FAANG metadata - experiment specification

This document describes the specification for all experiment metadata. You can find an overview of our metadata and archival plans in the overview document (faang_metadata_overview.md) . The sample (faang_sample_metadata.md) and analysis (faang_analysis_metadata.md) documents are also in this git repo (https://github.com/FAANG/faang-metadata) .

Experiments are expected to fall into two categories:

1. sequencing experiments, archived in an SRA database (hosted at (EMBL-EBI) [https://www.ebi.ac.uk/ena (https://www.ebi.ac.uk/ena)], NCBI (http://www.ncbi.nlm.nih.gov/sra/) and DDBJ (http://trace.ddbj.nig.ac.jp/dra/index_e.html)). Some of these submissions may be brokered by specialist services such as ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) and GEO (http://www.ncbi.nlm.nih.gov/geo/)

2. array experiments, archived in ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) or GEO (http://www.ncbi.nlm.nih.gov/geo/) .

Experiment metadata requirements

Requirements are laid out like this:

- `attribute name` (*data type*) a brief description

The data types will be described later in this document. The metadata group will seek guidance from the assay group on what needs to be recorded here for each assay type.

Each assay type will require metadata in addition to the core set of common attributes. The initial set proposed is based upon the IHEC metadata standards (http://ihec-epigenomes.org/research/reference-epigenome-standards/)

Common

Recquired:

These following elements must always be present in any experiment metadata

- Each experiment should reference one `sample`. The metadata for that sample should comply with the FAANG metadata specification.
  - This should be a reference to the BioSample ID for the specimen, purified cell, cultured cell or cell line the experiment was conducted on.
- `assay type` (*ontology term*) The class of experiment performed. e.g. RNA-Seq or expression array. This should be a child term of EFO:0002773 (http://www.ebi.ac.uk/efo/EFO_0002773)
- `experiment target` (*ontology term*) What is the experiment trying to find?

- ChIP-seq for histone modifications should use a child term of histone modification (http://purl.obolibrary.org/obo/SO_0001700)
    - ChIP-seq input should use the term input DNA (http://www.ebi.ac.uk/efo/EFO_0005031)
    - RNA-seq should use a child term of RNA (http://purl.obolibrary.org/obo/CHEBI_33697)
    - ATAC-seq and DNase-seq should use the term open_chromatin_region (http://purl.obolibrary.org/obo/SO_0001747)
    - Methylation assays should use the term DNA methylation (http://purl.obolibrary.org/obo/GO_0006306)
- `sample storage` (*text*) This should document how the sample was stored, from one of these values:

```
*   frozen, liquid nitrogen
*   frozen, -70 freezer
*   frozen, vapor phase
*   RNAlater, frozen
*   paraffin block
*   cut slide
*   fresh
```

  - `sample storage processing` (*text*) This should document how the sample was prepared for storage, from one of these values:
      - cryopreservation in liquid nitrogen (dead tissue)
      - cryopreservation in dry ice (dead tissue)
      - cryopreservation of live cells in liquid nitrogen
      - cryopreservation, other
      - formalin fixed, unbuffered
      - formalin fixed, buffered
      - formalin fixed and paraffin embedded
      - fresh
  - `sampling to preparation interval` (*number*) This should list how long between the sample being taken and used in the experiment. Units should be specified, and be either 'minutes','hours','days','weeks' or 'years'.
  - `experimental protocol` (*protocol*) a description of the experiment protocol
- `extraction protocol` (*protocol*) the protocol used to isolate the extract material

Recommended:

- library preparation location (*location*)
    - `library preparation location` (*text*) name of the library preparation location
    - `library preparation location latitude` (*number*) latitude of the library prep. location in decimal degrees. Units should be specified as 'decimal degrees'
    - `library preparation location longitude` (*number*) longitude of the library prep.

location in decimal degrees. Units should be specified as 'decimal degrees'

- `library preparation date` (*date*) Date on which the library was prepared, formatted as YYYY-MM-DD. Units should be specified as 'YYYY-MM-DD'
- sequencing location (*location*)
  - `sequencing location` (*text*) name of the sequencing location
  - `sequencing location latitude` (*number*) latitude of the sequencing location in decimal degrees. Units should be specified as 'decimal degrees'
  - `sequencing location longitude` (*number*) longitude of the sequencing. location in decimal degrees. Units should be specified as 'decimal degrees'
- `sequencing date` (*date*) date of sequencing

Whole-genome bisulfite sequencing

WGBS experiments should have an `assay type` of methylation profiling by high throughput sequencing (http://www.ebi.ac.uk/efo/EFO_0002761) and an `experiment target` of DNA methylation (http://purl.obolibrary.org/obo/GO_0006306) .

Required:

- `bisulfite conversion protocol` (*protocol*)
- `pcr product isolation protocol` (*protocol*) the protocol for isolating PCR products used for library generation
- `bisulfite conversion percent` (*number*) bisulfite conversion percent (between 0 and 100)

ChIP-seq

ChIP-seq experiments should have an `assay type` of ChIP-seq (http://www.ebi.ac.uk/efo/EFO_0002692) . The `experiment target` is variable. Use a child term of SO:0001700 (http://www.sequenceontology.org/browser/current_svn/term/SO:0001700) for histone modifications.

Examples of the antibody information are from the H3K4me3 antibody from Diagenode (https://www.diagenode.com/p/h3k4me3-polyclonal-antibody-premium-50-ug-50-ul) , used by the BLUEPRINT project.

Required:

- `chip protocol` (*protocol*) the ChIP protocol used
- `chip antibody provider` (*text*) the name of the company, laboratory or person that provided the antibody e.g. Diagneode
- `chip antibody catalog` (*text*) the catalog from which the antibody was purchased e.g. pAb-003-050
- `chip antibody lot` (*text*) the lot identifier of the antibody e.g. A5051-001P
- `library generation max fragment size range` (*number*) the maximum fragment size range of the preparation

- `library generation min fragment size range` (*number*) the minimum fragment size range of the preparation

ChIP-seq input

ChIP-seq experiments should have an `assay type` of ChIP-seq (http://www.ebi.ac.uk/efo/EFO_0002692) and an `experiment target` of Input DNA (http://www.ebi.ac.uk/efo/EFO_0005031) for ChIP input sequencing.

Required:

- `chip protocol` (*protocol*) the ChIP protocol used

- `library generation max fragment size range` (*number*) the maximum fragment size range of the preparation

- `library generation min fragment size range` (*number*) the minimum fragment size range of the preparation

RNA-seq

RNA-seq experiemnts should have an `assay type` of one of the following:

- RNA-seq of coding RNA (http://www.ebi.ac.uk/efo/EFO_0003738)
- RNA-seq of non coding RNA (http://www.ebi.ac.uk/efo/EFO_0003737)
- microRNA profiling by high throughput sequencing (http://www.ebi.ac.uk/efo/EFO_0002896)

The `experiment target` should be one of the following:

- polyA RNA (http://purl.obolibrary.org/obo/OBI_0000869)
- total RNA (http://www.ebi.ac.uk/efo/EFO_0004964)
- ncRNA (http://purl.obolibrary.org/obo/SO_0000655)
- microRNA (http://purl.obolibrary.org/obo/SBO_0000316)

Required:

- `rna preparation 3' adapter ligation protocol` (*protocol*) the protocol for 3' adapter ligation used in preparation

- `rna preparation 5' adapter ligation protocol` (*protocol*) the protocol for 5' adapter ligation used in preparation

- `library generation pcr product isolation protocol` (*protocol*) the protocol for isolating pcr products used for library generation

- `preparation reverse transcription protocol` (*protocol*) the protocol for reverse transcription used in preparation

- `library generation protocol` (*protocol*) the protocol used to generate the library

- `read strand` (*text*) where a strand specific protocol is used, specify which mate pair maps to the transcribed strand. Report 'not applicable' if the protocol is not strand specific. Possible values:

- 'not applicable' if the protocol is not strand specific
  - single-ended sequencing:
    - 'sense' if the reads should be on the same strand as the transcript
    - antisense if the read should be on the opposite strand of the transcript
  - paired-end sequencing:
    - 'mate 1 sense' if mate 1 should be on the same strand as the transcript
    - 'mate 2 sense' if mate 2 should be on the same strand as the transcript
- `rna purity - 260:280 ratio` (*number*) sample purity assesed with fluoresence ratio at 260 and 280nm, informative for protein contamination
- `rna purity - 260:230 ratio` (*number*) Sample purity assesed with fluoresence ratio at 260 and 230nm, informative for contamination by phenolate ion, thiocyanates, and other organic compounds
- `rna integrity number` (*number*) See Schroeder *et al* , 2006 (http://www.biomedcentral.com/1471-2199/7/3)

DNase-Hypersensitivity seq

DNase-seq experiments should have an `assay type` of DNase-Hypersensitivity seq (http://www.ebi.ac.uk/efo/EFO_0003752) and an `experiment target` of open chromatin region (http://purl.obolibrary.org/obo/SO_0001747)

Required:

- `dnase protocol` (*protocol*) the protocol used for DNAse treatment

ATAC-seq

ATAC-seq experiments should have an `assay type` of ATAC-seq (http://www.ebi.ac.uk/efo/EFO_0007045) (due in release 2.69) and an `experiment target` of open chromatin region (http://purl.obolibrary.org/obo/SO_0001747)

Required:

- `transposase protocol` (*protocol*) the protocol used for transposase treatment

Hi-C

Hi-C experiments should have an `assay type` of HiC (http://purl.obolibrary.org/obo/OBI_0002042) and an `experiment target` of chromosome conformation (http://purl.obolibrary.org/obo/OBI_0001917) .

Required:

- restriction enzyme (*text*)
- restriction site (*text*)

Data types for experiment attributes

SRA databases (ENA , NCBI, DDBJ) takes experiment records with a set of attributes. Each attribute has a name and a value, and can also have units. In contrast with the BioSamples (www.ebi.ac.uk/biosamples) database, they do not have direct support for ontology terms. The following section describe the expectations for each data type within FAANG.

date

Dates should be reported in the ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601) format, YYYY-MM-DD. To ensure clarity, the format should be reported as the 'units'.

number

A number, with units specified. BioSamples recommends that units are given without abbreviations .For example, a birth weight could have a value of 1.3 and the units specified as 'kilograms'.

protocol

A URL link to a protocol document on the FAANG FTP site. Please contact the FAANG data coordination centre (mailto:faang-dcc@ebi.ac.uk) to have your protocol documents added to the FTP site.

text

Text, using US English spellings.

URL

A URL, such as 'http://faang.org/ (http://faang.org/) '. Depending on the context, http, ftp, mailto links may be appropriate. Examples:

- ftp, ftp://ftp.faang.ebi.ac.uk/ftp/README
- http, http://faang.org/ (http://faang.org/)
- mailto, mailto:bob@example.org

ontology term

The text label of a term from an ontology. The attribute value should be the term label. Unlike for sample submissions, direct links to ontologies cannot be submitted as attributes. The attribute value should exactly match the term name in the ontology.

Missing data

Where data cannot be included in a submission, submit one of these text values instead

- 'not applicable'
- 'not collected' (i.e. will always be missing)
- 'not provided' (i.e. may be added later)
- 'restricted access' (i.e. it isn't missing, we just can't include it in a public document)

The use of these values will interact with the metadata validation system as follows:

- attribute is required
  - not applicable, not collected, not provided – validation will regard these as an error
  - restricted access – validation will generate a warning
- attribute is recommended
  - not applicable, not collected, not provided – validation will generate a warning
  - restricted access – pass
- attribute is optional
  - validation will pass with any of missing values terms

Submission

Each experiment record should reference a record in BioSamples. These have accessions like SAMEA1234567