Assay* and its *Sequencing Instrument* are tightly bound to this concept, because these influence the possible choices for the kits used for library preparation and constrains the intended read length and insert size.

An alternative naming that was considered is 'sample preparation for sequencing assay'. However, in the context of Omics/NGS sequencing, library preparation seemed more specific.

Note that as with other (experimental) processes, individual sub-steps can be provided, for instance for DNA extraction or amplification.

Since many of the composedOf concepts might not be known, for instance which kit was used, these have minimum cardinality 0.

# 6 Example

*Library preparation as part of a WGS sequencing assay*

preparation kit code: Illumina TruSeq DNA PCR-Free
start datetime: 2023-06-28
intended insert size
      value: 350
      unit: {#}
input:
      identifier: DNASample1
      collection datetime: 2023-06-26
      material type code: 258566005 |Deoxyribonucleic acid specimen|
output:
      identifier: Lib1
      collection datetime: 2023-06-28
      material type code: 258566005 |Deoxyribonucleic acid specimen|
predecessor:
      code: 104166004 |Nucleic acid molecular isolation or extraction method (procedure)|
      input:

identifier: BloodSample1
collection datetime: 2023-06-26
material type code: 19297000 |Blood specimen|

output:

identifier: DNASample1
collection datetime: 2023-06-26
material type code: 258566005 |Deoxyribonucleic acid specimen (specimen)|

standard operating procedure:

name: DNA isolation from blood sample
description: To describe how DNA from blood must be isolated and stored.
version: V.1.1
data file:

format code: EDAM: format_3508 |PDF|
code: EXAMPLE01