IMAGE metadata - experiment specification

This document describes the specification for all experiment metadata. The sample document is also available.

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.
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

These attributes should be present on every experiment record.

Mandatory:

- `Experiment type` (*ontology_id*) The type of experiment performed Allowed values are:
  - ATAC-seq
  - ChIP-seq
  - DNase-Hypersensitivity seq
  - Hi-C
  - bisulfite sequencing
  - microRNA profiling by high throughput sequencing
  - transcription profiling by high throughput sequencing
  - RNA-seq of coding RNA
  - RNA-seq of non coding RNA
  - whole genome sequencing
  - genotyping SNP

  - genotyping SSR
  - genotyping other markers
- `Experiment target` (*ontology_id*) What the experiment was trying to find Allowed values are:
  - open_chromatin_region
  - DNA methylation
  - input DNA
  - histone_modification
  - chromatin
  - RNA
  - deoxyribonucleic acid

Recommended:

- `Extraction protocol` (*uri_value*) Link to the protocol used to isolate the extract material
- `Library preparation location` (*text*) Location where library preperation was performed
- `Library preparation location longitude` (*number*) Longitude of the library prep location in decimal degrees
- `Library preparation location latitude` (*number*) Latitude of the library prep location in decimal degrees
- `Library preparation date` (*date*) Date on which the library was prepared
- `Sequencing location` (*text*) Location where sequencing was performed
- `Sequencing date` (*date*) Date on which sequencing was prepared

Optional:

- `Experimental protocol` (*uri_value*) Link to the description of the experiment protocol, an overview of the full experiment, that can refer to the order in which other protocols were performed and any intermediate steps not covered by the extraction or assay specific protocols
- `Sequencing location longitude` (*number*) Longitude of the location where sequencing was performed in decimal degrees
- `Sequencing location latitude` (*number*) Latitude of the location where sequencing was performed in decimal degrees

ATAC-seq

ATAC-seq experiments should have an `assay type` of ATAC-seq.

Recommended:

- `Transposase protocol` (*uri_value*) Link to 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.

Mandatory:

- `Library selection` (*ontology_id*) Allowed values are:
  - whole genome bisulfite sequencing
  - reduced representation bisulfite-seq
  - Tet-assisted bisulfite sequencing assay
  - MethylC-Capture sequencing assay
- `Bisulfite conversion protocol` (*uri_value*) Link to bisulfite conversion protocol
- `PCR product isolation protocol` (*uri_value*) Link to the protocol for isolating PCR products used for library generation
- `Bisulfite conversion percentage` (*number*)

Recommended:

- `Restriction enzyme` (*ec number*) Restriction enzyme used for Reduced representation bisulfite sequencing
- `Maximum fragment size selection range` (*number*)
- `Minimum fragment size selection range` (*number*)

ChIP-seq standard rules for both histone modifications and input DNA

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.

Mandatory:

- `ChIP protocol` (*uri_value*)
- `Library generation maximum fragment size range` (*number*)
- `Library generation minimum fragment size range` (*number*)

ChIP-seq for histone modifications

ChIP-seq histone modification experiments should have an `assay type` of ChIP-seq.

Mandatory:

- `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

DNase-Hypersensitivity seq

DNase-seq experiments should have an `assay type` of DNase-Hypersensitivity seq.

Mandatory:

- `DNase protocol` (*uri_value*)

Hi-C

Hi-C experiments should have an `assay type` of Hi-C.

Mandatory:

- `Restriction enzyme` (*ec number*) Restriction enzyme used
- `Restriction site` (*text*)

RNA-seq

RNA-seq experiemnts 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.

Mandatory:

- `RNA preparation 3' adapter ligation protocol` (*uri_value*)
- `RNA preparation 5' adapter ligation protocol` (*uri_value*)
- `Library generation pcr product isolation protocol` (*uri_value*)
- `Preparation reverse transcription protocol` (*uri_value*)
- `Library generation protocol` (*uri_value*)
- `Read strand` (*limited value*) For strand specific protocol, specify which mate pair maps to the transcribed strand or Report 'non-stranded' if the protocol is not strand specific. For single-ended sequencing: use 'sense' if the reads should be on the same strand as the transcript, 'antisense' if on opposite strand. For 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 Allowed values are:
  - not applicable
  - sense

  - antisense
  - mate 1 sense
  - mate 2 sense
  - non-stranded

Recommended:

- `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*) 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. For more information on RNA Integrity Numbers see Schroeder et al. (2006) http://bmcmolbiol.biomedcentral.com/articles/10.1186/1471-2199-7-3

Whole Genome Sequencing

Whole Genome Sequencing should have an `assay type` of whole genome sequencing assay.

Mandatory:

- `Library generation PCR product isolation protocol` (*uri_value*)
- `Library generation protocol` (*uri_value*)

Optional:

- `Library selection` (*limited value*) Allowed values are:
  - reduced representation
  - none

Mandatory:

- `Genotyping type` (*limited value*) Allowed values are:
  - SNP
  - SSR
  - Other markers
- `Genotyping protocol` (*uri_value*)