IHEC TrackHub - metadata requirements

Introduction

Many groups involved in IHEC intend to produce track hubs to allow their data to be viewed in genome browsers (UCSC & Ensembl). The TrackHub format includes a metadata field that allows each track to be described as a set of key/value pairs. We propose that IHEC members should use the metadata fields in a standardardized way to improve interoperability between groups.

Scope

This document will set out the metadata fields required for each track type. These fields will be a subset of those required by the IHEC metadata standards group. The metadata standards group is best placed to offer guidance on appropriate values for each field.

Is is expected that the sub group field will contain much of the same information as the metadata field. However, the sub groups are there to help a user select data and do not need to precisely mirror the meta data field. For example, it may make sense to merge the donor ID and line fields into a single sub group, ditto sex and donor sex.
Caveats
We have not tried to use these attributes yet. We need to establish that the UCSC browser can handle a large number of attributes.

The UCSC browser attempts to match key/value pairs in the metadata with those used by ENCODE. This is not helpful for non-ENCODE data.

The UCSC public hub guidelines note that support for metadata may be replaced by another system at some point in the future.

Ensembl does not use the metadata at the moment. This is something Blueprint is working with Ensembl to improve.

Hub Structure

For IHEC hubs to have a consistent look and feel here are some recommendations about the structure of the hub you produce.

Its aim is to provide a consistent grid representation for IHEC core assays.

The string delimiter should always be an underscore _. UCSC automatically replace underscores with spaces and this makes labels easier for human readability.

Groups should aim to present all their analyses in a single table so users can select from all the experimental data at once. This can be achieved using the

Composite tracks configuration to group similar tracks together in trackhub. Trackhub provides two different grouping style using "subgroups" and "views". More details on this grouping is available on the UCSC trackhub spec. http://genome.ucsc.edu/goldenPath/help/trackDb/trackDbHub.html#compositeTrack.

A caveat to this is if you have many different experiment types above and beyond the core IHEC experiments you may wish to have a table for the core experiments and another table for other experiments.

For tracks to be presented together, they need to be under a single composite parent. Tracks carrying similar types of information (eg, genomic regions in bed or bigBed or signal values in bigWig) will be grouped together using "view".

Its possible to specify the UI design for subtracks selection using subgroups settings. Dimension X (dimX) and Y (dimY) are used for defining a one or two dimensional array array of checkboxes (works well both in UCSC and Ensembl browser). Additional dimensions (abc) can be declared which provide added functionality of subtracks selection in UCSC. Currently Ensembl browser doesn't support these extra dimensions but its working on enabling this functionality in future releases.

Subgroups for subtrack selection:

Experiment (dimX) (e.g, Bisulphite-Seq)
Sample description (dimY) (e.g, cell type, cell line or disease info)
Analysis type (dimA) (sub category of experiment, e.g, hypo methylation)

Additional dimensions for different subgroups can be added.

Example Blueprint trackhub:

track bp
compositeTrack on
shortLabel Blueprint
longLabel Blueprint
subGroup1 experiment Experiment DNase=DNase
subGroup2 analysis_type Analysis_type HOTSPOT_peak=HOTSPOT_peak wiggler=wiggler
subGroup3 sample_description Sample_description DG-75=DG-75 U-266=U-266
subGroup5 view View Region=Region Signal=Signal
subGroup6 analysis_group Analysis_group EMBL-EBI=EMBL-EBI
dimensions dimX=experiment dimY=sample_description dimA=analysis_type dimB=analysis_group
filterComposite dimA dimB
dragAndDrop subTracks

sortOrder analysis_group=+ view=+ sample_description=+ analysis_type=+
experiment=+
priority 4
type bed 3
visibility full‘

track regions
parent bp
shortLabel Blueprint Regions
view Region
type bigBed
visibility dense

```
track bpDnaseRegionsBP_DG-75_d01DNaseHOTSPOT_peakEMBL-EBI
bigDataUrl http://.../BP_DG-75_d01.DNase.hotspot_v3.20130819.bb
parent regions on
type bigBed 6 .
shortLabel DG-75.DNase.DG-75
longLabel DG-75 DNase DG-75 peaks from NCMLS
color 8,104,172
subGroups experiment=DNase sample_description=DG-75 analysis_type=HOTSPOT_peak
      view=Region analysis_group=EMBL-EBI
metadata ...

track bpDnaseRegionsBP_U-266_d01DNaseHOTSPOT_peakEMBL-EBI
bigDataUrl http://.../BP_U-266_d01.DNase.hotspot_v3.20130819.bb
parent regions off
type bigBed 6 .
shortLabel U-266.DNase.U-266
longLabel U-266 DNase U-266 peaks from NCMLS
color 8,104,172
subGroups experiment=DNase sample_description=U-266 analysis_type=HOTSPOT_peak
      donor_id=U-266 view=Region analysis_group=EMBL-EBI
metadata ...
```

track signal
parent bp
shortLabel Blueprint Signal
view Signal
type bigWig
autoscale off
maxHeightPixels 64:32:16
visibility pack

```
track bpDnaseSignalBP_DG-75_d01DNasewigglerEMBL-EBIwiggler
bigDataUrl http://.../BP_DG-75_d01.DNase.wiggler.20130819.bw
parent signal off
type bigWig 0 100
```

```
shortLabel DG-75.DNase.DG-75
longLabel DG-75 DNase DG-75 signal from NCMLS
color 8,104,172
subGroups experiment=DNase sample_description=DG-75 analysis_type=wiggler
      view=Signal analysis_group=EMBL-EBI
metadata ...

track bpDnaseSignalBP_U-266_d01DNasewigglerEMBL-EBIwiggler
bigDataUrl http://.../BP_U-266_d01.DNase.wiggler.20130819.bw
parent signal off
type bigWig 0 100
shortLabel U-266.DNase.U-266
longLabel U-266 DNase U-266 signal from NCMLS
color 8,104,172
subGroups experiment=DNase sample_description=U-266 analysis_type=wiggler
      view=Signal analysis_group=EMBL-EBI
metadata ...
```

Required attributes

The aim is to include attributes that describe what the experiment represents, rather than how it was produced. Protocol information is too long to work well in this format. Full metadata, as per the IHEC metadata guidelines, should be present in the primary data archive (e.g. SRA). Many of these attributes should match what is also in the primary archive meta data.

Common attributes for all samples

- MOLECULE
- DISEASE
- BIOMATERIAL_TYPE
- SAMPLE_ONTOLOGY_URI
- DISEASE_ONTOLOGY_URI
- SAMPLE_ID - e.g. the archive ID for the sample

Cell line attributes

- LINE
- LINEAGE
- DIFFERENTIATION_STAGE
- MEDIUM
- SEX

Common attributes for primary cell, primary cell culture, primary tissue

- DONOR_ID
- DONOR_AGE
- DONOR_HEALTH_STATUS
- DONOR_SEX
- DONOR_ETHNICITY

Primary cell attributes

- CELL_TYPE

Primary cell culture attributes

- CELL_TYPE
- CULTURE_CONDITIONS

Primary tissue attributes

- TISSUE_TYPE
- TISSUE_DEPOT

Experiment Attributes

- EXPERIMENT_TYPE
- LIBRARY STRATEGY
- EXPERIMENT ID (e.g. SRX006237)
- REFERENCE REGISTRY ID (as assigned by the EpiRR registry service, when it is up and running)
- ANALYSIS_GROUP - e.g the centre who performed the analysis, which produced the file

Alignment

Analysis level 1

- ALIGNMENT_SOFTWARE (relabelled from SOFTWARE)
- ALIGNMENT_SOFTWARE_VERSION (relabelled from SOFTWARE_VERSION)

Analysis level 2

- ANALYSIS_SOFTWARE (relabelled from SOFTWARE)
- ANALYSIS_SOFTWARE_VERSION (relabelled from SOFTWARE_VERSION)

Track-specific metadata

- TRACK_TYPE (Defines the type of view that this track represents. Could be restricted with controlled vocabulary terms, in the following list: "signal", "peaks", "contigs", "profile")

Examples

This is an example of a single track with the metadata suitable for a primary cell sample:

track bpDnaseRegionsC0010K46DNaseEBI
bigDataUrl http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/Peripheral_blood/C0010K/Monocytes/DNase-Hypersensitivity/C0010K46.DNase.hotspot_v3.20130415.bb
parent bpDnaseRegions on
type bigBed 6 .
shortLabel C0010K.DNase.Mono
longLabel C0010K DNase Monocytes peaks (EBI)
color 8,104,172
subGroups experiment_type=DNase cell_type=CD14-positive,_CD16-negative_classical_monocyte donor_id=C0010K view=Region analysis_group=EBI
metadata MOLECULE=genomic_DNA DISEASE=None BIOMATERIAL_TYPE=Primary_cells
SAMPLE_ID=EGAN00001070025 DONOR_ID=C0010K46 DONOR_AGE=60-65 DONOR_HEALTH_STATUS=Healthy DONOR_SEX=Female
DONOR_ETHNICITY=Northern_European CELL_TYPE=CD14-positive,_CD16-negative_classical_monocyte TISSUE_TYPE=Venous_blood
LIBRARY_STRATEGY=DNAse-Seq EXPERIMENT_TYPE=Chromatin_accessibility
Experiment_ID=EGAX00001084791 ALIGNMENT_SOFTWARE=BWA
ALIGNMENT_SOFTWARE_VERSION=0.5.9 ANALYSIS_SOFTWARE=Hotspot
ANALYSIS_SOFTWARE_VERSION=v3 ANALYSIS_GROUP=EMBL-EBI

This is an example of a metadata suitable for cell line sample:

metadata LINE=BL-2 DIFFERENTIATION_STAGE=B_cell
MEDIUM=RPMI_1640_+*10%FBS*+_1%_Glutamine SEX=Male MOLECULE=genomic_DNA
DISEASE=Sporadic_Burkitt_lymphoma DISEASE_ONTOLOGY_URI=http://ncimeta.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20MetaThesaurus&code=C1336077 BIOMATERIAL_TYPE=Cell_line
SAMPLE_ONTOLOGY_URI=http://www.ebi.ac.uk/efo/EFO_0001639
SAMPLE_ID=ERS333897 LIBRARY_STRATEGY=DNAse-Seq EXPERIMENT_TYPE=H3K27ac Experiment_ID=EGAX00001084792 ALIGNMENT_SOFTWARE=BWA ALIGNMENT_SOFTWARE_VERSION=0.5.9
ANALYSIS_SOFTWARE=Hotspot ANALYSIS_SOFTWARE_VERSION=v3
ANALYSIS_GROUP=EMBL-EBI

This is an example of a metadata suitable for primary tissue sample:

metadata MOLECULE=genomic_DNA DISEASE=None BIOMATERIAL_TYPE=Primary_tissue SAMPLE_ID=EGAN00001070027 DONOR_ID=C0010K46 DONOR_AGE=60-65 DONOR_HEALTH_STATUS=Healthy DONOR_SEX=Female DONOR_ETHNICITY=Northern_European TISSUE_TYPE=Venous_blood TISSUE_DEPOT=median_cubital_vein LIBRARY_STRATEGY=DNAse-Seq EXPERIMENT_TYPE=Chromatin_accessibility Experiment_ID=EGAX00001084792 ALIGNMENT_SOFTWARE=BWA ALIGNMENT_SOFTWARE_VERSION=0.5.9 ANALYSIS_SOFTWARE=Hotspot ANALYSIS_SOFTWARE_VERSION=v3 ANALYSIS_GROUP=EMBL-EBI SAMPLE_ONTOLOGY_URI=http://purl.obolibrary.org/obo/CL_0000775

This is an example of a metadata suitable for a primary cell culture sample:

metadata MOLECULE=genomic_DNA DISEASE=None BIOMATERIAL_TYPE=Primary_Cell_Culture CELL_TYPE=macrophage SAMPLE_ID=EGAN00001070028 DONOR_ID=C0010K46 DONOR_AGE=60-65 DONOR_HEALTH_STATUS=Healthy DONOR_SEX=Female DONOR_ETHNICITY=Northern_European CULTURE_CONDITIONS=http://www.blueprint-epigenome.eu/UserFiles/file/Protocols/UCAM_BluePrint_Macrophage.pdf LIBRARY_STRATEGY=DNAse-Seq EXPERIMENT_TYPE=Chromatin_accessibility Experiment_ID=EGAX00001084792 ALIGNMENT_SOFTWARE=BWA ALIGNMENT_SOFTWARE_VERSION=0.5.9 ANALYSIS_SOFTWARE=Hotspot ANALYSIS_SOFTWARE_VERSION=v3 ANALYSIS_GROUP=EMBL-EBI SAMPLE_ONTOLOGY_URI=http://purl.obolibrary.org/obo/CL_0000235