

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](#). The [sample](#) and [analysis](#) documents are also in this [git repo](#). Further guidance can be found on the [FAANG wiki pages](#), with specific guidance for submission of [sequencing data](#).

Experiments are expected to fall into two categories:

1. sequencing experiments, archived in an SRA (Sequence Read Archive) database (hosted at [EMBL-EBI](#), [NCBI](#) and [DDBJ](#)). Some of these submissions may be brokered by specialist services such as [ArrayExpress](#) and [GEO](#)
2. array experiments, archived in [ArrayExpress](#) or [GEO](#).

Experiment metadata requirements

Requirements are laid out like this:

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

The details of data types can be found [here](#).

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](#) database, they do not have direct support for ontology terms.

The metadata & data sharing (M&DS) group will seek guidance from the animals, samples and assays (ASA) 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](#)

Common

Required:

These following elements must always be present in any experiment metadata

- **sample** (*BioSample ID*) the BioSamples ID for the specimen (specimen from organism, cell specimen, cell cultur etc.) the experiment was conducted on. Each experiment must reference at least one FAANG BioSample
- **assay type** (*ontology term*) The class of experiment performed. e.g. RNA-seq of coding RNA. This should be one of the following terms:
 - ATAC-seq
 - ChIP-seq

- DNase-Hypersensitivity seq
- HiC
- methylation profiling by high throughput sequencing
- microRNA profiling by high throughput sequencing
- RNA-seq of total RNA
- RNA-seq of coding RNA
- RNA-seq of non coding RNA
- transcription profiling by high throughput sequencing
- WGS
- CAGE-seq
- **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’.
- **extraction protocol** (*protocol*) the protocol used to isolate the extract material

Recommended:

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

Optional:

- **sample storage** (*text*) This should document how the sample was stored, from one of these values:
 - ambient temperature
 - cut slide
 - fresh
 - frozen, -70 freezer
 - frozen, -150 freezer
 - frozen, liquid nitrogen
 - frozen, vapor phase
 - RNAlater, frozen
 - TRIzol, frozen
 - paraffin block
- **experimental protocol** (*protocol*) a description of the experiment protocol

ATAC-seq

ATAC-seq experiments should have an **assay type** of [ATAC-seq](#)

Required:

- **experiment target** (*ontology term*) Should use the term [open chromatin region](#)
- **transposase protocol** (*protocol*) the protocol used for transposase treatment

Bisulfite sequencing

Whole Genome Bisulfite Sequencing (WGBS) and Reduced Representation Bisulfite Sequencing (RRBS) experiments should have an **assay type** of [methylation profiling by high throughput sequencing](#)

Required:

- **experiment target** (*ontology term*) Should use the term [DNA methylation](#)
- **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)

Recommended:

- **restriction enzyme** (*text*) Restriction enzyme used for Reduced representation bisulfite sequencing

- **max fragment size selection range** (*number*) The maximum fragment size of the fragment selection range
- **min fragment size selection range** (*number*) The minimum fragment size of the fragment selection range

ChIP-seq standard rules

ChIP-seq experiments should have an **assay type** of [ChIP-seq](#).

Examples of the antibody information are from the [H3K4me3 antibody from Diagenode](#), used by the BLUEPRINT project.

Required:

- **experiment target** (*ontology term*)
- ChIP-seq for histone modifications should use [histone modification](#)
- ChIP-seq for Transcription factor should use [TF binding site](#)
- ChIP-seq input should use the term [input DNA](#)
- **chip protocol** (*protocol*) the ChIP protocol used

ChIP-seq for DNA-binding proteins

ChIP-seq DNA-binding proteins experiments should have an **assay type** of [ChIP-seq](#)

Required:

- **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

Recommended:

- **control experiment** (*text*) Experiment alias (in this submission) or ENA experiment accession (if submitted previously) of the ChIP-seq input DNA experiment which acts as the control/background for the analysis.

Optional:

- **target protein name** (*ontology term*) State the name of targeted protein

ChIP-seq input DNA

ChIP-seq input experiments should have an **assay type** of [ChIP-seq](#) and an **experiment target** of [Input DNA](#) for ChIP input sequencing.

Required:

- **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

DNase-Hypersensitivity seq

DNase-seq experiments should have an **assay type** of [DNase-Hypersensitivity seq](#)

Required:

- **experiment target** (*ontology term*) Should use the term [open chromatin region](#)
- **dnase protocol** (*protocol*) the protocol used for DNase treatment

Hi-C

Hi-C experiments should have an **assay type** of [Hi-C](#)

Required:

- **experiment target** (*ontology term*) Should use the term of [chromosome conformation](#)
- **restriction enzyme** (*text*)
- **restriction site** (*text*)

RNA-seq

RNA-seq experiments should have an **assay type** of one of the following:

- [RNA-seq of coding RNA](#)
- [RNA-seq of non coding RNA](#)
- [microRNA profiling by high throughput sequencing](#)

Required:

- **experiment target** (*ontology term*) Should be one of the following:
- [polyA RNA](#)
- [total RNA](#)
- [ncRNA](#)

- [microRNA](#)
- **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

Recommended:

- **rna purity - 260:280 ratio** (*number*) sample purity assessed with fluorescence ratio at 260 and 280nm, informative for protein contamination
- **rna purity - 260:230 ratio** (*number*) Sample purity assessed with fluorescence ratio at 260 and 230nm, informative for contamination by phenolate ion, thiocyanates, and other organic compounds
- **rna integrity number** (*number*) It is important to obtain this value, but if you are unable to supply this number (e.g. due to machine failure) then by submitting you are asserting the quality by visual inspection of traces and agreeing that the samples were suitable for sequencing. See [Schroeder et al , 2006](#)

WGS

Whole Genome Sequencing should have an **assay type** of [whole genome sequencing assay](#)

Required:

- **experiment target** (*ontology term*) should use the term [input DNA](#)

- **library generation pcr product isolation protocol** (*protocol*) the protocol for isolating pcr products used for library generation
- **library generation protocol** (*protocol*) link to the protocol used to generate the library

Optional:

- **library selection** (*text*) State whether reduced representation was used in the protocol from one of these values:
 - reduced representation
 - none

CAGE-seq

CAGE-seq (Cap analysis gene expression sequencing) should have **assay type** of CAGE-seq

Required:

- **experiment target** (*ontology term*) should use the term [ribonucleic acid](#)
- **restrict enzyme target sequence** (*protocol*) the DNA sequence targeted by the restrict enzyme

Recommended:

- **sequencing primer provider** (*text*) The name of the company, laboratory or person that provided the sequencing primer library
- **sequencing primer catalog** (*text*) The catalog from which the sequencing primer library was purchased
- **sequencing primer lot** (*text*) The lot identifier of the sequencing primer library
- **rna purity - 260:280 ratio** (*number*) sample purity assessed with fluorescence ratio at 260 and 280nm, informative for protein contamination
- **rna purity - 260:230 ratio** (*number*) Sample purity assessed with fluorescence ratio at 260 and 230nm, informative for contamination by phenolate ion, thiocyanates, and other organic compounds
- **rna integrity number** (*number*) It is important to obtain this value, but if you are unable to supply this number (e.g. due to machine failure) then by submitting you are asserting the quality by visual inspection of traces and agreeing that the samples were suitable for sequencing. See [Schroeder et al , 2006](#)

Missing data

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

- ‘not applicable’ (i.e. does not apply to this experiment)

- ‘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 collected, not provided - validation will generate a warning
- restricted access, not applicable - 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. As described above, experiments themselves should be submitted to the appropriate EMBL-EBI, NCBI or DDBJ assay archives.

For submissions to the [The European Nucleotide Archive](#) you can follow the FAANG supported submission process. Your submission should be prepared following the guidance on the [FAANG wiki pages](#). This will guide you through:

- Downloading the empty Excel template to record your metadata
- Completing the template following the [instructions](#) and referring to the [latest experiment ruleset specification](#). The rules for each attribute define if it is mandatory or optional, what sort of data is expected (numeric, date, text, etc.), what units are permitted, and whether or not an ontology term is required.
- Visiting the [FAANG validation service](#) where you can validate that your template complies with the metadata specifications.
- Resolving any errors or warnings that it provides, referring to the [instructions](#) and referring to the [latest metadata rules specification](#) for advice.
- Converting your template into XML ready for submission using the [FAANG conversion tool](#)
- Follow the [upload and verification instructions for the ENA](#)

Links to the different submission systems can be found below.

- [ArrayExpress](#)
- [The Sequence read archive at NCBI](#)
- [GEO](#)

- [DDBJ](#)

Further guidance can be found on the [FAANG wiki](#) pages.