BLUEPRINT proposed data submission schemas

Draft v0.2.3.4-a6979008015043a94ca2fa20d6178ec46524e858

(model SHA1 9854a4a492849d0ebfd9f5578cc8d0e417f7509d)

(external controlled vocabulary SHA1 5cc6e3c2b41962a47a9563a3691e4e6b940d9797)

José María Fernández

March 26, 2013

Contents

1 Data Submission 1
- 1.1 Overview of Data Submission Process . . . 1
- 1.2 Preparing Sample Tracking Data and Analyzed Contents for their submission . . . 1
  - 1.2.1 File Naming Conventions . . . 2
  - 1.2.2 Tabular File Structure . . . 3
- 1.3 File Submission Procedure . . . 4

2 DCC Submission Tabular Formats 5
- 2.1 Sample Tracking Submission File Specifications . . . 5
  - 2.1.1 Donor Data File . . . 6
  - 2.1.2 Specimen Data File . . . 6
  - 2.1.3 Analyzed Sample Data File . . . 8
- 2.2 Gene Expression . . . 9
  - 2.2.1 Expression - Metadata File . . . 9
  - 2.2.2 Expression - Gene File . . . 11
- 2.3 Exon Junction . . . 12
  - 2.3.1 Exon Junction - Metadata File . . . 12
  - 2.3.2 Exon Junction - Primary Analysis File . . . 14
- 2.4 Protein-DNA interactions . . . 16
  - 2.4.1 Protein-DNA interaction - Metadata File . . . 16
  - 2.4.2 Protein-DNA interaction - Primary Analysis File . . . 18
  - 2.4.3 Protein-DNA interaction - Secondary Analysis File . . . 19
- 2.5 Regulatory Regions . . . 19
  - 2.5.1 Regulatory regions - Metadata File . . . 19
  - 2.5.2 Regulatory regions - Primary Analysis File . . . 21
  - 2.5.3 Regulatory regions - Secondary Analysis File . . . 22
- 2.6 DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...) . . . 23
  - 2.6.1 DNA *-lation - Metadata File . . . 23
  - 2.6.2 DNA *-lation - Primary Analysis File . . . 24
  - 2.6.3 DNA *-lation - Secondary Analysis File . . . 26

A Controlled Vocabularies 27
- A.1 Ensembl Genes . . . 27
- A.2 Ensembl Transcripts . . . 27
- A.3 Cell Ontology . . . 27
- A.4 ISO 3166-1 and ISO 3166-2 . . . 27
- A.5 Institution ID . . . 27
- A.6 Value Codes for Platform or Validation Platform . . . 29
- A.7 Chromosome Names for Reference Genome GRCh37 . . . 31
- A.8 Value Codes for Reference Genome Assembly Version . . . 34
- A.9 Value Codes for Raw Data Repository . . . 35

List of Figures

1.1 Overview of BLUEPRINT data model 2

2.1 Sample Tracking Data Sub-Schema 5
2.2 Gene Expression Sub-Schema 9
2.3 Junction Read Count explanation 12
2.4 Exon Junction Sub-Schema 13
2.5 Protein-DNA interactions Sub-Schema 16
2.6 Regulatory Regions Sub-Schema 20
2.7 DNA Methylation , Hydroxy-Methylation, Formylation, etc... Sub-Schema 23

List of Tables

2.1.1 Donor Data File . . . 6
2.1.2 Specimen Data File . . . 6
2.1.3 Analyzed Sample Data File . . . 8
2.2.1 Expression - Metadata File . . . 10
2.2.2 Expression - Gene File . . . 11
2.3.1 Exon Junction - Metadata File . . . 13
2.3.2 Exon Junction - Primary Analysis File . . . 14
2.4.1 Protein-DNA interaction - Metadata File . . . 17
2.4.2 Protein-DNA interaction - Primary Analysis File . . . 18
2.4.3 Protein-DNA interaction - Secondary Analysis File . . . 19
2.5.1 Regulatory regions - Metadata File . . . 20
2.5.2 Regulatory regions - Primary Analysis File . . . 21
2.5.3 Regulatory regions - Secondary Analysis File . . . 22
2.6.1 DNA *-lation - Metadata File . . . 23
2.6.2 DNA *-lation - Primary Analysis File . . . 24
2.6.3 DNA *-lation - Secondary Analysis File . . . 26

A.4.1 ISO 3166-1 and ISO 3166-2 aliases . . . 27
A.5.1 Institution ID . . . 28
A.6.1 Value Codes for Platform or Validation Platform . . . 29
A.7.1 Chromosome Names for Reference Genome GRCh37 . . . 31
A.8.1 Value Codes for Reference Genome Assembly Version . . . 34
A.9.1 Value Codes for Raw Data Repository . . . 35

Chapter 1

Data Submission

1.1 Overview of Data Submission Process

There are four major steps in the data submission process:

1. Submit raw sequence data to the European Genome-phenome Archive
2. Prepare the BLUEPRINT submission files according to DCC data format specifications
3. Verify conformity of the submission files
4. Submit files to the DCC Secure FTP server

All submitted data must be based on **Human reference genome assembly GRCh37** and GENCODE 15 (which uses **EnsEMBL gene set version 70**)

When submitting experimental data, please make sure you've already deposited your raw data to the appropriate public data repositories (eg: sequencing reads to EBI EGA) and then populate in your submission files the data elements **raw_data_repository** and **raw_data_accession** with the correct repository and accession number respectively.

1.2 Preparing Sample Tracking Data and Analyzed Contents for their submission

Submitted experimental data files must be from any one of these categories:

- Sample Tracking
- Gene Expression
- Exon Junctions
- DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...)
- Protein-DNA interactions
- Regulatory regions

BLUEPRINT DCC is hosting both sample tracking data and analyzed contents. Contents must be sent following the textual tabular formats defined below. Files with those contents must also follow the BLUEPRINT DCC file naming convention.

Each submitter must have a unique signing key, provided by DACO and DCC. Each file in a submitted archive must be accompanied by its SHA1 **uncompressed** content digest file, digitally signed with the submitter's signing key.

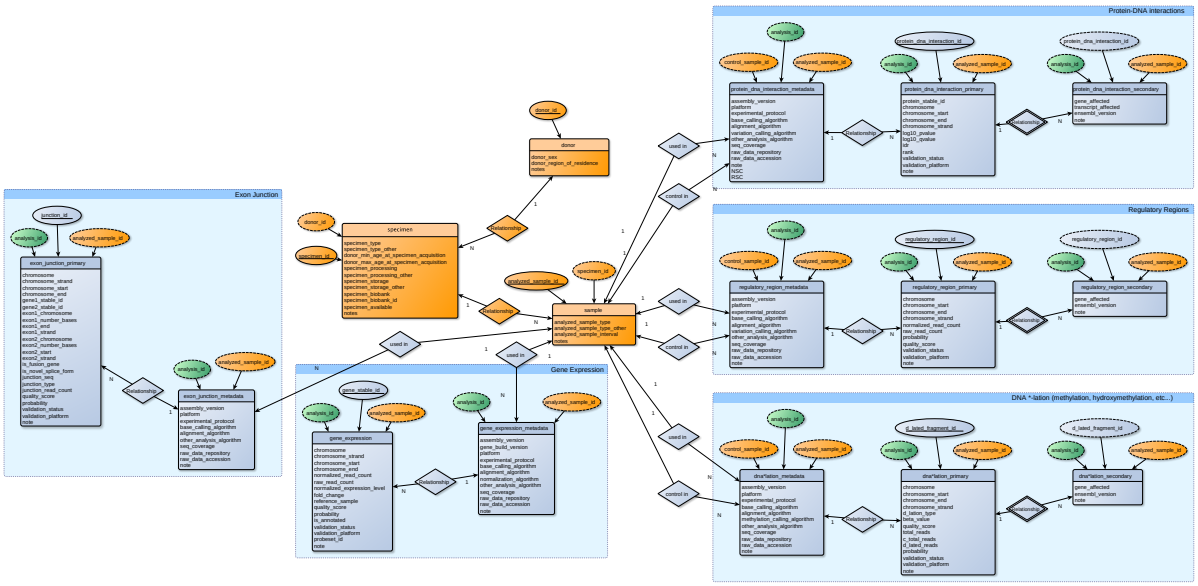

Figure 1.1: Overview of BLUEPRINT data model

Signed digest generation and verification using OpenSSL

```
# Signed digest of uncompressed contents, will be dlat-p--001-20120920--mycode.txt.sha1
openssl dgst -sha1 -sign subKey.pem -out dlat-p--001-20120920--mycode.txt.sha1 \
        dlat-p--001-20120920--mycode.txt

# Signed digest of already compressed contents
bunzip2 -c dlat-p--001-20120920--mycode.txt.bz2 | openssl dgst -sha1 -sign subKey.pem \
        -out dlat-p--001-20120920--mycode.txt.sha1

# Verification of uncompressed contents using
# signed digest dlat-p--001-20120920--mycode.txt.sha1
openssl dgst -sha1 -verify subKey.pem.pub -signature dlat-p--001-20120920--mycode.txt.sha1 \
        dlat-p--001-20120920--mycode.txt

# Verification of compressed contents
bunzip2 -c dlat-p--001-20120920--mycode.txt.bz2 | openssl dgst -sha1 -verify subKey.pem.pub \
        -signature dlat-p--001-20120920--mycode.txt.sha1
```

The procedure to submit analyzed contents to BLUEPRINT DCC also involves first having the raw data used for the analysis in the European Genome-phenome Archive (EGA), as all the metadata entries from the analyzed contents to be stored in BLUEPRINT DCC **must point** to the original raw data.

1.2.1 File Naming Conventions

Submitted files, containing either sample tracking data or analyzed experiment contents, must follow next file naming convention

```
featureType-fileType--institutionCode-dateFileCreated--freeField.txt
```

```
featureType-fileType--institutionCode-dateFileCreated--freeField.txt.sha1
```

The file name components are mapped in the next way:

Components	Description	Key
featureType	Sample Tracking data	sdata
	Gene Expression	exp
	Exon Junctions	jcn
	DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...)	dlat
	Protein-DNA interactions	pdna
	Regulatory regions	rreg
fileType	Metadata file	m
	Primary data file	p
	Secondary data file	s
	Gene expression file	g
	Donor file	donor
	Specimen file	specimen
	Sample file	sample
	Donor's Family file	family
institutionCode	Institution submitting data	CV Table A.5.1
dateFileCreated	The date on which the file is created	*YYYYMMDD* (ISO-8601)
freeField	An alphanumeric field (max length of 16 characters) where submitters can put internal codes, file sequence numbers, etc...	*e.g.: mysample, 0B1845J*

Different file types of the same feature type are interrelated, because the data they are storing is intertwined. Specific relations are defined on the documentation of each feature type and their file types. For instance, information stored in a primary data file is related and depends on the data from its corresponding metadata file, and the same happens to secondary data files and primary data files. Metadata file contents are related to sample tracking data sample files.

1.2.2 Tabular File Structure

The submitted analyzed contents are kept in tab-delimited text files. General comments may be added to the beginning of the file with a hash ('#') prefixed at beginning of each comment line. The first non-comment line is the header containing the names of the columns. Each column corresponds to a data element defined in DCC Submission Tabular Formats specification (Chapter 2).

There is a subset of comment lines used to attach data labels to the text files. These data labels follow the form '`##labelName value [value ...]`'. Currently acknowledged data labels are:

- **`format`**: This label is **required**, and its value defines the BLUEPRINT data formatting schema used on the file.
- **`depends`**: Although this label is not always required, it is important to validate the data coherence of the whole data set, because it ensures related data is not corrupted. The values of this label are the file on the same submission this file is related to (for instance, the name of a metadata file), and the SHA1 digest value (in its hexadecimal representation) of that file's contents.

There are several ways to generate the SHA1 digest of a file, like libraries in most of the programming languages and command-line tools:

SHA1 digest generation

```
# Getting the SHA1 digest value of uncompressed contents using OpenSSL
openssl dgst -sha1 dlat-p--001-20120920--mycode.txt

SHA1(dlat-p--001-20120920--mycode.txt)= 81ae49a7014d2d0260625d3535fa6e2a4a0bc06f

# Getting the SHA1 digest value of uncompressed contents using sha1sum
sha1sum dlat-p--001-20120920--mycode.txt

81ae49a7014d2d0260625d3535fa6e2a4a0bc06f  dlat-p--001-20120920--mycode.txt
```

An example file is shown below (note that parts of the lines are omitted for readability):

dlat-p--001-20120920--mycode.txt

```
# This is an example of a primary analysis file for simple somatic mutations.
# File name: dlat-p--001-20120920--mycode.txt
#
# And it has its labels
##format 0.2.3.4-a6979008015043a94ca2fa20d6178ec46524e858
##depends dlat-m--001-20120920--mycode.txt 03366af5145107cc818f4827e86b61dcf998ff29
analysis_id    ⤑analyzed_sample_id    ⤑d_lated_fragment_id    ⤑chromosome    ⤑...    ⤑note
an:001:000124  ⤑sample:001:000035      ⤑dlat:001:1234ff33      ⤑1             ⤑...    ⤑#FF#
an:001:000124  ⤑sample:001:000035      ⤑dlat:001:00019878      ⤑1             ⤑...    ⤑#FF#
an:001:000124  ⤑sample:001:000092      ⤑dlat:001:a712838       ⤑21            ⤑...    ⤑#FF#
an:001:000124  ⤑sample:001:000092      ⤑dlat:001:abebdZZZZZ    ⤑4             ⤑...    ⤑#FF#
```

All the declared columns for each file type must be set. Data columns are labeled as identifier or reference (**I**), required (**R**), desirable (**D**) or optional (**O**). Data providers (i.e. submitters) must put all the efforts in order to provide values for the idref and required data columns. The exception for this rule are the desirable fields, required fields which can be unknown on the first submissions, but in that case the fields these exceptions are properly documented.

There are several possible reasons why a column value (either desirable or optional) has not been provided. Next reserved codes must be used to describe the reason:

Code	Meaning
`#FF#`	Data not supplied at this time *(for future fill)*
`#NA#`	Not applicable for the context of the surrounding knowledge
`#VO#`	Data verified to be unknown (void, undef, null)
`#DE#`	Data derived from a required or idref field

Some data columns described in this submission manual contain values used as identifiers on BLUEPRINT DCC (e.g. `analysis_id`, `regulatory_region_id`, ...). As such, these identifiers should uniquely identify the entity they are referring (an analysis, a regulatory region, ...), and the identifier's value should be globally unique within a center's data submission. Also, these identifiers should be consistent along the different data submissions and releases. If you have to generate your own identifiers, there are some general recommendations, like using the same prefix for the identifiers of the same kind.

When you are submitting string values for columns which can contain URLs or multiple values delimited by commas, each separate value string, before being joined, should be URI encoded.

1.3 File Submission Procedure

Files with the contents to be submitted, along with their corresponding signed disgest, must be sent in a single tar archive. Either the tar archive or its embedded contents should be submitted compressed, using gzip, bzip2 or xz formats.

To be finished/defined

Chapter 2

DCC Submission Tabular Formats

2.1 Sample Tracking Submission File Specifications

Overview

There are three **required** sample and tissue annotation submission files, and one **optional** template file:

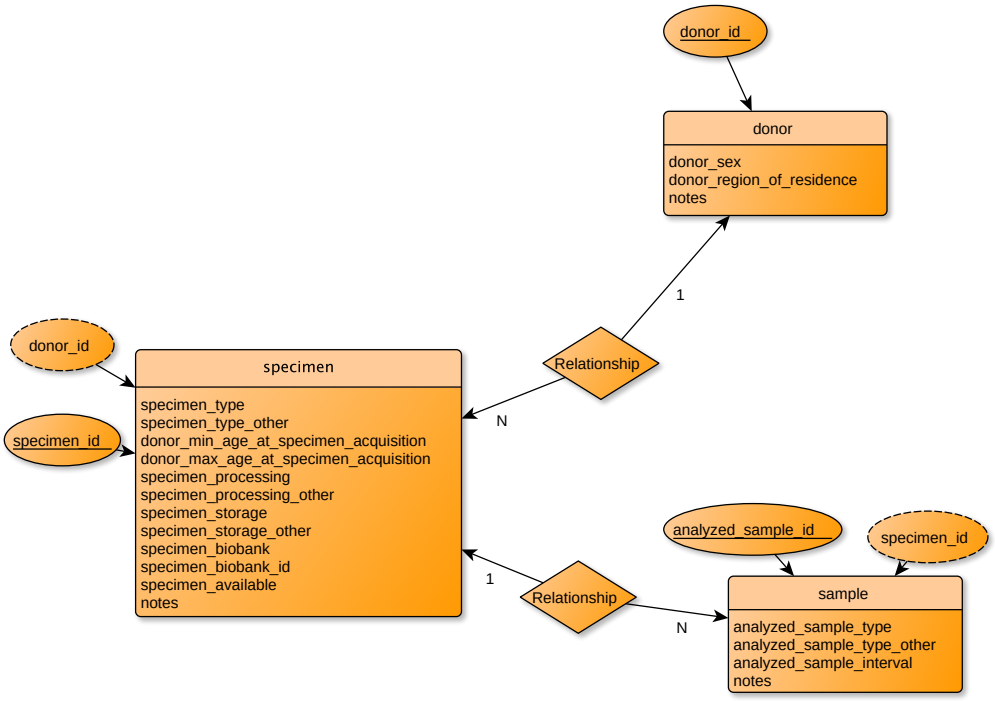

Figure 2.1: Sample Tracking Data Sub-Schema

Core Sample Tracking Data Files

1. *Donor Data File (donor)*
   Mandatory information about the donor's age, gender and vital status.

2. *Specimen Data File (specimen)*
   Mandatory information about a specimen that was obtained from a donor. There may be several specimens per donor that were obtained concurrently or at different times.

3. *Analyzed Sample Data File (sample)*
   Mandatory information about an analyzed sample that was subjected to molecular analysis. There may be several analyzed samples per specimen, for example, blood samples at different ages.

All data submissions to the DCC **must include the three core sample tracking data files.**

Optional Template Files

1. *Donor Family History (family)*
   Optional details about family history of the donor

Coding of donor IDs

The three mandatory data files contain donor, specimen and analyzed sample IDs, respectively. These IDs are to be coded specifically for BLUEPRINT purposes and only the submitting group will keep the key that will permit to link back the data to the individual donors. The key must not be communicated to the data users. It should not be derived from other IDs such as biobank or hospital identifiers. These IDs are to be coded in such a way that they cannot be tracked back to the individual donors, except by the submitting group. IDs are assigned by each submitting group, and must be unique within all the data submitted by that group (i.e. no duplicate IDs allowed). The DCC will prevent collisions between similar IDs submitted by different groups by including the project source column by default in all BioMart queries.

2.1.1 Donor Data File

Donor Data File [donor] (required)

This submission file describes a donor from which one or more specimens were obtained.

Table 2.1.1: Donor Data File

Name	Type	Need	Description / Values
donor_id	**string**	I	Unique identifier for the donor; assigned by data provider.
donor_region_of_residence	**string[]** (array seps **,**)	R	Country, and optionally state or province code, but not city. *ISO3166-1-alpha-2 or ISO3166-2 codes, eg: "CA" or "CA-ON" (See external CV description A.4)*
donor_sex	**string**	D	Donor biological sex. *"Other" has been removed from the controlled vocabulary due to identifiability concerns.* **m** = male **f** = female
notes	**string**	O	Any additional non-identifying information can be included here.

2.1.2 Specimen Data File

Specimen Data File [specimen] (required)

This submission file describes a specimen from which one or more samples were derived. Use additional rows for more than one specimen from the same patient. If more than one specimen was extracted during the same procedure, each gets a distinct ID.

Table 2.1.2: Specimen Data File

Name	Type	Need	Description / Values
Continued on next page			

Table 2.1.2 – continued from previous page

Name	Type	Need	Description / Values
specimen_id	**string**	I	Unique identifier for the specimen assigned by data provider.
donor_id	**string**	I	Unique identifier for the donor; assigned by data provider.
donor_max_age_at_specimen_acquisition	**duration**	D	Donor max age when the specimen was acquired. If it is '#DE#', then it is the same value as "donor_min_age_at_specimen_acquisition"
donor_min_age_at_specimen_acquisition	**duration**	R	Donor minimal age when the specimen was acquired, in ISO-8601 duration (basic format)
notes	**string**	O	Any additional non-identifying information can be included here.
specimen_available	**boolean**	O	Whether additional tissue is available for followup studies.
specimen_biobank	**string**	O	If the specimen was obtained from a biobank, provide the biobank name here
specimen_biobank_id	**string**	O	If the specimen was obtained from a biobank, provide the biobank accession number here.
specimen_processing	**string**	R	Description of technique used to process specimen **1** = cryopreservation in liquid nitrogen (dead tissue) **2** = cryopreservation in dry ice (dead tissue) **3** = cryopreservation of live cells in liquid nitrogen **4** = cryopreservation, other **5** = formalin fixed, unbuffered **6** = formalin fixed, buffered **7** = formalin fixed & paraffin embedded **8** = fresh **9** = other technique

Continued on next page

Table 2.1.2 – concluded from previous page

Name	Type	Need	Description / Values
specimen_processing_other	**string**	0	If "other" specified for specimen_processing, may indicate technique here.
specimen_storage	**string**	R	Description of how specimen was stored. For specimens that were extracted freshly or immediately cultured, answer (1) "NA". **1** = frozen, liquid nitrogen **2** = frozen, -70 freezer **3** = frozen, vapor phase **4** = RNA later frozen **5** = paraffin block **6** = cut slide **7** = other
specimen_storage_other	**string**	0	If "other" specified for specimen_storage, may indicate technique here.
specimen_type	**string**	R	Controlled vocabulary description of specimen type. **1** = primary tumour **2** = tumour local recurrence **3** = tumour metastasis to local lymph node **4** = tumour metastasis to distant location **per_blood** = peripheral blood **6** = bone marrow **7** = lymph node **c_tissue** = normal control (tissue adjacent to primary) **c_blood** = normal control (blood) **c_other** = normal control (other) **d_tissue** = disease tissue (other) **cord_blood** = cord blood
specimen_type_other	**string**	0	Free text description of site of specimen if "normal control (other)" or "disease tissue (other)" was specified in specimen_type field.

2.1.3 Analyzed Sample Data File

Analyzed Sample Data File [sample] (required)

This submission file describes an analyzed sample on which molecular characterization was performed. It includes both control samples (from healthy people) and samples from ill people.

Table 2.1.3: Analyzed Sample Data File

Name	Type	Need	Description / Values
analyzed_sample_id	**string**	I	Unique identifier for the sample assigned by data provider

Continued on next page

Table 2.1.3 – concluded from previous page

Name	Type	Need	Description / Values
specimen_id	**string**	I	Unique identifier for the specimen assigned by data provider.
analyzed_sample_interval	**integer**	0	Interval from specimen acquisition to sample use in an analytic procedure (e.g. DNA extraction), in days
analyzed_sample_type	**string**	R	Controlled vocabulary description of sample type **n_blood** = Normal blood **l_blood** = Leukemic blood **3** = Normal control adjacent to primary **4** = Normal control from non-tumour site **5** = Control from cell line derived from normal tissue **6** = Normal mouse host **7** = Primary tumour **8** = Mouse xenograft derived from tumour **9** = Cell line derived from tumour **10** = Cell line derived from xenograft **11** = Other (specify)
analyzed_sample_type_other	**string**	0	Free text description of site of sample if "other" was specified in sample_type field
notes	**string**	0	Any additional non-identifying information can be included here.
purified_cell_type	**string**	R	Purified cell type for the sample *(See external CV description A.3)*

2.2 Gene Expression

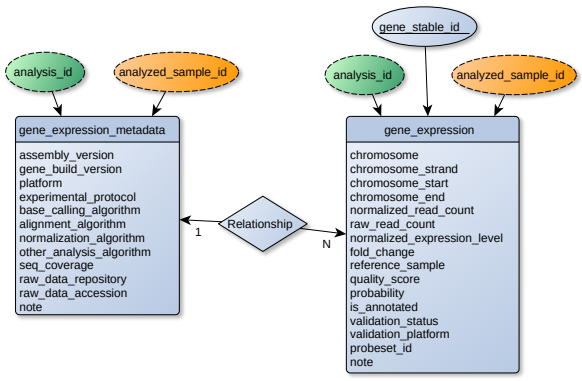

Figure 2.2: Gene Expression Sub-Schema

2.2.1 Expression - Metadata File

Expression [exp] – Metadata File [m]

Table 2.2.1: Expression - Metadata File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
alignment_algorithm	**compound** *name;url*	R	Name of alignment algorithm and URL to written protocol
analysis_group_id	**string**	R	Identifier of the analysis group (i.e. the one who prepared/run the pipeline) *(See CV A.5)*
assembly_version	**integer**	R	Version of reference genome assembly *(See CV A.8)*
data_status	**integer**	R	The status of the analysis over the associated raw data **0** = Raw data available at the EGA, but not more **1** = Raw data available at the EGA, analysis in process **2** = Analysis results obtained (analysis finished)
experimental_group_id	**string**	R	Identifier of the experimental group who did the experimental analysis *(See CV A.5)*
experimental_protocol	**compound** *name;url*	O	Name of experimental protocol and URL to written protocol
note	**string**	O	Optional field to leave notes
other_analysis_algorithm	**compound[]** *name;url* (array seps **,**)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	**integer**	R	Platform or technology used in the detection phase *(See CV A.6)*
program_versions	**compound[]** *program:version* (array seps **;**)	D	The versions of (some of) the programs used for the analysis
raw_data_accession	**compound** *accession;url*	O	Accession and URL for referencing the raw data at the public repository

Continued on next page

Table 2.2.1 – concluded from previous page

Name	Type	Need	Description / Values
raw_data_repository	**integer**	R	Public repository where raw data is submitted (#) *(See CV A.9)*
seq_coverage	**decimal**	O	Sequence coverage if analyzed by sequencing platforms

2.2.2 Expression - Gene File

Expression [exp] – Gene File [g]

Table 2.2.2: Expression - Gene File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
gene_stable_id	**string**	I	For annotated gene, use Ensembl gene ID. Otherwise, use assemblyBuild_chr_start_end where assemblyBuild is hg19.
chromosome	**string**	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV A.7)*
chromosome_end	**integer**	R	End position of the mutation/variation on the chromosome
chromosome_start	**integer**	R	Start position of the mutation/variation on the chromosome
chromosome_strand	**integer**	O	Strand where it was detected the mutation/variation on the chromosome **-1** = Reverse strand **1** = Forward strand
fold_change	**decimal**	O	Expressed fold change if differential expression is measured
is_annotated	**boolean**	R	If it is true, it indicate if the expressed fragment is annotated in GENCODE/Ensembl (i.e. gene_stable_id contains a Ensembl Gene Identifier)

Continued on next page

Table 2.2.2 – concluded from previous page

Name	Type	Need	Description / Values
normalized_expression_level	**decimal**	0	Normalized value of expression level if analyzed by microarray platforms
normalized_read_count	**decimal**	R	Normalized count of sequencing reads if analyzed by sequencing platforms
note	**string**	0	Optional field to leave notes
probability	**decimal**	0	Probability of the mutation/variation call
probeset_id	**string**	0	ID of the probeset used in microarray if analyzed by microarray platform
quality_score	**decimal**	0	Average quality score for the mutation/variation call
raw_read_count	**integer**	R	Raw count of sequencing reads if analyzed by sequencing platforms
reference_sample	**string**	0	ID of the reference analyzed sample if differential expression is measured
validation_platform	**integer**	0	Platform or technology used in validation *(See CV A.6)*
validation_status	**integer**	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.3 Exon Junction

The following diagram, based on the one from ICGC DCC manual, illustrates how junction_id should be generated, how junction_read_count, exon1_number_bases and exon2_number_bases are calculated:

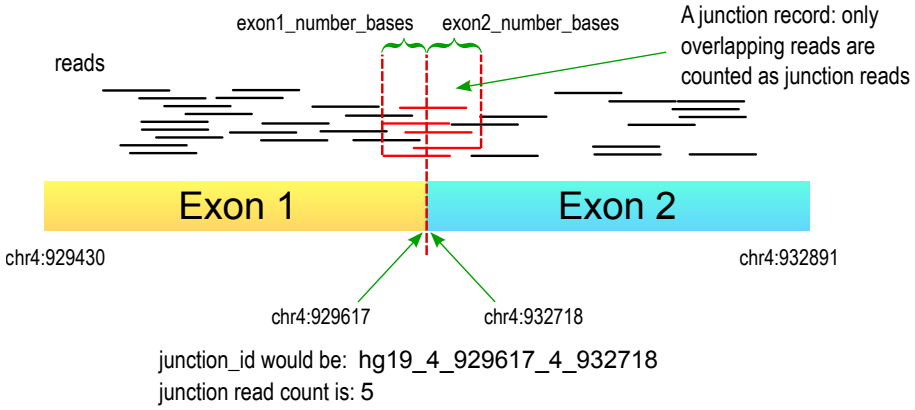

Figure 2.3: Junction Read Count explanation

2.3.1 Exon Junction - Metadata File

Exon Junction [jcn] – Metadata File [m]

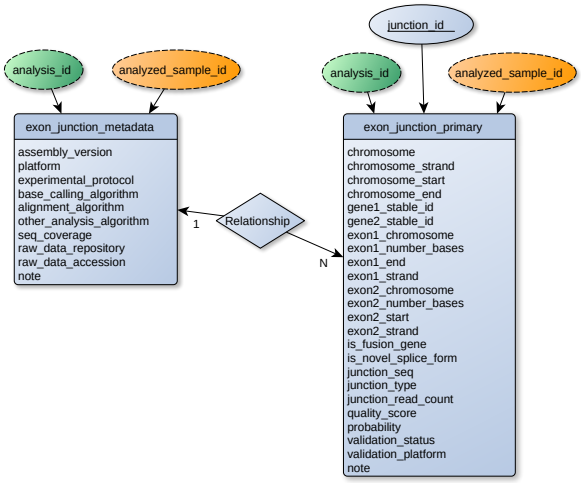

Figure 2.4: Exon Junction Sub-Schema

Table 2.3.1: Exon Junction - Metadata File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
alignment_algorithm	**compound** *name;url*	R	Name of alignment algorithm and URL to written protocol
analysis_group_id	**string**	R	Identifier of the analysis group (i.e. the one who prepared/run the pipeline) *(See CV A.5)*
assembly_version	**integer**	R	Version of reference genome assembly *(See CV A.8)*
data_status	**integer**	R	The status of the analysis over the associated raw data **0** = Raw data available at the EGA, but not more **1** = Raw data available at the EGA, analysis in process **2** = Analysis results obtained (analysis finished)

Continued on next page

Table 2.3.1 – concluded from previous page

Name	Type	Need	Description / Values
experimental_group_id	**string**	R	Identifier of the experimental group who did the experimental analysis *(See CV A.5)*
experimental_protocol	**compound** *name;url*	O	Name of experimental protocol and URL to written protocol
note	**string**	O	Optional field to leave notes
other_analysis_algorithm	**compound[]** *name;url* (array seps ,)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	**integer**	R	Platform or technology used in the detection phase *(See CV A.6)*
program_versions	**compound[]** *program:version* (array seps ;)	D	The versions of (some of) the programs used for the analysis
raw_data_accession	**compound** *accession;url*	O	Accession and URL for referencing the raw data at the public repository
raw_data_repository	**integer**	R	Public repository where raw data is submitted (#) *(See CV A.9)*
seq_coverage	**decimal**	O	Sequence coverage if analyzed by sequencing platforms

2.3.2 Exon Junction - Primary Analysis File

Exon Junction [jcn] – Primary Analysis File [p]

Table 2.3.2: Exon Junction - Primary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
junction_id	**string**	I	For known exons, use exonID1_exonID2 where exonID1 and exonID2 are Ensembl IDs of the 5' and 3' exons, respectively. For novel or putative exons,use assemblyBuild_exon1chr_exon1end_exon2chr_exon2start where assemblyBuild is hg18 or hg19; exon1chr and exon2chr are the chromosomes of the 5' and 3' exons, respectively; exon1end is the end position of the 5' exon; exon2start is the start position of the 3' exon.

Continued on next page

Table 2.3.2 – continued from previous page

Name	Type	Need	Description / Values
chromosome	**string**	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV A.7)*
chromosome_end	**integer**	R	End position of the mutation/variation on the chromosome
chromosome_start	**integer**	R	Start position of the mutation/variation on the chromosome
chromosome_strand	**integer**	0	Strand where it was detected the mutation/variation on the chromosome **-1** = Reverse strand **1** = Forward strand
exon1_chromosome	**string**	R	Name of the chromosome containing the 5' exon (#) *(See CV A.7)*
exon1_end	**integer**	R	End position of the 5' exon on the chromosome
exon1_number_bases	**integer**	R	Number of bases from 5' exon
exon1_strand	**integer**	0	Chromosome strand of the 5' exon **-1** = Reverse strand **1** = Forward strand
exon2_chromosome	**string**	R	Name of the chromosome containing the 3' exon (#) *(See CV A.7)*
exon2_number_bases	**integer**	R	Number of bases from 3' exon
exon2_start	**integer**	R	Start position of the 3' exon on the chromosome
exon2_strand	**integer**	0	Chromsome strand of the 3' exon **-1** = Reverse strand **1** = Forward strand
gene1_stable_id	**string**	R	Stable ID of the gene containing the 5' exon at the junction. For GENCODE/Ensembl annotated gene, use Ensembl gene ID. For putative and novel gene, use assemblyBuild_chr_start_end where assemblyBuild can be hg18 or hg19.
gene2_stable_id	**string**	0	In the case of a fusion gene, provide the Stable ID of the gene containing the 3' exon at the junction. For GENCODE/Ensembl annotated genes, use Ensembl gene ID. For putative and novel genes, use assemblyBuild_chr_start_end where assemblyBuild can be hg18 or hg19.

Continued on next page

Table 2.3.2 – concluded from previous page

Name	Type	Need	Description / Values
is_fusion_gene	**boolean**	0	Indicate if the function is the result of a fusion gene
is_novel_splice_form	**boolean**	0	Indicate if the splice form is novel
junction_read_count	**integer**	R	Count of sequencing reads that span across exons
junction_seq	**string**	0	Provide junction sequence if either is_fusion_gene or is_novel_splice_form is true
junction_type	**integer**	0	Type of junction **1** = Canonical **2** = Non-canonical **3** = U12
note	**string**	0	Optional field to leave notes
probability	**decimal**	0	Probability of the mutation/variation call
quality_score	**decimal**	0	Average quality score for the mutation/variation call
validation_platform	**integer**	0	Platform or technology used in validation *(See CV A.6)*
validation_status	**integer**	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.4 Protein-DNA interactions

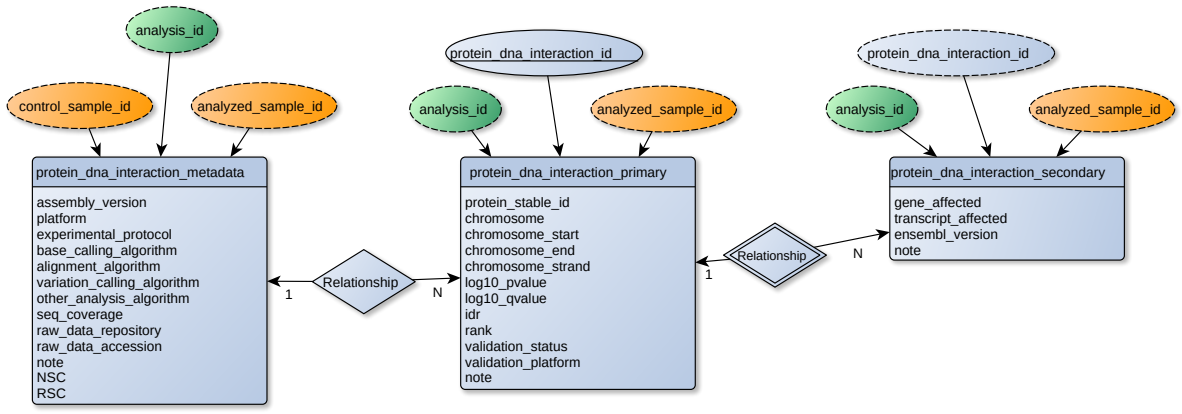

Figure 2.5: Protein-DNA interactions Sub-Schema

2.4.1 Protein-DNA interaction - Metadata File

Protein-DNA [pdna] – Metadata File [m]

Table 2.4.1: Protein-DNA interaction - Metadata File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
NSC	**decimal**	O	Normalized strand-cross correlation of the analysis
RSC	**decimal**	O	Relative strand-cross correlation of the analysis
alignment_algorithm	**compound** *name;url*	R	Name of alignment algorithm and URL to written protocol
analysis_group_id	**string**	R	Identifier of the analysis group (i.e. the one who prepared/run the pipeline) *(See CV A.5)*
assembly_version	**integer**	R	Version of reference genome assembly *(See CV A.8)*
control_sample_id	**string**	R	Unique identifier for the analyzed control/matched sample
data_status	**integer**	R	The status of the analysis over the associated raw data **0** = Raw data available at the EGA, but not more **1** = Raw data available at the EGA, analysis in process **2** = Analysis results obtained (analysis finished)
experimental_group_id	**string**	R	Identifier of the experimental group who did the experimental analysis *(See CV A.5)*
experimental_protocol	**compound** *name;url*	O	Name of experimental protocol and URL to written protocol
note	**string**	O	Optional field to leave notes
other_analysis_algorithm	**compound[]** *name;url* (array seps ,)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	**integer**	R	Platform or technology used in the detection phase *(See CV A.6)*

Continued on next page

Table 2.4.1 – concluded from previous page

Name	Type	Need	Description / Values
program_versions	**compound[]** *program:version* (array seps ;)	D	The versions of (some of) the programs used for the analysis
raw_data_accession	**compound** *accession;url*	O	Accession and URL for referencing the raw data at the public repository
raw_data_repository	**integer**	R	Public repository where raw data is submitted (#) *(See CV A.9)*
seq_coverage	**decimal**	O	Sequence coverage if analyzed by sequencing platforms

2.4.2 Protein-DNA interaction - Primary Analysis File

Protein-DNA [pdna] – Primary Analysis File [p]

Table 2.4.2: Protein-DNA interaction - Primary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
protein_dna_interaction_id	**string**	I	Unique identifier for the protein-DNA interaction
chromosome	**string**	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV A.7)*
chromosome_end	**integer**	R	End position of the mutation/variation on the chromosome
chromosome_start	**integer**	R	Start position of the mutation/variation on the chromosome
chromosome_strand	**integer**	O	Strand where it was detected the mutation/variation on the chromosome **-1** = Reverse strand **1** = Forward strand
idr	**decimal**	D	Irreproducible discovery rate
log10_pvalue	**decimal**	R	-log10(p-value)
log10_qvalue	**decimal**	D	-log10(q-value) , which available for peaks, but not for broad peaks

Continued on next page

Table 2.4.2 – concluded from previous page

Name	Type	Need	Description / Values
note	**string**	O	Optional field to leave notes
protein_stable_id	**string**	R	Stable id of the interacting protein, antibody or protein complex
rank	**compound[]** *rank:value* (array seps **;**)	O	Kind of used ranking and its value, in the form "rank;value". As it can hold more than one value, they are separated by bars
validation_platform	**integer**	O	Platform or technology used in validation *(See CV A.6)*
validation_status	**integer**	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.4.3 Protein-DNA interaction - Secondary Analysis File

Protein-DNA [pdna] – Secondary Analysis File [s]

Table 2.4.3: Protein-DNA interaction - Secondary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
ensembl_version	**integer**	R	Version of Ensembl gene build used for annotation (or the version of Ensembl gene build integrated into used GENCODE build)
gene_affected	**string[]** (array seps **|**)	R	Gene affected. Use Ensembl gene id, separated by | when there is more than one. If no gene is affected, don't put an entry *(See external CV description A.1)*
note	**string**	O	Optional field to leave notes
protein_dna_interaction_id	**string**	I	Unique identifier for the protein-DNA interaction
transcript_affected	**string[]** (array seps **|**)	O	Transcript on the protein-DNA interaction area. Use Ensembl transcript id. Separate multiple transcripts with vertical bars in the form of trasncriptA|trasncriptB|trasncriptC *(See external CV description A.2)*

2.5 Regulatory Regions

2.5.1 Regulatory regions - Metadata File

Regulatory regions [rreg] – Metadata File [m]

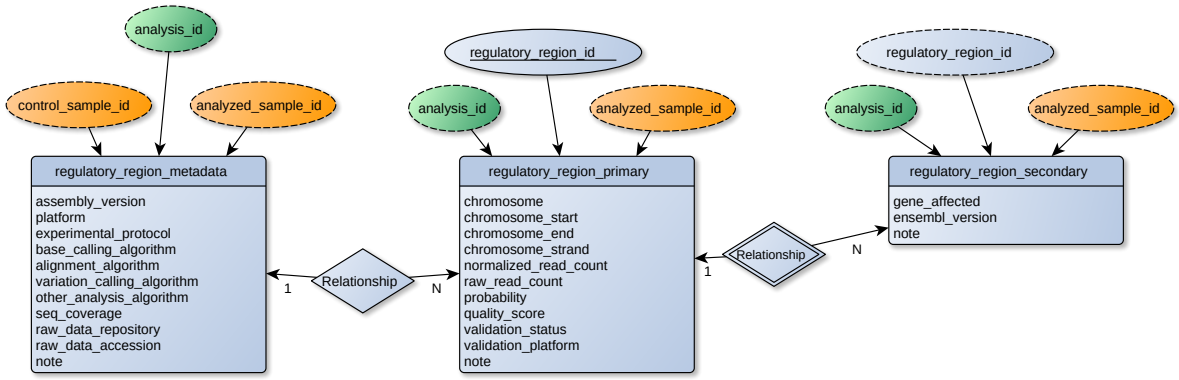

Figure 2.6: Regulatory Regions Sub-Schema

Table 2.5.1: Regulatory regions - Metadata File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
alignment_algorithm	**compound** *name;url*	R	Name of alignment algorithm and URL to written protocol
analysis_group_id	**string**	R	Identifier of the analysis group (i.e. the one who prepared/run the pipeline) *(See CV A.5)*
assembly_version	**integer**	R	Version of reference genome assembly *(See CV A.8)*
control_sample_id	**string**	R	Unique identifier for the analyzed control/matched sample
data_status	**integer**	R	The status of the analysis over the associated raw data **0** = Raw data available at the EGA, but not more **1** = Raw data available at the EGA, analysis in process **2** = Analysis results obtained (analysis finished)

Continued on next page

Table 2.5.1 – concluded from previous page

Name	Type	Need	Description / Values
experimental_group_id	**string**	R	Identifier of the experimental group who did the experimental analysis *(See CV A.5)*
experimental_protocol	**compound** *name;url*	O	Name of experimental protocol and URL to written protocol
note	**string**	O	Optional field to leave notes
other_analysis_algorithm	**compound[]** *name;url* (array seps ,)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	**integer**	R	Platform or technology used in the detection phase *(See CV A.6)*
program_versions	**compound[]** *program:version* (array seps ;)	D	The versions of (some of) the programs used for the analysis
raw_data_accession	**compound** *accession;url*	O	Accession and URL for referencing the raw data at the public repository
raw_data_repository	**integer**	R	Public repository where raw data is submitted (#) *(See CV A.9)*
seq_coverage	**decimal**	O	Sequence coverage if analyzed by sequencing platforms

2.5.2 Regulatory regions - Primary Analysis File

Regulatory regions [rreg] – Primary Analysis File [p]

Table 2.5.2: Regulatory regions - Primary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
regulatory_region_id	**string**	I	Unique identifier for the identified regulatory region
chromosome	**string**	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV A.7)*
chromosome_end	**integer**	R	End position of the mutation/variation on the chromosome

Continued on next page

Table 2.5.2 – concluded from previous page

Name	Type	Need	Description / Values
chromosome_start	**integer**	R	Start position of the mutation/variation on the chromosome
chromosome_strand	**integer**	0	Strand where it was detected the mutation/variation on the chromosome **-1** = Reverse strand **1** = Forward strand
normalized_read_count	**decimal**	R	Normalized count of sequencing reads if analyzed by sequencing platforms
note	**string**	0	Optional field to leave notes
probability	**decimal**	0	Probability of the mutation/variation call
quality_score	**decimal**	0	Average quality score for the mutation/variation call
raw_read_count	**integer**	R	Raw count of sequencing reads if analyzed by sequencing platforms
validation_platform	**integer**	0	Platform or technology used in validation *(See CV A.6)*
validation_status	**integer**	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.5.3 Regulatory regions - Secondary Analysis File

Regulatory regions [rreg] – Secondary Analysis File [s]

Table 2.5.3: Regulatory regions - Secondary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
ensembl_version	**integer**	R	Version of Ensembl gene build used for annotation (or the version of Ensembl gene build integrated into used GENCODE build)
gene_affected	**string[]** (array seps **|**)	R	Gene affected. Use Ensembl gene id, separated by | when there is more than one. If no gene is affected, don't put an entry *(See external CV description A.1)*

Continued on next page

Table 2.5.3 – concluded from previous page

Name	Type	Need	Description / Values
note	**string**	0	Optional field to leave notes
regulatory_region_id	**string**	I	Unique identifier for the identified regulatory region

2.6 DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...)

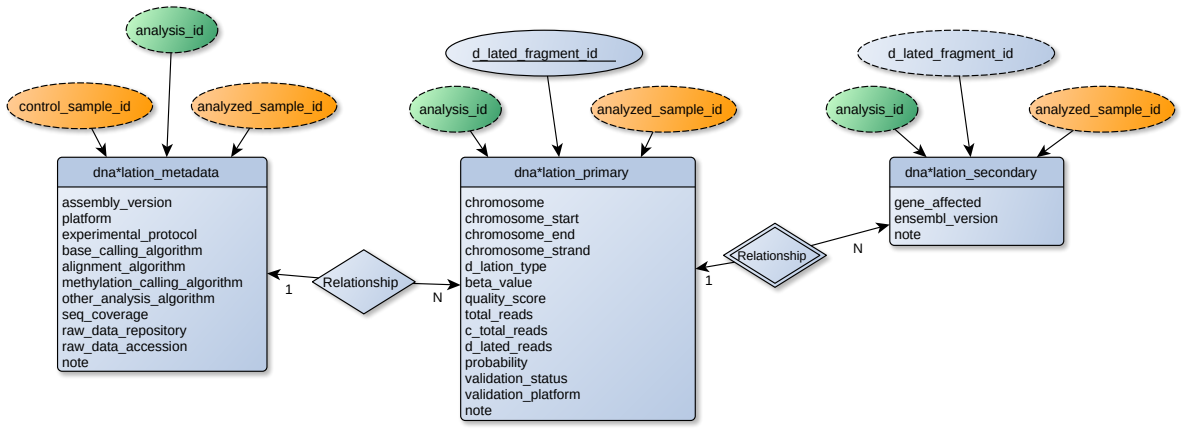

Figure 2.7: DNA Methylation , Hydroxy-Methylation, Formylation, etc... Sub-Schema

2.6.1 DNA *-lation - Metadata File

DNA *-lation [dlat] - Metadata File [m]

Table 2.6.1: DNA *-lation - Metadata File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
alignment_algorithm	**compound** *name;url*	R	Name of alignment algorithm and URL to written protocol
analysis_group_id	**string**	R	Identifier of the analysis group (i.e. the one who prepared/run the pipeline) *(See CV A.5)*
assembly_version	**integer**	R	Version of reference genome assembly *(See CV A.8)*

Continued on next page

Table 2.6.1 – concluded from previous page

Name	Type	Need	Description / Values
control_sample_id	**string**	R	Unique identifier for the analyzed control/matched sample
data_status	**integer**	R	The status of the analysis over the associated raw data **0** = Raw data available at the EGA, but not more **1** = Raw data available at the EGA, analysis in process **2** = Analysis results obtained (analysis finished)
experimental_group_id	**string**	R	Identifier of the experimental group who did the experimental analysis *(See CV A.5)*
experimental_protocol	**compound** *name;url*	O	Name of experimental protocol and URL to written protocol
methylation_calling_algorithm	**compound** *name;url*	O	Name of variation calling algorithm and URL to written protocol
note	**string**	O	Optional field to leave notes
other_analysis_algorithm	**compound[]** *name;url* (array seps **,**)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	**integer**	R	Platform or technology used in the detection phase *(See CV A.6)*
program_versions	**compound[]** *program:version* (array seps **;**)	D	The versions of (some of) the programs used for the analysis
raw_data_accession	**compound** *accession;url*	O	Accession and URL for referencing the raw data at the public repository
raw_data_repository	**integer**	R	Public repository where raw data is submitted (#) *(See CV A.9)*
seq_coverage	**decimal**	O	Sequence coverage if analyzed by sequencing platforms

2.6.2 DNA *-lation - Primary Analysis File

DNA *-lation [dlat] - Primary Analysis File [p]

Table 2.6.2: DNA *-lation - Primary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples

Continued on next page

Table 2.6.2 – continued from previous page

Name	Type	Need	Description / Values
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
d_lated_fragment_id	**string**	I	Unique identifier for the methylated fragment, in the form d'*lationType|chromosome_chromosomeStart_chromosomeEnd
beta_value	**decimal**	O	DNA *-lation beta value for interrogated site
c_total_reads	**decimal**	R	Reads which has identified this position as a cytosine
chromosome	**string**	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV A.7)*
chromosome_end	**integer**	R	End position of the mutation/variation on the chromosome
chromosome_start	**integer**	R	Start position of the mutation/variation on the chromosome
chromosome_strand	**integer**	O	Strand where it was detected the mutation/variation on the chromosome **-1** = Reverse strand **1** = Forward strand
d_lated_reads	**decimal**	R	Reads which has identified this position as a DNA *lated cytosine
d_lation_type	**string**	R	Type of DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...) **m** = Methylation (cytosine) **hm** = Hydroxy-Methylation (cytosine) **hmU** = Hydroxy-Methylation (uracil) **f** = Formylation (cytosine) **ca** = Carboxylation (cytosine)
note	**string**	O	Optional field to leave notes
probability	**decimal**	O	Probability of the DNA *-lation call
quality_score	**decimal**	O	Quality score for the DNA *-lation call
total_reads	**decimal**	R	Total number of reads over this position/segment, either identifying or not a cytosine, for sequencing platforms. Mean reads depth for other technologies

Continued on next page

Table 2.6.2 – concluded from previous page

Name	Type	Need	Description / Values
validation_platform	**integer**	0	Platform or technology used in validation *(See CV A.6)*
validation_status	**integer**	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.6.3 DNA *-lation - Secondary Analysis File

DNA *-lation [dlat] - Secondary Analysis File [s]

Table 2.6.3: DNA *-lation - Secondary Analysis File

Name	Type	Need	Description / Values
analysis_id	**string**	I	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	**string**	I	Unique identifier for the analyzed sample
d_lated_fragment_id	**string**	I	Unique identifier for the methylated fragment, in the form d'*lationType|chromosome_chromosomeStart_chromosomeEnd
ensembl_version	**integer**	R	Version of Ensembl gene build used for annotation (or the version of Ensembl gene build integrated into used GENCODE build)
gene_affected	**string[]** (array seps **|**)	R	Gene affected. Use Ensembl gene id, separated by | when there is more than one. If no gene is affected, don't put an entry *(See external CV description A.1)*
note	**string**	O	Optional field to leave notes

Appendix A

Controlled Vocabularies

A.1 Ensembl Genes

Valid Ensembl Genes identifiers
(See it at http://jan2013.archive.ensembl.org/Homo_sapiens/Info/Index)

A.2 Ensembl Transcripts

Valid Ensembl Transcript identifiers
(See it at http://jan2013.archive.ensembl.org/Homo_sapiens/Info/Index)

A.3 Cell Ontology

The Cell Ontology is designed as a structured controlled vocabulary for cell types
(See it at http://cellontology.org/)

A.4 ISO 3166-1 and ISO 3166-2

ISO 3166 is the International Standard for country codes and codes for their subdivisions. The purpose of ISO 3166 is to establish internationally recognised codes for the representation of names of countries, territories or areas of geographical interest, and their subdivisions.
(See them at http://www.iso.org/iso/country_codes.htm and http://en.wikipedia.org/wiki/ISO_3166-2)

Table A.4.1: ISO 3166-1 and ISO 3166-2 aliases

Alias	Key	Description
ALIAS:EAL	GB-CAM GB-ESS GB-HRT GB-NFK GB-SFK	East Anglia: United Kingdom region composed of the administrative counties of Norfolk to the north, Suffolk to the south, Cambridgeshire and Essex to the west.

A.5 Institution ID

Please contact BLUEPRINT DCC if your institution is not listed, or you wish to modify the text

Table A.5.1: Institution ID

ID	Institution
1	Radboud University Nijmegen (H.G. Stunnenberg)
2a	University College London (S. Beck)
2b	University College London (T. Enver)
3a	University of Cambridge (A. Ferguson-Smith)
3b	University of Cambridge (W. H.Ouwehand)
4	Friedrich Miescher Institute (D. Schübeler)
5	Christian Albrechts University of Kiel (R. Siebert)
6	National Cancer Research Centre Spain (A. Valencia)
7a	Institute of Molecular Oncology Foundation - European Institute of Oncology (P.G. Pelicci)
7b	Institute of Molecular Oncology Foundation - European Institute of Oncology (S. Minucci)
8	European Bioinformatics Institute (P. Flicek)
9a	Wellcome Trust Sanger Institute (M. Stratton)
9b	Wellcome Trust Sanger Institute (D. Adams)
9c	Wellcome Trust Sanger Institute (N. Soranzo)
10	Bellvitge Institute for Biomedical Research (M. Esteller)
11	Centro Nacional de Analysis Genómico (I. Gut)
12a	Max Planck Institute for Bioinformatics (T. Lengauer/C.Bock)
12b	Max Planck Institute for Molecular Genetics (H. Lehrach)
12c	Max Planck Institute for Molecular Genetics (M. Vingron)
13	University of Saarland (J. Walter)
14	Second University of Naples (L. Altucci)
15a	Centre for Genomic Regulation (X. Estivill)
15b	Centre for Genomic Regulation (R. Guigo)
15c	Centre for Genomic Regulation (T. Graf)
16a	Queen Mary, University of London (D. Leslie/V. Rakyan)
16b	Queen Mary, University of London (J. Fitzgibbon)
17	The Babraham Institute (W. Reik)
18	Cellzome AG (D. Simmons)
19	Diagenode SA (D. Allaer)
20	Olink Genomics (F. Dahl)
21	Genomatix Software GmbH (M. Seifert)

Continued on next page

Table A.5.1 – concluded from previous page

ID	Institution
22	Oxford Nanopore Technologies Ltd (S. Willcocks)
23	Siena Biotech SpA (A. Caricasole)
24	Centre of Immunology of Marseille-Luminy (S. Spicuglia)
25	Institut d'Investigacions Biomèdique August Pi i Sunyer (E. Campo)
26	Weizmann Institute of Science (A. Tanay)
27	Erasmus University Medical Centre Rotterdam (F. Grosveld)
28	Universitaetsklinikum Ulm (B. Böhm)
29	University of Edinburgh (A. Bird)
30	Lund University (A. Lernmark)
31	University of Copenhagen (K. Helin)
32	Sapienza University of Rome (A. Mai)
33	Vivia Biotech S.L. (J. Ballesteros)
34	University of Geneva (M. Dermitzakis, S. Antonorakis)
35	University Medical Centre Groningen (E. Vellenga)
36	Neckar Hospital (Elizbeth Macintyre)
37	Epigenomics AG (R. Wasserkort)
38	University of Duisburg-Essen (R. Küppers)
39	University of Leipzig (M. Loffler)
40	Barcelona Supercomputing Center (D. Torrents)
41	Sigolis (J. Jarvius)
42	Eurice (V. Siegmund)

A.6 Value Codes for Platform or Validation Platform

Please contact the DCC if your platform/technology is not listed here.

Table A.6.1: Value Codes for Platform or Validation Platform

Key	Platform or Validation Platform
1	PCR
2	qPCR
3	capillary sequencing
4	SOLiD sequencing
5	Illumina GA sequencing
6	454 sequencing
7	Helicos sequencing
8	Affymetrix Genome-Wide Human SNP Array 6.0
9	Affymetrix Genome-Wide Human SNP Array 5.0
10	Affymetrix Mapping 100K Array Set
11	Affymetrix Mapping 500K Array Set
12	Affymetrix Mapping 10K 2.0 Array Set
13	Affymetrix EMET Plus Premier Pack

Continued on next page

Table A.6.1 – continued from previous page

Key	Platform or Validation Platform
14	Agilent Whole Human Genome Oligo Microarray Kit
15	Agilent Human Genome 244A
16	Agilent Human Genome 105A
17	Agilent Human CNV Association 2x105K
18	Agilent Human Genome 44K
19	Agilent Human CGH 1x1M
20	Agilent Human CGH 2x400K
21	Agilent Human CGH 4x180K
22	Agilent Human CGH 8x60K
23	Agilent Human CNV 2x400K
24	Agilent Human miRNA Microarray Kit (v2)
25	Agilent Human CpG Island Microarray Kit
26	Agilent Human Promoter ChIP-on-chip Microarray Set
27	Agilent Human SpliceArray
28	Illumina human1m-duo
29	Illumina human660w-quad
30	Illumina humancytosnp-12
31	Illumina human510s-duo
32	Illumina humanmethylation27
33	Illumina goldengate methylation
34	Illumina HumanHT-12 v4.0 beadchip
35	Illumina HumanWG-6 v3.0 beadchip
36	Illumina HumanRef-8 v3.0 beadchip
37	Illumina microRNA Expression Profiling Panel
38	Illumina humanht-16
39	Illumina humanht-17
40	Nimblegen Human CGH 3x720 Whole-Genome v3.0 Array
41	Nimblegen Human CGH 2.1M Whole-Genome v2.0D Array
42	Nimblegen Gene Expression 385K
43	Nimblegen Gene Expression 4x72K
44	Nimblegen Gene Expression 12x135K
45	Nimblegen Human Methylation 2.1M Whole-Genome sets
46	Nimblegen Human Methylation 385K Whole-Genome sets
47	Nimblegen CGS
48	Illumina Human1M OmniQuad chip
49	PCR and capillary sequencing
50	Custom-designed gene expression array
51	Affymetrix HT Human Genome U133A Array Plate Set

Continued on next page

Table A.6.1 – concluded from previous page

Key	Platform or Validation Platform
52	Agilent 244K Custom Gene Expression G4502A-07-1
53	Agilent 244K Custom Gene Expression G4502A-07-2
54	Agilent 244K Custom Gene Expression G4502A-07-3
55	Agilent Human Genome CGH Custom Microaary 2x415K
56	Affymetrix Human U133 Plus PM
57	Affymetrix Human U133 Plus 2.0
58	Affymetrix Human Exon 1.0 ST
59	Almac Human CRC
60	Illumina HiSeq
61	Affymetrix Human MIP 330K
62	Affymetrix Human Gene 1.0 ST
63	Illumina Human Omni1-Quad beadchip
64	Sequenom MassARRAY
65	Custom-designed cDNA array
66	Illumina HumanHap550
67	Ion Torrent PGM
68	Illumina GoldenGate Methylation Cancer Panel I
69	Illumina Infinium HumanMethylation450
70	Agilent 8 x 15K Human miRNA-specific microarray
71	M.D. Anderson Reverse Phase Protein Array Core
72	Microsatellite Instability Analysis
73	Agilent 244K Custom Gene Expression G4502A-07
74	Illumina HumanCNV370-Duo v1.0 BeadChip
75	Illumina HumanOmniExpress BeadChip

A.7 Chromosome Names for Reference Genome GRCh37

Table A.7.1: Chromosome Names for Reference Genome GRCh37

Key	Chromosome Name
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13

Continued on next page

Table A.7.1 – continued from previous page

Key	Chromosome Name
14	14
15	15
16	16
17	17
18	18
19	19
20	20
21	21
22	22
X	X
Y	Y
MT	MT
c5_H2	c5_H2
c6_COX	c6_COX
c6_QBL	c6_QBL
NT_113870	NT_113870
NT_113871	NT_113871
NT_113872	NT_113872
NT_113874	NT_113874
NT_113878	NT_113878
NT_113880	NT_113880
NT_113881	NT_113881
NT_113884	NT_113884
NT_113885	NT_113885
NT_113886	NT_113886
NT_113888	NT_113888
NT_113889	NT_113889
NT_113890	NT_113890
NT_113898	NT_113898
NT_113899	NT_113899
NT_113901	NT_113901
NT_113902	NT_113902
NT_113903	NT_113903
NT_113906	NT_113906
NT_113908	NT_113908
NT_113909	NT_113909
NT_113910	NT_113910
NT_113911	NT_113911
NT_113912	NT_113912
NT_113915	NT_113915
NT_113916	NT_113916
NT_113917	NT_113917

Continued on next page

Table A.7.1 – continued from previous page

Key	Chromosome Name
NT_113923	NT_113923
NT_113924	NT_113924
NT_113925	NT_113925
NT_113926	NT_113926
NT_113927	NT_113927
NT_113929	NT_113929
NT_113930	NT_113930
NT_113931	NT_113931
NT_113932	NT_113932
NT_113933	NT_113933
NT_113934	NT_113934
NT_113935	NT_113935
NT_113936	NT_113936
NT_113937	NT_113937
NT_113939	NT_113939
NT_113943	NT_113943
NT_113944	NT_113944
NT_113946	NT_113946
NT_113949	NT_113949
NT_113951	NT_113951
NT_113953	NT_113953
NT_113954	NT_113954
NT_113956	NT_113956
NT_113957	NT_113957
NT_113958	NT_113958
NT_113960	NT_113960
NT_113961	NT_113961
NT_113962	NT_113962
NT_113963	NT_113963
NT_113964	NT_113964
NT_113965	NT_113965
NT_113966	NT_113966
HSCHR17_1	HSCHR17_1
HSCHR17_RANDOM_CTG2	HSCHR17_RANDOM_CTG2
HSCHR17_RANDOM_CTG3	HSCHR17_RANDOM_CTG3
HSCHR19_RANDOM_CTG2	HSCHR19_RANDOM_CTG2
HSCHR1_RANDOM_CTG12	HSCHR1_RANDOM_CTG12
HSCHR1_RANDOM_CTG5	HSCHR1_RANDOM_CTG5
HSCHR4_RANDOM_CTG2	HSCHR4_RANDOM_CTG2
HSCHR4_RANDOM_CTG3	HSCHR4_RANDOM_CTG3
HSCHR6_MHC_APD	HSCHR6_MHC_APD
HSCHR6_MHC_COX	HSCHR6_MHC_COX

Continued on next page

Table A.7.1 – concluded from previous page

Key	Chromosome Name
HSCHR6_MHC_DBB	HSCHR6_MHC_DBB
HSCHR6_MHC_MANN	HSCHR6_MHC_MANN
HSCHR6_MHC_MCF	HSCHR6_MHC_MCF
HSCHR6_MHC_QBL	HSCHR6_MHC_QBL
HSCHR6_MHC_SSTO	HSCHR6_MHC_SSTO
HSCHR7_RANDOM_CTG1	HSCHR7_RANDOM_CTG1
HSCHR8_RANDOM_CTG1	HSCHR8_RANDOM_CTG1
HSCHR8_RANDOM_CTG4	HSCHR8_RANDOM_CTG4
HSCHR9_RANDOM_CTG2	HSCHR9_RANDOM_CTG2
HSCHR9_RANDOM_CTG4	HSCHR9_RANDOM_CTG4
HSCHR9_RANDOM_CTG5	HSCHR9_RANDOM_CTG5
HSCHRUN_RANDOM_CTG1	HSCHRUN_RANDOM_CTG1
HSCHRUN_RANDOM_CTG10	HSCHRUN_RANDOM_CTG10
HSCHRUN_RANDOM_CTG11	HSCHRUN_RANDOM_CTG11
HSCHRUN_RANDOM_CTG13	HSCHRUN_RANDOM_CTG13
HSCHRUN_RANDOM_CTG14	HSCHRUN_RANDOM_CTG14
HSCHRUN_RANDOM_CTG15	HSCHRUN_RANDOM_CTG15
HSCHRUN_RANDOM_CTG16	HSCHRUN_RANDOM_CTG16
HSCHRUN_RANDOM_CTG17	HSCHRUN_RANDOM_CTG17
HSCHRUN_RANDOM_CTG2	HSCHRUN_RANDOM_CTG2
HSCHRUN_RANDOM_CTG20	HSCHRUN_RANDOM_CTG20
HSCHRUN_RANDOM_CTG21	HSCHRUN_RANDOM_CTG21
HSCHRUN_RANDOM_CTG22	HSCHRUN_RANDOM_CTG22
HSCHRUN_RANDOM_CTG23	HSCHRUN_RANDOM_CTG23
HSCHRUN_RANDOM_CTG26	HSCHRUN_RANDOM_CTG26
HSCHRUN_RANDOM_CTG29	HSCHRUN_RANDOM_CTG29
HSCHRUN_RANDOM_CTG3	HSCHRUN_RANDOM_CTG3
HSCHRUN_RANDOM_CTG30	HSCHRUN_RANDOM_CTG30
HSCHRUN_RANDOM_CTG31	HSCHRUN_RANDOM_CTG31
HSCHRUN_RANDOM_CTG32	HSCHRUN_RANDOM_CTG32
HSCHRUN_RANDOM_CTG33	HSCHRUN_RANDOM_CTG33
HSCHRUN_RANDOM_CTG34	HSCHRUN_RANDOM_CTG34
HSCHRUN_RANDOM_CTG35	HSCHRUN_RANDOM_CTG35
HSCHRUN_RANDOM_CTG36	HSCHRUN_RANDOM_CTG36
HSCHRUN_RANDOM_CTG4	HSCHRUN_RANDOM_CTG4
HSCHRUN_RANDOM_CTG40	HSCHRUN_RANDOM_CTG40
HSCHRUN_RANDOM_CTG5	HSCHRUN_RANDOM_CTG5
HSCHRUN_RANDOM_CTG6	HSCHRUN_RANDOM_CTG6
HSCHRUN_RANDOM_CTG9	HSCHRUN_RANDOM_CTG9
HSCHR4_1	HSCHR4_1

A.8 Value Codes for Reference Genome Assembly Version

Table A.8.1: Value Codes for Reference Genome Assembly Version

Key	Reference Genome Assembly Version

Continued on next page

Table A.8.1 – concluded from previous page

Key	Reference Genome Assembly Version
1	GRCh37
2	NCBI36
3	GRCh37.p1
4	GRCh37.p2
5	GRCh37.p3
6	GRCh37.p4
7	GRCh37.p5

A.9 Value Codes for Raw Data Repository

Table A.9.1: Value Codes for Raw Data Repository

Key	Raw Data Repository
1	EGA
2	dbSNP
3	TCGA
4	CGHub
5	GEO