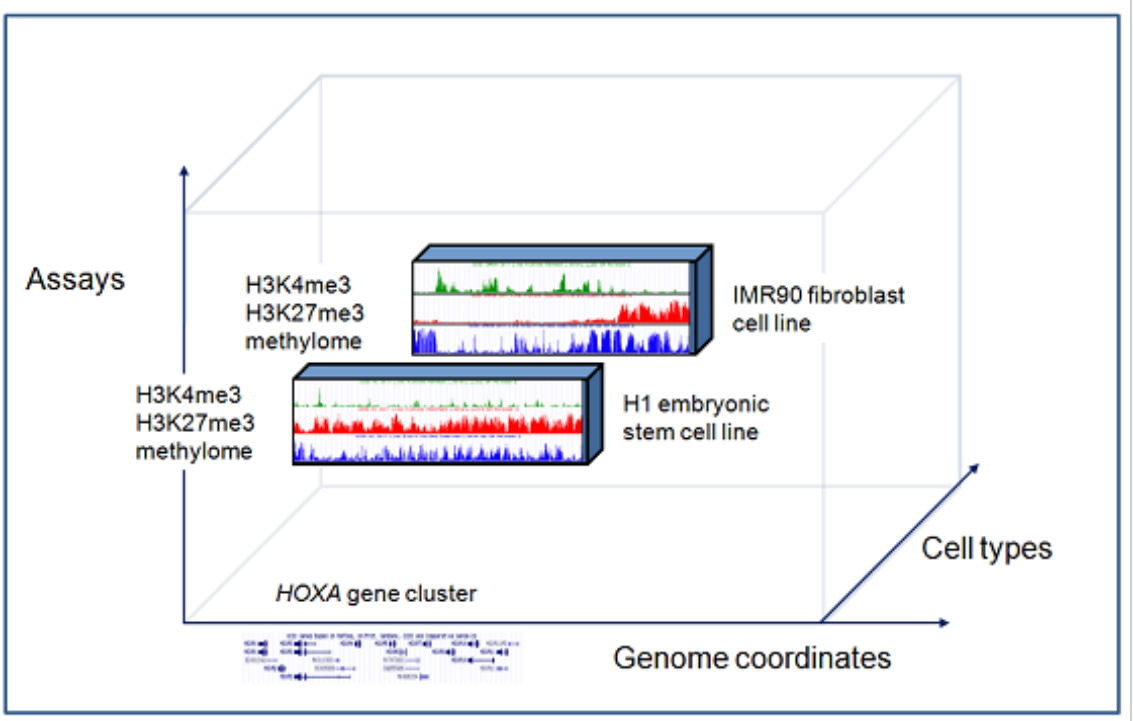
Data and Metadata Models Recommendations Version 1.2
Developed by the IHEC Metadata Standards Workgroup

1. Introduction

The data produced by IHEC is illustrated in **Figure 1**.

Figure 1. The space of epigenomic variation.



The two-dimensional plane in **Figure 1** corresponding to Assay and Cell Type combinations is displayed on the Epigenome Atlas portal page (www.epigenomeatlas.org) in the form of an interactive grid, as illustrated in **Figure 2**.

Each cell in the grid in **Figure 2** corresponds to the same Cell Type–Assay combination. Multiple sequencing Runs are combined by Library ID into the same technical replicate and, at the next higher level by Donor ID into a biological replicate.

Counts within each grid cell indicate the combined number of technical and biological replicates. By selecting individual cells in the grid, the user may navigate along the Genome coordinate axis by genomic element, gene, pathway and by other means using tools provided on the Epigenome Atlas portal page.

The grid in **Figure 2** is constructed automatically from the submitted metadata. The metadata may be viewed by clicking on column headers, row headers or individual cells within the grid.

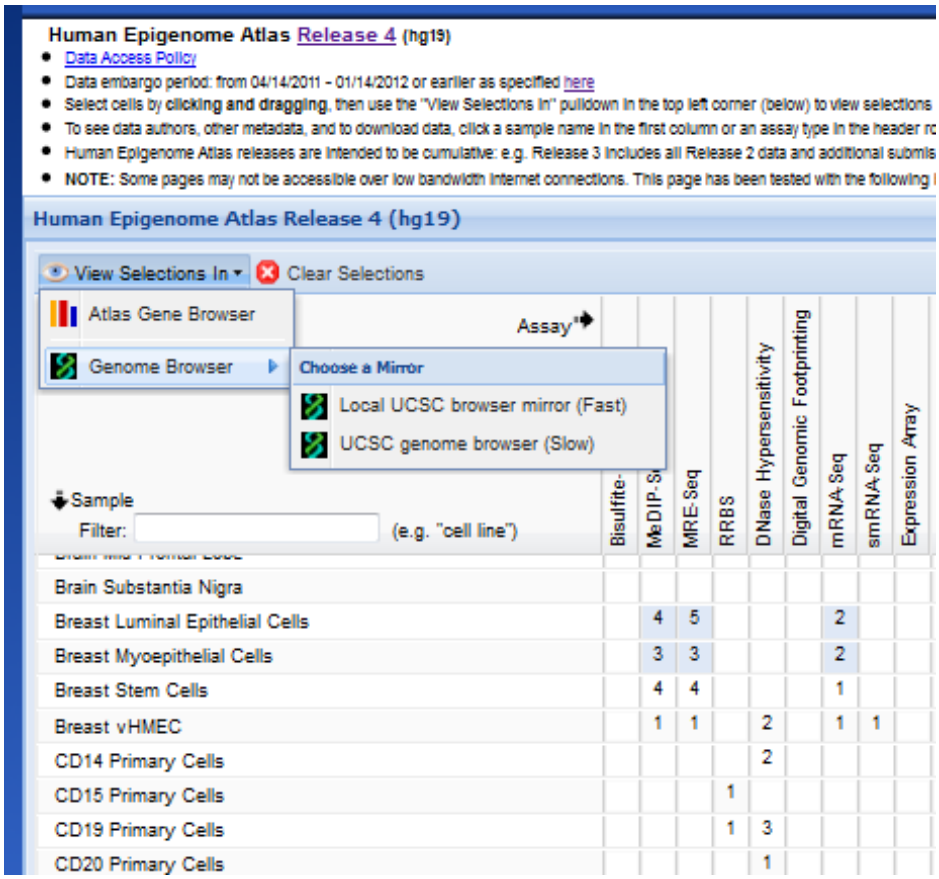


Figure 2. Human Epigenome Atlas portal (at www.epigenomeatlas.org).

2. Data and Metadata Model

The data / metadata model in **Fig. 3** is used to capture the data from mapping centers and generate data and metadata for submission to the GEO/SRA archives at the NCBI and organize the Human Epigenome Atlas portal.

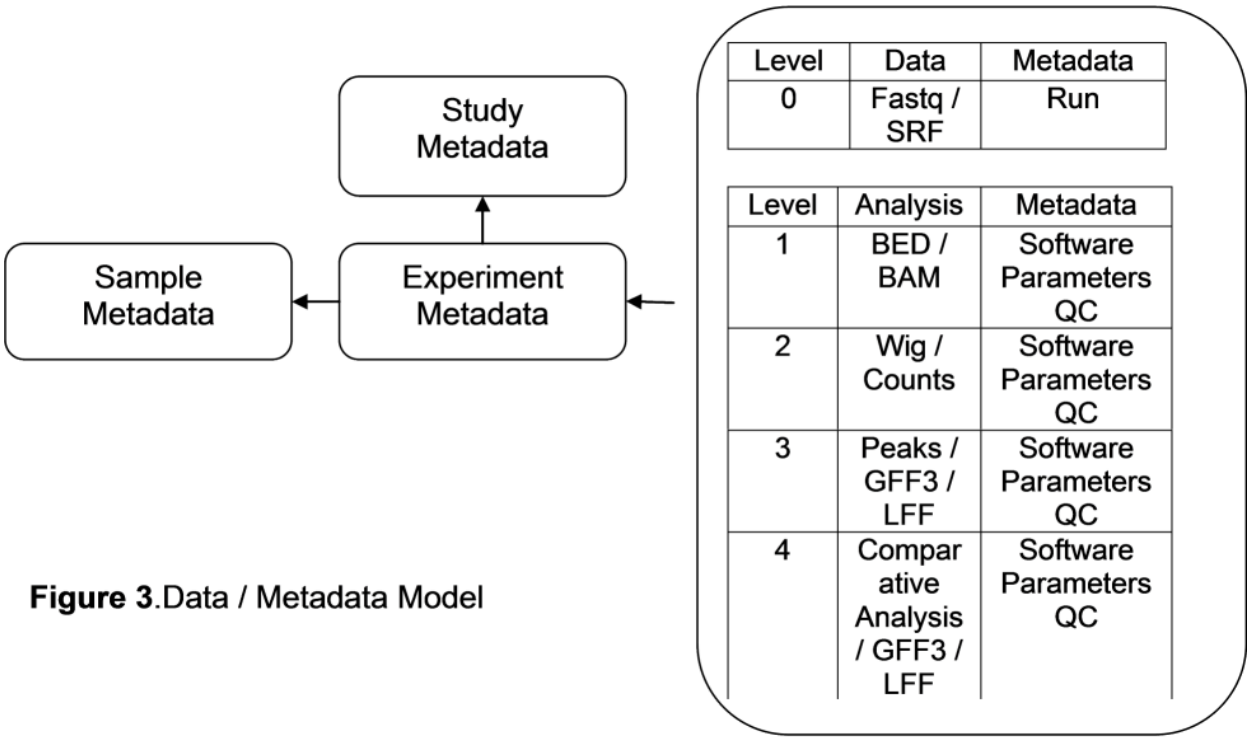


Figure 3.Data / Metadata Model

The arrows indicate references ("foreign key" references). Specifically, the arrows indicate that a sequencing experiment corresponds to a specific **Run,** an event which denotes production of a specific unit of data (such as the production of DNA sequencing reads from a single lane of an Illumina sequencer or hybridization of a sample to a specific arrays) in the context of a specific **Experiment** on a specific **Sample** in the context of a specific **Study**.

The data and metadata models are based on the SRA XML Schema Version 1.2. This schema was co-developed and is shared by the EBI and NCBI archives, facilitating exchange of data and international collaboration.

The SRA XML Schema Version 1.2 allows extensions with data fields that may be required to capture relevant data for specific applications such as epigenomic assays. The NIH Roadmap Epigenomics Initiative has extended the schema with specific data fields for **Study, Sample, Experiment, Run** and various **Analysis Levels** and types, as described in the following section.

3. Metadata Elements Extending SRA XML Schema 1.2

The core SRA XML elements are augmented by additional attributes defined for purposes of the NIH Roadmap Epigenomics as described in this section. The same attribute may be used multiple times in a single XML record. This may be most useful for supplying URIs to multiple ontologies or for supplying multiple references to a single ontology such as in the case of DISEASE_ONTOLOGY_URI.

Documentation for the core SRA XML elements is here:

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=doc

The SRA XML schemas are here:

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=xml_schemas

Level 0 Data

SAMPLES

Cell Line

MOLECULE - The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

SAMPLE_ONTOLOGY_URI - links to sample ontology information.

DISEASE_ONTOLOGY_URI - links to disease ontology information.

DISEASE: Free form field for more specific disease information

BIOMATERIAL_PROVIDER - The name of the company, laboratory or person that provided the biological material.

BIOMATERIAL_TYPE: Cell Line

LINE – The name of the cell line.

LINEAGE – The developmental lineage to which the cell line belongs.

DIFFERENTIATION_STAGE - The stage in cell differentiation to which the cell line belongs.

DIFFERENTIATION_METHOD – The protocol used to differentiation the cell line.

PASSAGE – The number of times the cell line has been re-plated and allowed to grow back to confluency or to some maximum density if using suspension cultures.

MEDIUM – The medium in which the cell line has been grown.

SEX: "Male", "Female" or "Unknown"

BATCH – The batch from which the cell line is derived. Primarily applicable to initial H1 cell line batches. NA if not applicable.

Primary Cell
MOLECULE - The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

SAMPLE_ONTOLOGY_URI - links to sample ontology information.

DISEASE_ONTOLOGY_URI - links to disease ontology information.

DISEASE: Free form field for more specific disease information

BIOMATERIAL_PROVIDER - The name of the company, laboratory or person that provided the biological material.

BIOMATERIAL_TYPE: Primary Cell

CELL_TYPE – The type of cell.

MARKERS – Markers used to isolate and identify the cell type.

DONOR_ID - An identifying designation for the donor that provided the primary cell.

DONOR_AGE - The age of the donor that provided the primary cell. NA if not available.

DONOR_HEALTH_STATUS - The health status of the donor that provided the primary cell. NA if not available.

DONOR_SEX: "Male", "Female" or "Unknown"

DONOR_ETHNICITY - The ethnicity of the donor that provided the primary cell. NA if not available.

PASSAGE_IF_EXPANDED – If the primary cell has been expanded, the number

of times the primary cell has been re-plated and allowed to grow back to confluency or to some maximum density if using suspension cultures. NA if no expansion.

Primary Cell Culture

MOLECULE- The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

SAMPLE_ONTOLOGY_URI - links to sample ontology information.

DISEASE_ONTOLOGY_URI - links to disease ontology information.

DISEASE: Free form field for more specific disease information.

BIOMATERIAL_PROVIDER - The name of the company, laboratory or person that provided the biological material.

BIOMATERIAL_TYPE: Primary Cell Culture

CELL_TYPE – The type of cell.

MARKERS – Markers used to isolate and identify the cell type.

CULTURE_CONDITIONS – The conditions under which the primary cell was cultured.

DONOR_ID - An identifying designation for the donor that provided the primary cell.

DONOR_AGE - The age of the donor that provided the primary cell. NA if not available.

DONOR_HEALTH_STATUS - The health status of the donor that provided the primary cell. NA if not available.

DONOR_SEX: "Male", "Female" or "Unknown"

DONOR_ETHNICITY - The ethnicity of the donor that provided the primary cell. NA if not available.

PASSAGE_IF_EXPANDED – If the primary cell culture has been expanded, the number of times the cell culture has been re-plated and allowed to grow back to confluency or to some maximum density if using suspension cultures. NA if no expansion.

Primary Tissue
MOLECULE - The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

SAMPLE_ONTOLOGY_URI - links to sample ontology information.

DISEASE_ONTOLOGY_URI - links to disease ontology information.

DISEASE: Free form field for more specific disease information.

BIOMATERIAL_PROVIDER - The name of the company, laboratory or person that provided the biological material.

BIOMATERIAL_TYPE: Primary Tissue

TISSUE_TYPE – The type of tissue.

TISSUE_DEPOT – Details about the anatomical location from which the primary tissue was collected.

COLLECTION_METHOD – The protocol for collecting the primary tissue.

DONOR_ID - An identifying designation for the donor that provided the primary tissue.

DONOR_AGE - The age of the donor that provided the primary tissue. NA if not available.

DONOR_HEALTH_STATUS - The health status of the donor that provided the primary tissue. NA if not available.

DONOR_SEX: "Male", "Female" or "Unknown"

DONOR_ETHNICITY - The ethnicity of the donor that provided the primary tissue. NA if not available.

EXPERIMENTS

Chromatin Accessibility
EXPERIMENT_TYPE: Chromatin Accessibility

EXPERIMENT_ONTOLOGY_URI links to experiment ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

DNASE_PROTOCOL – The protocol used for DNAse treatment.

WGBS (NOTE: this is a new name to be used instead of Bisulfite-Seq)
EXPERIMENT_TYPE: DNA Methylation

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_TYPE_OF_SONICATOR - The type of sonicator used for extraction.

EXTRACTION_PROTOCOL_SONICATION_CYCLES - The number of sonication cycles used for extraction.

DNA_PREPARATION_INITIAL_DNA_QNTY – The initial DNA quantity used in preparation.

DNA_PREPARATION_FRAGMENT_SIZE_RANGE – The DNA fragment size range used in preparation.

DNA_PREPARATION_ADAPTOR_SEQUENCE – The sequence of the adaptor used in preparation.

DNA_PREPARATION_ADAPTOR_LIGATION_PROTOCOL – The protocol used for adaptor ligation.

DNA_PREPARATION_POST-LIGATION_FRAGMENT_SIZE_SELECTION – The fragment size selection after adaptor ligation.

BISULFITE_CONVERSION_PROTOCOL – The bisulfite conversion protocol.

BISULFITE_CONVERSION_PERCENT – The bisulfite conversion percent and how it was determined.

LIBRARY_GENERATION_PCR_TEMPLATE_CONC – The PCR template concentration for library generation.

LIBRARY_GENERATION_PCR_POLYMERASE_TYPE – The PCR polymerase used for library generation

LIBRARY_GENERATION_PCR_THERMOCYCLING_PROGRAM – The thermocycling program used for library generation.

LIBRARY_GENERATION_PCR_NUMBER_CYCLES – The number of PCR cycles used for library generation.

LIBRARY_GENERATION_PCR_F_PRIMER_SEQUENCE – The sequence of the PCR forward primer used for library generation.

LIBRARY_GENERATION_PCR_R_PRIMER_SEQUENCE – The sequence of the PCR reverse primer used for library generation.

LIBRARY_GENERATION_PCR_PRIMER_CONC – The concentration of the PCR primers used for library generation.

LIBRARY_GENERATION_PCR_PRODUCT_ISOLATION_PROTOCOL – The protocol for isolating PCR products used for library generation.

MeDIP-Seq
EXPERIMENT_TYPE: DNA Methylation

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_TYPE_OF_SONICATOR - The type of sonicator used for extraction.

EXTRACTION_PROTOCOL_SONICATION_CYCLES - The number of sonication cycles used for extraction.

MeDIP_PROTOCOL – The MeDIP protocol used.

MeDIP_PROTOCOL_DNA_AMOUNT – The amount of DNA used in the MeDIP protocol.

MeDIP_PROTOCOL_BEAD_TYPE – The type of bead used in the MeDIP protocol.

MeDIP_PROTOCOL_BEAD_AMOUNT – The amount of beads used in the MeDIP protocol.

MeDIP_PROTOCOL_ANTIBODY_AMOUNT – The amount of antibody used in the MeDIP protocol.

MeDIP_ANTIBODY – The specific antibody used in the MeDIP protocol.

MeDIP_ANTIBODY_PROVIDER - The name of the company, laboratory or person that provided the antibody.

MeDIP_ANTIBODY_CATALOG – The catalog from which the antibody was purchased.

MeDIP_ANTIBODY_LOT – The lot identifier of the antibody.

MRE-Seq
EXPERIMENT_TYPE: DNA Methylation

EXPERIMENT_ONTOLOGY_URI elements that contain links to experiment ontology information.

MRE_PROTOCOL – The MRE protocol.

MRE_PROTOCOL_CHROMATIN_AMOUNT – The amount of chromatin used in the MRE protocol.

MRE_PROTOCOL_RESTRICTION_ENZYME – The restriction enzyme(s) used in the MRE protocol.

MRE_PROTOCOL_SIZE_FRACTION – The size of the fragments selected in the MRE protocol.

Chip-Seq Input
EXPERIMENT_TYPE: ChIP-Seq Input

EXPERIMENT_ONTOLOGY_URIS - container element for EXPERIMENT_ONTOLOGY_URI elements that contain links to sample ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_TYPE_OF_SONICATOR - The type of sonicator used for extraction.

EXTRACTION_PROTOCOL_SONICATION_CYCLES - The number of sonication cycles used for extraction.

CHIP_PROTOCOL: Input

CHIP_PROTOCOL_CHROMATIN_AMOUNT– The amount of chromatin used in the ChIP protocol.

Chip-Seq

EXPERIMENT_TYPE: 'Histone H3K4me1','Histone H3K4me3','Histone H3K9me3','Histone H3K9ac','Histone H3K27me3', 'Histone H3K36me3', etc.

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_TYPE_OF_SONICATOR - The type of sonicator used for extraction.

EXTRACTION_PROTOCOL_SONICATION_CYCLES - The number of sonication cycles used for extraction.

CHIP_PROTOCOL – The ChIP protocol used.

CHIP_PROTOCOL_CHROMATIN_AMOUNT - The amount of chromatin used in the ChIP protocol.

CHIP_PROTOCOL_BEAD_TYPE - The type of bead used in the ChIP protocol.

CHIP_PROTOCOL_BEAD_AMOUNT - The amount of beads used in the ChIP protocol.

CHIP_PROTOCOL_ANTIBODY_AMOUNT – The amount of antibody used in the ChIP protocol.

CHIP_ANTIBODY - The specific antibody used in the ChIP protocol.

CHIP_ANTIBODY_PROVIDER - The name of the company, laboratory or person that provided the antibody.

CHIP_ANTIBODY_CATALOG – The catalog from which the antibody was purchased.

CHIP_ANTIBODY_LOT – The lot identifier of the antibody.

CHIP_PROTOCOL_CROSSLINK_TIME - The timespan in which the chromatin is crosslinked
LIBRARY_GENERATION_FRAGMENT_SIZE_RANGE – The fragment size range of the preparation.

<u>mRNA-seq</u>
EXPERIMENT_TYPE: mRNA-Seq

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_MRNA_ENRICHMENT – The mRNA enrichment method used in the extraction protocol.

EXTRACTION_PROTOCOL_FRAGMENTATION – The fragmentation method used in the extraction protocol.

MRNA_PREPARATION_FRAGMENT_SIZE_RANGE – The mRNA fragment size range of the preparation.

RNA_PREPARATION_5'_RNA_ADAPTER_SEQUENCE – The sequence of the 5’ RNA adapter used in preparation.

RNA_PREPARATION_3'_RNA_ADAPTER_SEQUENCE - The sequence of the 3’ RNA adapter used in preparation.

RNA_PREPARATION_REVERSE_TRANSCRIPTION_PRIMER_SEQUENCE – The sequence of the primer for reverse transcription used in preparation.

RNA_PREPARATION_5'_DEPHOSPHORYLATION – The protocol for 5’ dephosphorylation used in preparation.

RNA_PREPARATION_5'_PHOSPHORYLATION – The protocol for 5’ phosphorylation used in preparation.

RNA_PREPARATION_3'_RNA ADAPTER_LIGATION_PROTOCOL – The protocol for 3’ adapter ligation used in preparation.

RNA_PREPARATION_5'_RNA_ADAPTER_LIGATION_PROTOCOL - The protocol for 5’ adapter ligation used in preparation.

LIBRARY_GENERATION_PCR_TEMPLATE_CONC – The PCR template concentration for library generation.

LIBRARY_GENERATION_PCR_POLYMERASE_TYPE – The PCR polymerase used for library generation

LIBRARY_GENERATION_PCR_THERMOCYCLING_PROGRAM – The thermocycling program used for library generation.

LIBRARY_GENERATION_PCR_NUMBER_CYCLES – The number of PCR cycles used for library generation.

LIBRARY_GENERATION_PCR_F_PRIMER_SEQUENCE – The sequence of the PCR forward primer used for library generation.

LIBRARY_GENERATION_PCR_R_PRIMER_SEQUENCE – The sequence of the PCR reverse primer used for library generation.

LIBRARY_GENERATION_PCR_PRIMER_CONC – The concentration of the PCR primers used for library generation.

LIBRARY_GENERATION_PCR_PRODUCT_ISOLATION_PROTOCOL – The protocol for isolating PCR products used for library generation.

TEMPLATE_TYPE - mRNA or cDNA - The type of template.

AMPLIFIED - True or False - Is the sample amplified?

PREPARATION_INITIAL_MRNA_QNTY -The initial mRNA quantity used in preparation.

PREPARATION_REVERSE_TRANSCRIPTION_PROTOCOL - The protocol for reverse transcription used in preparation.

PREPARATION_PCR_NUMBER_CYCLES - The number of PCR cycles used to amplify.

LIBRARY_GENERATION_PROTOCOL - The protocol used to generate the library.

LIBRARY_GENERATION_FRAGMENTATION - The fragmentation method used in the library protocol.

LIBRARY_GENERATION_FRAGMENT_SIZE_RANGE – The fragment size range of the preparation.

LIBRARY_GENERATION_3'_ADAPTER_SEQUENCE – The sequence of the 3' adapter used for library generation.

LIBRARY_GENERATION_5'_ADAPTER_SEQUENCE – The sequence of the 5' adapter used for library generation.

<u>smRNA-Seq</u>
EXPERIMENT_TYPE:smRNA-Seq

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

EXTRACTION_PROTOCOL - The protocol used to isolate the extract material.

EXTRACTION_PROTOCOL_SMRNA_ENRICHMENT - The smRNA enrichment method used in the extraction protocol.

SMRNA_PREPARATION_INITIAL_SMRNA_QNTY - The initial smRNA quantity

used in preparation.

RNA_PREPARATION_5'_RNA_ADAPTER_SEQUENCE – The sequence of the 5’ RNA adapter used in preparation.

RNA_PREPARATION_3'_RNA_ADAPTER_SEQUENCE - The sequence of the 3’ RNA adapter used in preparation.

RNA_PREPARATION_REVERSE_TRANSCRIPTION_PRIMER_SEQUENCE – The sequence of the primer for reverse transcription used in preparation.

RNA_PREPARATION_3'_RNA ADAPTER_LIGATION_PROTOCOL – The protocol for 3’ adapter ligation used in preparation.

RNA_PREPARATION_5'_RNA_ADAPTER_LIGATION_PROTOCOL - The protocol for 5’ adapter ligation used in preparation.

RNA_PREPARATION_REVERSE_TRANSCRIPTION_PROTOCOL - The protocol for reverse transcription used in preparation.

LIBRARY_GENERATION_PCR_TEMPLATE_CONC – The PCR template concentration for library generation.

LIBRARY_GENERATION_PCR_POLYMERASE_TYPE – The PCR polymerase used for library generation

LIBRARY_GENERATION_PCR_THERMOCYCLING_PROGRAM – The thermocycling program used for library generation.

LIBRARY_GENERATION_PCR_NUMBER_CYCLES – The number of PCR cycles used for library generation.

LIBRARY_GENERATION_PCR_F_PRIMER_SEQUENCE – The sequence of the PCR forward primer used for library generation.

LIBRARY_GENERATION_PCR_R_PRIMER_SEQUENCE – The sequence of the PCR reverse primer used for library generation.

LIBRARY_GENERATION_PCR_PRIMER_CONC – The concentration of the PCR primers used for library generation.

LIBRARY_GENERATION_PCR_PRODUCT_ISOLATION_PROTOCOL – The protocol for isolating PCR products used for library generation.

Level 1 Data (SRA: ANALYSIS_TYPE -REFERENCE_ALIGNMENT)

DATA_ANALYSIS_LEVEL - 1

EXPERIMENT_TYPE - The type of experiment (Chromatin Accessibility, Bisulfite-Seq, MeDIP-Seq, MRE-Seq, ChIP-Seq, mRNA-Seq, smRNA-Seq).

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

GENOME_ASSEMBLY – The genome assembly to which the reads are mapped.

SOFTWARE – The name of the software used for mapping.

SOFTWARE_VERSION – The version of the software used for mapping.

SOFTWARE_COMMAND_LINE - The command line used to run the software.

ANALYSIS_PROTOCOL - Description of how the analysis was performed.

MAXIMUM_ALIGNMENT_LENGTH – The maximum read alignment length supported by the software. If the software aligns the entire read use “Read Length”.

MISMATCHES_ALLOWED – The number of mismatches allowed in an alignment.

ALIGNMENTS_ALLOWED – The number of locations to which a read is allowed to align.

TREATMENT_OF_MULTIPLE_ALIGNMENTS – How reads aligning to multiple locations are treated.

TREATMENT_OF_IDENTICAL_ALIGNMENTS_OF_MULTIPLE_READS – How multiple reads aligning to the same location are treated. This applies to clonal duplicate removal.

ALIGNMENT_POSTPROCESSING – Any postprocessing applied to the alignments.

NUMBER_OF_MAPPED_READS – The number of mapped reads.

Quality Control – Quality control related analysis attributes. These are dependent on the type of experiment.

Level 2 Data (SRA: ANALYSIS_TYPE - ABUNDANCE_MEASUREMENT)

DATA_ANALYSIS_LEVEL: 2

EXPERIMENT_TYPE - The type of experiment (Chromatin Accessibility, Bisulfite-Seq, MeDIP-Seq, MRE-Seq, ChIP-Seq, mRNA-Seq, smRNA-Seq).

EXPERIMENT_ONTOLOGY_URI - links to experiment ontology information.

GENOME_ASSEMBLY – The genome assembly to which the reads are mapped.

SOFTWARE – The name of the software used for determining signal (read density).

SOFTWARE_VERSION – The version of the software used for determining signal (read density).

SOFTWARE_COMMAND_LINE - The command line used to run the software.

ANALYSIS_PROTOCOL - Description of how the analysis was performed.

READ_EXTENSION – If read mappings are extended before determining signal (read density), the length to which the read is extended in bp. NA if not applicable.

GENOMIC_WINDOW – The bp size of the window in which the signal (read density) is calculated.

TREATMENT_OF_REGIONS_PRONE_TO_MULTIPLE_ALIGNMENTS – Any treatment of regions, such as repeats, which are prone to multiple alignments. NA if not applicable.

NUMBER_OF_MAPPED_READS – The number of mapped reads.

Quality Control – Quality control related analysis attributes. These are dependent on the type of experiment.

4. Accepted Ontologies

Sample Ontologies

- Cell Lines - Experimental Factor Ontology (EFO)
- Primary Cells - Cell Ontology (CL)
- Primary Tissue - Uberon

Disease Ontologies

- Disease - NCI Metathesaurus

Experiment Ontologies

- Assays and Platforms - Ontology for Biomedical Investigations (OBI)