BLUEPRINT proposed data submission schemas
Draft v0.2.0

José María Fernández

January 26, 2013

Contents

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 **Ensembl gene set version 68**

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 Clinical Data and Analyzed Contents for their submission

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

- Clinical data
- Copy Number Variations
- Gene Expression
- Exon Junctions
- DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...)
- Protein-DNA interactions
- Regulatory regions

BLUEPRINT DCC is hosting both clinical 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.

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 clinical 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	Clinical data	cdata
	Copy Number Variations	cngv
	Gene Expression	exp
	Exon Junctions	jcn
	DNA *-lation (Methylation, Hydroxy-Methylation, Formylation, etc...)	dlat
	Protein-DNA interactions	pdna
	Regulatory regions	rreg
institutionCode	Institution submitting data	CV Table A.1
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
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 clinical 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 0.2.0
##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  ⤑...  ⤑-999
an:001:000124  ⤑sample:001:000035  ⤑dlat:001:00019878  ⤑1  ⤑...  ⤑-999
an:001:000124  ⤑sample:001:000092  ⤑dlat:001:a712838  ⤑21  ⤑...  ⤑-999
an:001:000124  ⤑sample:001:000092  ⤑dlat:001:abebdZZZZZ  ⤑4  ⤑...  ⤑-999
```

All the declared columns for each file type must be set. Data columns are labeled as either required (R) or optional (O). Data providers (i.e. submitters) must put all the efforts in order to provide values for the required data columns. There are some exceptions for this rule, where required fields are unknown on the first submission, but these exceptions are properly documented.

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

Code	Meaning
-999	Data not supplied at this time
-888	Not applicable
-777	Data verified to be unknown

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.

The following steps are involved in submitting your project's data files to DCC:

1. Contact xxx and notify the DCC of your intent to submit data.
2. The DCC will provide an SFTP or Aspera account for uploading your data to the DCC's secure server xxx.
3. Prepare an archive containing the set of data files (along with their signed digests) comprising your submission.
4. Generate a signed digest of the archive your are going to upload, which is also going to be uploaded.
5. Login into your SFTP or Aspera account and upload both the data archive and its signed digest.
6. Notify the DCC of your successful upload, so they can start the internal validation and processing.

To be finished/defined

Chapter 2

DCC Submission Tabular Formats

2.1 Clinical Data Submission File Specifications

Overview

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

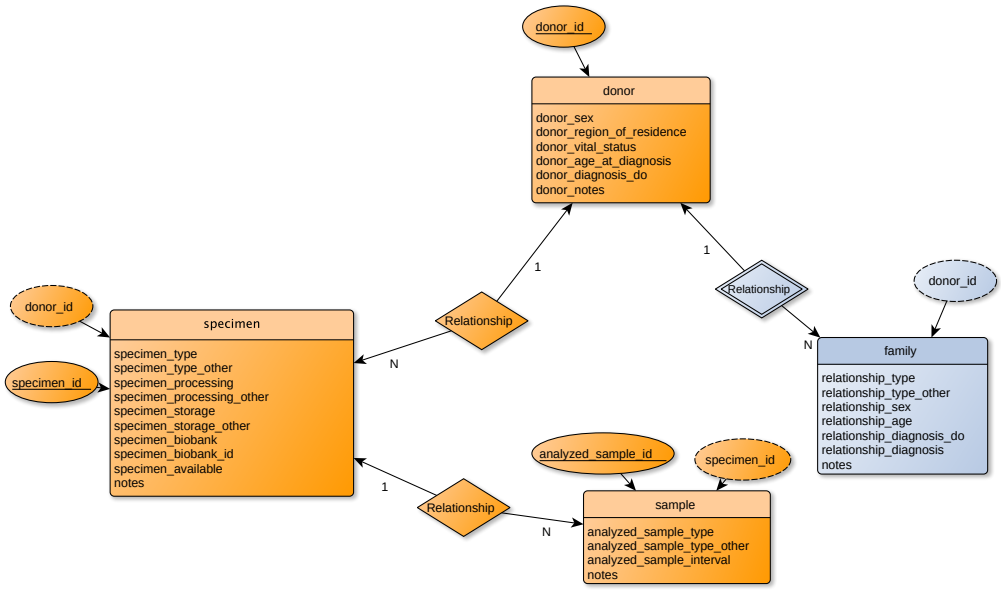

Figure 2.1: Clinical Data Sub-Schema

Core Clinical 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, when a tumour is used to derive xenografts and cell lines.

All data submissions to the DCC **must include the three core clinical 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: Donor Data File

Name	Type	R/O	Description / Values
donor_id	string	R	Unique identifier for the donor; assigned by data provider.
donor_age_at_diagnosis	integer	R	Age at primary diagnosis *Use "90" for patients >=90*
donor_diagnosis_do	string	R	Disease Ontology code *Disease Ontology code*
donor_region_of_residence	string	R	Country, and optionally state or province code, but not city. *ISO3166-1-alpha-2 or ISO3166-2 codes, eg: "CA" or "CA-ON"*
donor_sex	integer	R	Donor biological sex. *"Other" has been removed from the controlled vocabulary due to identifiability concerns.* 1 = male 2 = female
donor_vital_status	integer	O	Donor's last known vital status 1 = alive 2 = deceased
notes	string	O	Any additional non-identifying information can be included here.

2.1.2 Donor Family History

Donor Family History [family] (optional)

This file describes the family history of the donor.

Table 2.2: Donor Family History

Name	Type	R/O	Description / Values
Continued on next page			

Table 2.2 – concluded from previous page

Name	Type	R/O	Description / Values
donor_id	string	R	Unique identifier for the donor; assigned by data provider.
notes	string	O	Any additional non-identifying information can be included here.
relationship_age	integer	R	Age of relative at primary diagnosis (years) *Use 90 for ages >= 90 years.*
relationship_diagnosis	string	R	Diagnosis (disease or healthy status) *e.g. "breast cancer"*
relationship_diagnosis_do	string	R	Disease Ontology code for the relative's diagnosis status
relationship_sex	integer	R	Biological sex of related individual **1** = male **2** = female
relationship_type	integer	R	Relationship to the donor **1** = sibling **2** = parent **3** = grandparent **4** = uncle/aunt **5** = cousin **6** = other
relationship_type_other	string	R	If "other" answered in previous column, specify the relationship type here

2.1.3 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 2.3: Specimen Data File

Name	Type	R/O	Description / Values
specimen_id	string	R	Unique identifier for the specimen assigned by data provider.
donor_id	string	R	Unique identifier for the donor; assigned by data provider.
notes	string	O	Any additional non-identifying information can be included here.

Continued on next page

Table 2.3 – continued from previous page

Name	Type	R/O	Description / Values
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
specimen_processing_other	string	O	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	O	If "other" specified for specimen_storage, may indicate technique here.
specimen_type	integer	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 **5** = peripheral blood **6** = bone marrow **7** = lymph node **8** = normal control (tissue adjacent to primary) **9** = normal control (blood) **10** = normal control (other) **11** = disease tissue (other)

Continued on next page

Table 2.3 – concluded from previous page

Name	Type	R/O	Description / Values
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.

2.1.4 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 and tumour samples.

Table 2.4: Analyzed Sample Data File

Name	Type	R/O	Description / Values
analyzed_sample_id	string	R	Unique identifier for the sample assigned by data provider
specimen_id	string	R	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	integer	R	Controlled vocabulary description of sample type **1** = Normal blood **2** = 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)

Continued on next page

Table 2.4 – concluded from previous page

Name	Type	R/O	Description / Values
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.

2.2 Copy Number Germline Variations

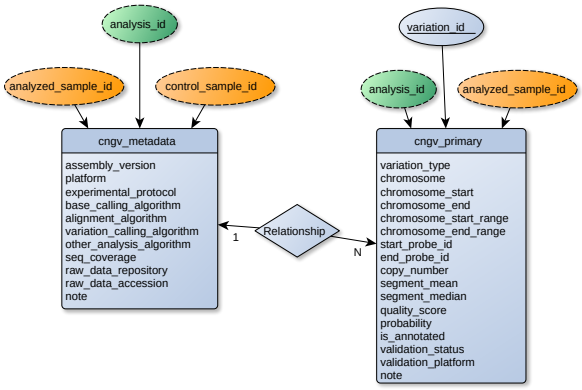

Figure 2.2: Copy Number Germline Variations Sub-Schema

2.2.1 Copy Number Germline Variations - Metadata File

Copy Number Germline Variations [cngv] – Metadata File [m]

Table 2.5: Copy Number Germline Variations - Metadata File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
alignment_algorithm	compound	R	Name of alignment algorithm and URL to written protocol
assembly_version	integer	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	compound	R	Name of base calling algorithm and URL to written protocol

Continued on next page

Table 2.5 – concluded from previous page

Name	Type	R/O	Description / Values
control_sample_id	string	R	Unique identifier for the analyzed control/matched sample
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms
variation_calling_algorithm	compound	R	Name of variation calling algorithm and URL to written protocol

2.2.2 Copy Number Germline Variations – Primary Analysis File

Copy Number Germline Variations [cngv] – Primary Analysis File [p]

Table 2.6: Copy Number Germline Variations – Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
variation_id	string	R	Unique identifier for the variation
chromosome	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
chromosome_end	integer	R	End position of the mutation/variation on the chromosome

Continued on next page

Table 2.6 – continued from previous page

Name	Type	R/O	Description / Values
chromosome_end_range	integer	R	Number of bases around chromosome_end that may contain the end position *0 if end position is exactly at chromosome_end; positive integer for +/- number of bases around chromosome_end*
chromosome_start	integer	R	Start position of the mutation/variation on the chromosome
chromosome_start_range	integer	R	Number of bases around chromosome_start that may contain the start position *0 if start position is exactly at chromosome_start; positive integer for +/- number of bases around chromosome_start*
chromosome_strand	integer	O	Strand where it was detected the mutation/variation on the chromosome **-1** = -1 **1** = 1
copy_number	decimal	O	DNA copy number estimated
end_probe_id	string	O	Probe id containing the chromosome_end (if array platform was used)
is_annotated	boolean	O	Indicate if the variation is annotated in the database of Genomic Variations
note	string	O	Optional field to leave notes
probability	decimal	O	Probability of the mutation/variation call
quality_score	decimal	O	Average quality score for the mutation/variation call
segment_mean	decimal	O	Mean LRR per segment
segment_median	decimal	O	Median LRR per segment
start_probe_id	string	O	Probe id containing the chromosome_start (if array platform was used)
validation_platform	integer	O	Platform or technology used in validation *(See CV Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

Continued on next page

Table 2.6 – concluded from previous page

Name	Type	R/O	Description / Values
variation_type	`string`	R	Type of variation **1** = gain **2** = loss **3** = copy neutral LOH **4** = copy neutral **5** = hemizigous del LOH **6** = apm LOH

2.3 Gene Expression

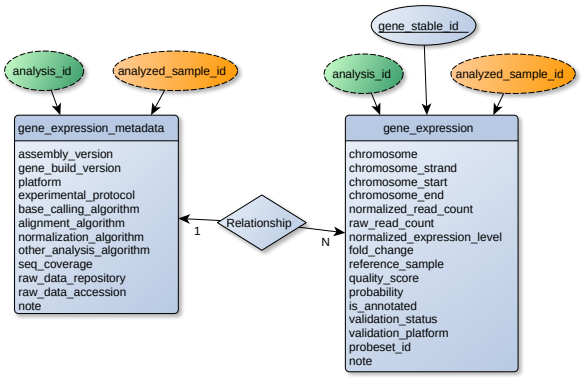

Figure 2.3: Gene Expression Sub-Schema

2.3.1 Expression - Metadata File

Expression [exp] – Metadata File [m]

Table 2.7: Expression - Metadata File

Name	Type	R/O	Description / Values
analysis_id	`string`	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	`string`	R	Unique identifier for the analyzed sample
alignment_algorithm	`compound`	R	Name of alignment algorithm and URL to written protocol
assembly_version	`integer`	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	`compound`	R	Name of base calling algorithm and URL to written protocol

Continued on next page

Table 2.7 – concluded from previous page

Name	Type	R/O	Description / Values
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms

2.3.2 Expression - Gene File

Expression [exp] – Gene File [g]

Table 2.8: Expression - Gene File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
gene_stable_id	string	R	For annotated gene, use Ensembl gene ID. Otherwise, use assemblyBuild_chr_start_end where assemblyBuild is hg18 or hg19.
chromosome	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
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

Continued on next page

Table 2.8 – concluded from previous page

Name	Type	R/O	Description / Values
chromosome_strand	integer	O	Strand where it was detected the mutation/variation on the chromosome **-1** = -1 **1** = 1
fold_change	decimal	O	Expressed fold change if differential expression is measured
is_annotated	boolean	R	Indicate if the expressed fragment is annotated in Ensembl
normalized_expression_level	decimal	O	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	O	Optional field to leave notes
probability	decimal	O	Probability of the mutation/variation call
probeset_id	string	O	ID of the probeset used in microarray if analyzed by microarray platform
quality_score	decimal	O	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	O	ID of the reference analyzed sample if differential expression is measured
validation_platform	integer	O	Platform or technology used in validation *(See CV Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.4 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:

2.4.1 Exon Junction - Metadata File

Exon Junction [jcn] – Metadata File [m]

Table 2.9: Exon Junction - Metadata File

Name	Type	R/O	Description / Values
Continued on next page			

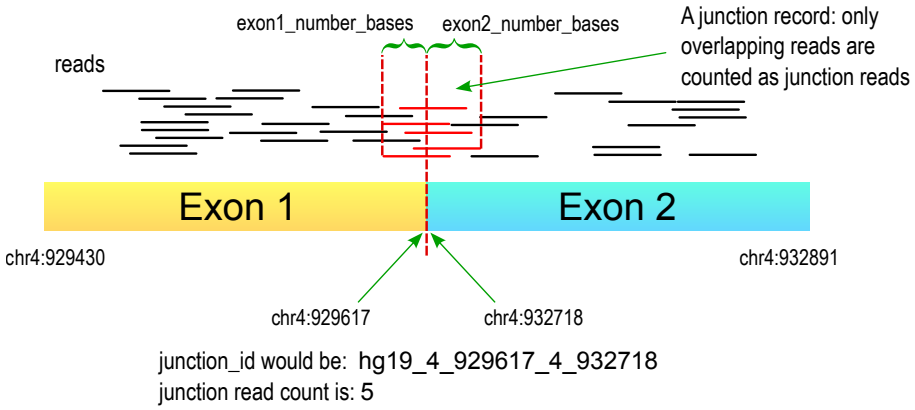

Figure 2.4: Junction Read Count explanation

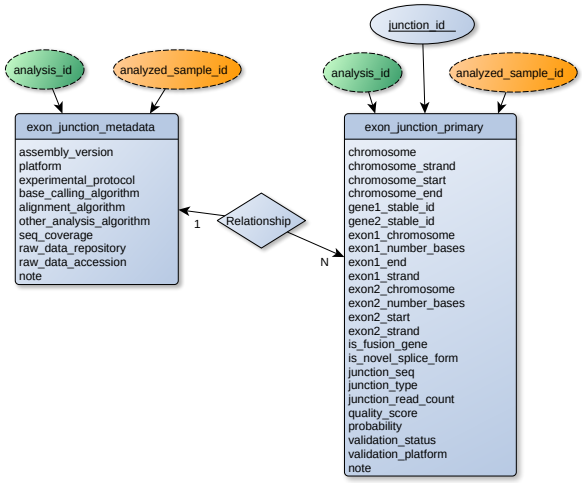

Figure 2.5: Exon Junction Sub-Schema

Table 2.9 – continued from previous page

Name	Type	R/O	Description / Values
analysis_id	`string`	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	`string`	R	Unique identifier for the analyzed sample
alignment_algorithm	`compound`	R	Name of alignment algorithm and URL to written protocol
assembly_version	`integer`	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	`compound`	R	Name of base calling algorithm and URL to written protocol

Continued on next page

Table 2.9 – concluded from previous page

Name	Type	R/O	Description / Values
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms

2.4.2 Exon Junction - Primary Analysis File

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

Table 2.10: Exon Junction - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
junction_id	string	R	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 2.10 – continued from previous page

Name	Type	R/O	Description / Values
chromosome	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
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** = -1 **1** = 1
exon1_chromosome	integer	R	Name of the chromosome containing the 5' exon (#) *(See CV Table A.3)*
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** = -1 **1** = 1
exon2_chromosome	integer	R	Name of the chromosome containing the 3' exon (#) *(See CV Table A.3)*
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** = -1 **1** = 1
gene1_stable_id	string	R	Stable ID of the gene containing the 5' exon at the junction. For 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 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 2.10 – concluded from previous page

Name	Type	R/O	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 Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.5 Protein-DNA interactions

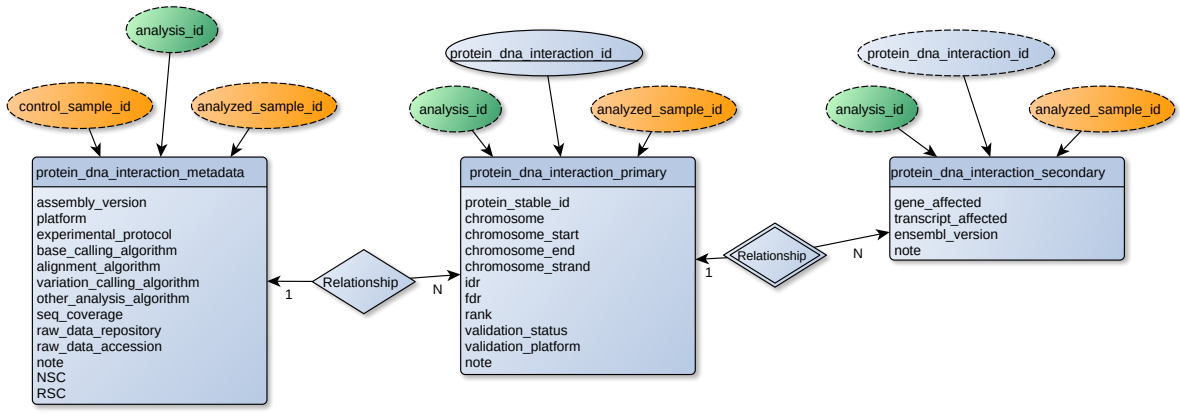

Figure 2.6: Protein-DNA interactions Sub-Schema

2.5.1 Protein-DNA interaction - Metadata File

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

Table 2.11: Protein-DNA interaction - Metadata File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	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	R	Name of alignment algorithm and URL to written protocol
assembly_version	integer	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	compound	R	Name of base calling algorithm and URL to written protocol
control_sample_id	string	R	Unique identifier for the analyzed control/matched sample
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms

2.5.2 Protein-DNA interaction - Primary Analysis File

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

Table 2.12: Protein-DNA interaction - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples

Continued on next page

Table 2.12 – concluded from previous page

Name	Type	R/O	Description / Values
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
protein_dna_interaction_id	string	R	Unique identifier for the protein-DNA interaction
chromosome	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
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** = -1 **1** = 1
fdr	decimal	R	False discovery rate
idr	decimal	R	Irreproducible discovery rate
note	string	0	Optional field to leave notes
protein_stable_id	string	R	Stable id of the interacting protein, antibody or protein complex
rank	compound	0	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	0	Platform or technology used in validation *(See CV Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.5.3 Protein-DNA interaction - Secondary Analysis File

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

Table 2.13: Protein-DNA interaction - Secondary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples

Continued on next page

Table 2.13 – concluded from previous page

Name	Type	R/O	Description / Values
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
ensembl_version	integer	R	Version of Ensembl gene build used for annotation
gene_affected	string	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
note	string	O	Optional field to leave notes
protein_dna_interaction_id	string	R	Unique identifier for the protein-DNA interaction
transcript_affected	string	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

2.6 Regulatory Regions

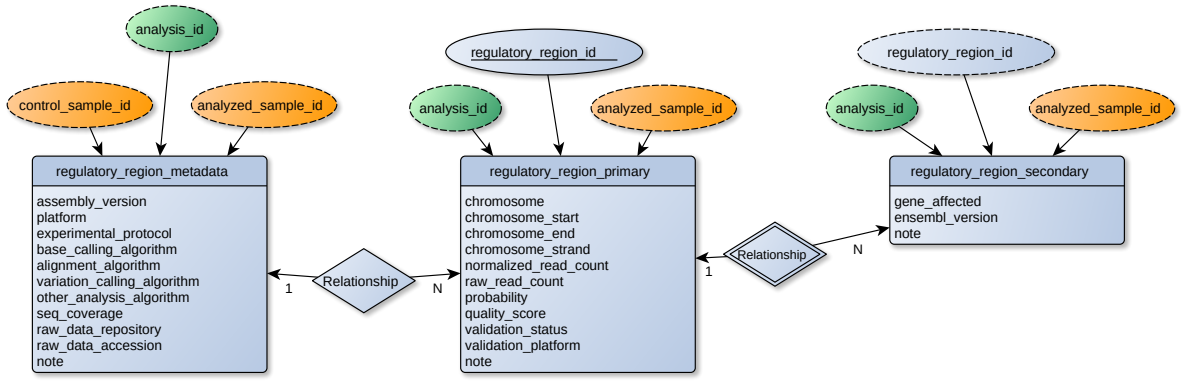

Figure 2.7: Regulatory Regions Sub-Schema

2.6.1 Regulatory regions - Metadata File

Regulatory regions [rreg] – Metadata File [m]

Table 2.14: Regulatory regions - Metadata File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples

Continued on next page

Table 2.14 – concluded from previous page

Name	Type	R/O	Description / Values
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
alignment_algorithm	compound	R	Name of alignment algorithm and URL to written protocol
assembly_version	integer	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	compound	R	Name of base calling algorithm and URL to written protocol
control_sample_id	string	R	Unique identifier for the analyzed control/matched sample
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms

2.6.2 Regulatory regions - Primary Analysis File

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

Table 2.15: Regulatory regions - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
regulatory_region_id	string	R	Unique identifier for the identified regulatory region

Continued on next page

Table 2.15 – concluded from previous page

Name	Type	R/O	Description / Values
chromosome	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
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** = -1 **1** = 1
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 Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.6.3 Regulatory regions - Secondary Analysis File

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

Table 2.16: Regulatory regions - Secondary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
ensembl_version	integer	R	Version of Ensembl gene build used for annotation

Continued on next page

Table 2.16 – concluded from previous page

Name	Type	R/O	Description / Values
gene_affected	string	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
note	string	O	Optional field to leave notes
regulatory_region_id	string	R	Unique identifier for the identified regulatory region

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

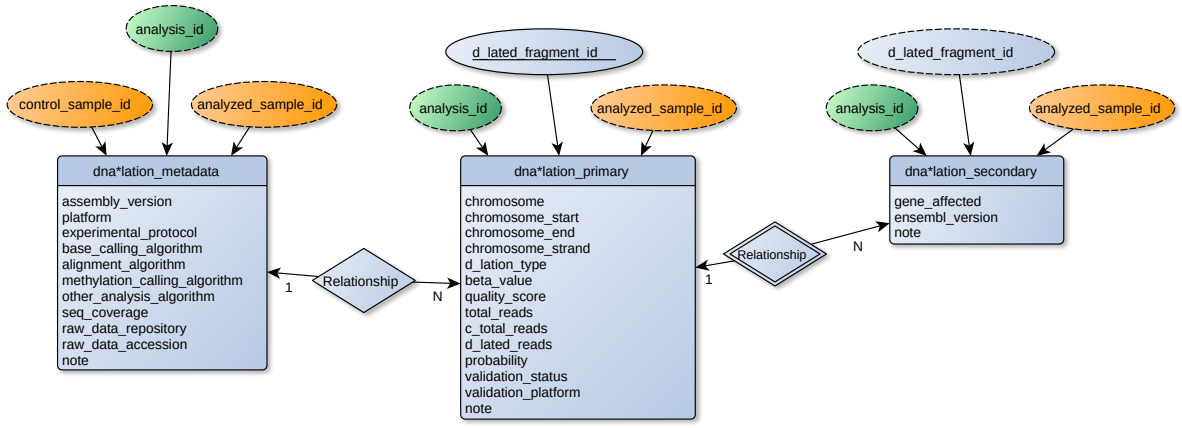

Figure 2.8: DNA Methylation , Hydroxy-Methylation, Formylation, etc... Sub-Schema

2.7.1 DNA *-lation - Metadata File

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

Table 2.17: DNA *-lation - Metadata File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
alignment_algorithm	compound	R	Name of alignment algorithm and URL to written protocol

Continued on next page

Table 2.17 – concluded from previous page

Name	Type	R/O	Description / Values
assembly_version	integer	R	Version of reference genome assembly *(See CV Table A.4)*
base_calling_algorithm	compound	R	Name of base calling algorithm and URL to written protocol
control_sample_id	string	R	Unique identifier for the analyzed control/matched sample
experimental_protocol	compound	O	Name of experimental protocol and URL to written protocol
methylation_calling_algorithm	compound	O	Name of variation calling algorithm and URL to written protocol
note	string	O	Optional field to leave notes
other_analysis_algorithm	compound	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
platform	integer	R	Platform or technology used in the detection phase *(See CV Table A.2)*
raw_data_accession	compound	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 Table A.5)*
seq_coverage	decimal	O	Sequence coverage if analyzed by sequencing platforms

2.7.2 DNA *-lation - Primary Analysis File

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

Table 2.18: DNA *-lation - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
d_lated_fragment_id	string	R	Unique identifier for the methylated fragment, in the form d'*lationType|chromosome_chromosomeStart_chromosomeEnd

Continued on next page

Table 2.18 – concluded from previous page

Name	Type	R/O	Description / Values
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	integer	R	Name of the chromosome containing the experimentally detected feature (mutation, variation, expression, ...) *(See CV Table A.3)*
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** = -1 **1** = 1
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
validation_platform	integer	O	Platform or technology used in validation *(See CV Table A.2)*
validation_status	integer	R	Indicate if the mutation/variation has been validated **-1** = Not valid **0** = Not tested **1** = Validated

2.7.3 DNA *-lation - Secondary Analysis File

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

Table 2.19: DNA *-lation - Secondary Analysis File

Name	Type	R/O	Description / Values
analysis_id	string	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	string	R	Unique identifier for the analyzed sample
d_lated_fragment_id	string	R	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
gene_affected	string	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
note	string	O	Optional field to leave notes

Appendix A

Controlled Vocabulary Tables

A.1 Institution ID

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

Table A.1: Institution ID

ID	Institution
001	Spanish National Cancer Research Centre (CNIO, Madrid)
002	Barcelona Supercomputing Center (BSC-CNS, Madrid)
003	EMBL-EBI (Hinxton)

A.2 Value Codes for Platform or Validation Platform

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

Table A.2: 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.2 – 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.2 – 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.3 Chromosome Names for Reference Genome GRCh37

Table A.3: 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.3 – 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
23	X
24	Y
25	MT
26	c5_H2
27	c6_COX
28	c6_QBL
29	NT_113870
30	NT_113871
31	NT_113872
32	NT_113874
33	NT_113878
34	NT_113880
35	NT_113881
36	NT_113884
37	NT_113885
38	NT_113886
39	NT_113888
40	NT_113889
41	NT_113890
42	NT_113898
43	NT_113899
44	NT_113901
45	NT_113902
46	NT_113903
47	NT_113906
48	NT_113908
49	NT_113909
50	NT_113910
51	NT_113911
52	NT_113912
53	NT_113915
54	NT_113916
55	NT_113917

Continued on next page

Table A.3 – continued from previous page

Key	Chromosome Name
56	NT_113923
57	NT_113924
58	NT_113925
59	NT_113926
60	NT_113927
61	NT_113929
62	NT_113930
63	NT_113931
64	NT_113932
65	NT_113933
66	NT_113934
67	NT_113935
68	NT_113936
69	NT_113937
70	NT_113939
71	NT_113943
72	NT_113944
73	NT_113946
74	NT_113949
75	NT_113951
76	NT_113953
77	NT_113954
78	NT_113956
79	NT_113957
80	NT_113958
81	NT_113960
82	NT_113961
83	NT_113962
84	NT_113963
85	NT_113964
86	NT_113965
87	NT_113966
88	HSCHR17_1
89	HSCHR17_RANDOM_CTG2
90	HSCHR17_RANDOM_CTG3
91	HSCHR19_RANDOM_CTG2
92	HSCHR1_RANDOM_CTG12
93	HSCHR1_RANDOM_CTG5
94	HSCHR4_RANDOM_CTG2
95	HSCHR4_RANDOM_CTG3
96	HSCHR6_MHC_APD
97	HSCHR6_MHC_COX

Continued on next page

Table A.3 – concluded from previous page

Key	Chromosome Name
98	HSCHR6_MHC_DBB
99	HSCHR6_MHC_MANN
100	HSCHR6_MHC_MCF
101	HSCHR6_MHC_QBL
102	HSCHR6_MHC_SSTO
103	HSCHR7_RANDOM_CTG1
104	HSCHR8_RANDOM_CTG1
105	HSCHR8_RANDOM_CTG4
106	HSCHR9_RANDOM_CTG2
107	HSCHR9_RANDOM_CTG4
108	HSCHR9_RANDOM_CTG5
109	HSCHRUN_RANDOM_CTG1
110	HSCHRUN_RANDOM_CTG10
111	HSCHRUN_RANDOM_CTG11
112	HSCHRUN_RANDOM_CTG13
113	HSCHRUN_RANDOM_CTG14
114	HSCHRUN_RANDOM_CTG15
115	HSCHRUN_RANDOM_CTG16
116	HSCHRUN_RANDOM_CTG17
117	HSCHRUN_RANDOM_CTG2
118	HSCHRUN_RANDOM_CTG20
119	HSCHRUN_RANDOM_CTG21
120	HSCHRUN_RANDOM_CTG22
121	HSCHRUN_RANDOM_CTG23
122	HSCHRUN_RANDOM_CTG26
123	HSCHRUN_RANDOM_CTG29
124	HSCHRUN_RANDOM_CTG3
125	HSCHRUN_RANDOM_CTG30
126	HSCHRUN_RANDOM_CTG31
127	HSCHRUN_RANDOM_CTG32
128	HSCHRUN_RANDOM_CTG33
129	HSCHRUN_RANDOM_CTG34
130	HSCHRUN_RANDOM_CTG35
131	HSCHRUN_RANDOM_CTG36
132	HSCHRUN_RANDOM_CTG4
133	HSCHRUN_RANDOM_CTG40
134	HSCHRUN_RANDOM_CTG5
135	HSCHRUN_RANDOM_CTG6
136	HSCHRUN_RANDOM_CTG9
137	HSCHR4_1

A.4 Value Codes for Reference Genome Assembly Version

Table A.4: Value Codes for Reference Genome Assembly Version

Key	Reference Genome Assembly Version
	Continued on next page

Table A.4 – 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.5 Value Codes for Raw Data Repository

Table A.5: Value Codes for Raw Data Repository

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