BLUEPRINT proposed data submission schemas
Draft v0.1.3

José María Fernández

September 17, 2012

Contents

Chapter 1

Tabular format of input files

1.1 Gene Expression

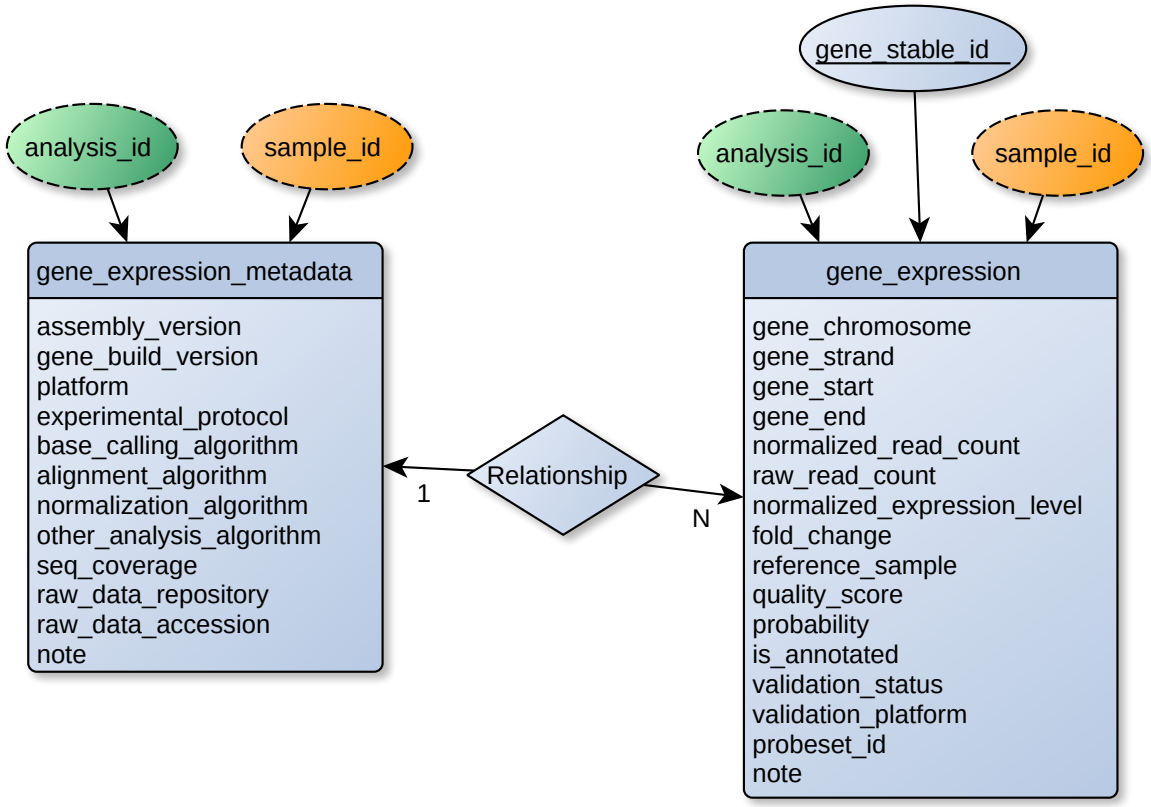

Figure 1.1: Test

1.1.1 Expression - Metadata File

Expression [exp] – Metadata File [m]

Table 1.1: Expression - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples
sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample
assembly_version	VARCHAR(64)	R	Version of reference genome assembly *(See CV Table A.1)*
gene_build_version	INTEGER	R	Version of Ensembl gene build used for annotation
platform	VARCHAR(512)	R	Platform or technology used in detecting the expression *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
normalization_algorithm	VARCHAR(512)	R	Name of normalization algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
seq_coverage	FLOAT(5,2)	O	Sequence coverage if analyzed by sequencing platforms
raw_data_repository	VARCHAR(128)	O	Public repository where raw data is submitted (#) *(See CV Table A.3)*
raw_data_accession	VARCHAR(128)	O	Accession and URL for referencing the raw data at the public repository
note	TEXT	O	Optional field to leave notes

1.1.2 Expression - Gene File

Expression [exp] – Gene File [g]

Table 1.2: Expression - Gene File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples

Continued on next page

Table 1.2 – continued from previous page

Name	Type	R/O	Description / Values
sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample
gene_stable_id	VARCHAR(64)	R	For annotated gene, use Ensembl gene ID. Otherwise, use assemblyBuild_chr_start_end where assemblyBuild is hg18 or hg19.
gene_chromosome	VARCHAR(64)	R	Name of the chromosome containing the expressed gene/region interrogated *(See CV Table A.4)*
gene_strand	INTEGER	R	Strand of the chromosome containing the expressed gene/region **1** = 1 **-1** = -1
gene_start	INTEGER	R	Start position of the gene on the chromosome
gene_end	INTEGER	R	End position of the gene on the chromosome
normalized_read_count	FLOAT(5,2)	R	Normalized count of sequencing reads if analyzed by sequencing platforms
raw_read_count	INTEGER	R	Raw count of sequencing reads if analyzed by sequencing platforms
normalized_expression_level	FLOAT(5,2)	O	Normalized value of expression level if analyzed by microarray platforms
fold_change	FLOAT(5,2)	O	Expressed fold change if differential expression is measured
reference_sample	VARCHAR(64)	O	ID of the reference analyzed sample if differential expression is measured
quality_score	INTEGER	O	Quality score for the expression call
probability	FLOAT(3,2)	O	Probability of the expression call
is_annotated	VARCHAR(64)	O	Indicate if the expressed fragment is annotated in Ensembl **1** = annotated **2** = not annotated
validation_status	VARCHAR(64)	R	Indicate if the expressed fragment has been validated **1** = validated **2** = not tested **3** = not valid

Continued on next page

Table 1.2 – concluded from previous page

Name	Type	R/O	Description / Values
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
probeset_id	VARCHAR(128)	O	ID of the probeset used in microarray
note	TEXT	O	Optional field to leave notes

1.2 Copy Number Germline Variations

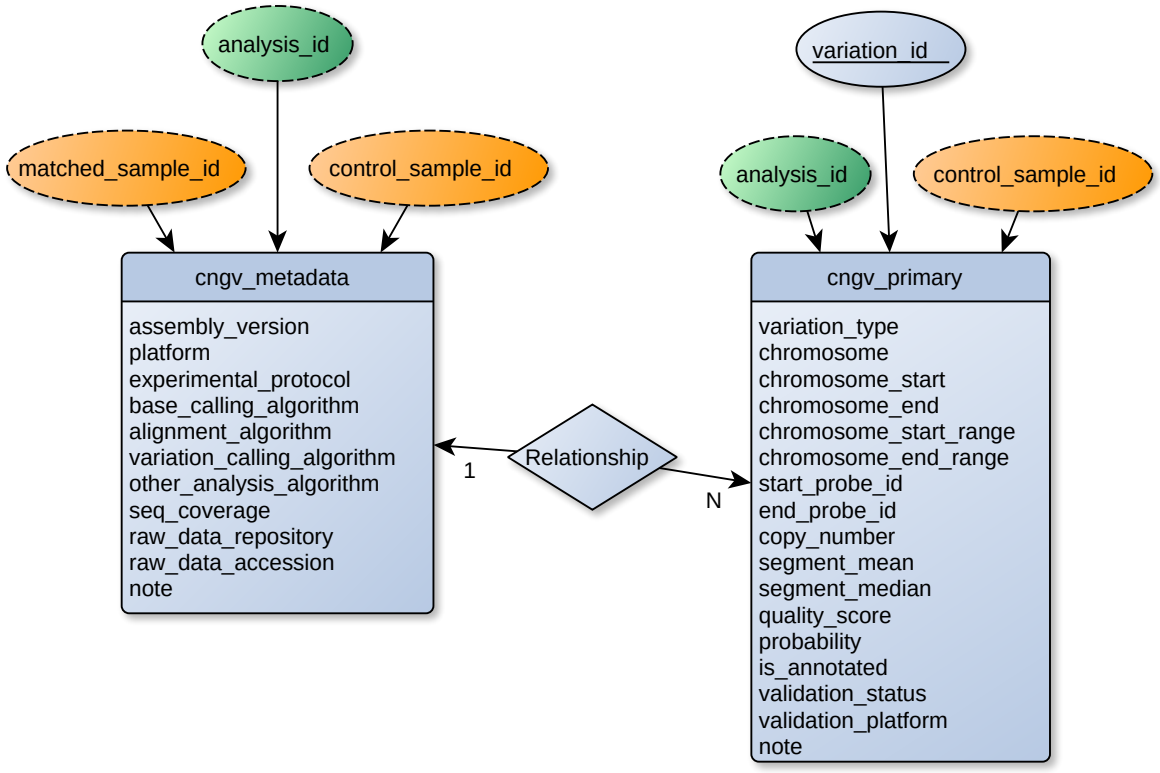

Figure 1.2: Test

1.2.1 Simple Germline Variations - Metadata File

Simple Germline Variations [sgv] – Metadata File [m]

Table 1.3: Simple Germline Variations - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular group of samples

Continued on next page

Table 1.3 – concluded from previous page

Name	Type	R/O	Description / Values
control_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed control sample
matched_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed matched sample
assembly_version	VARCHAR(64)	R	Version of reference genome assembly *(See CV Table A.1)*
platform	VARCHAR(512)	R	Platform or technology used in detecting the mutation/variation *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
variation_calling_algorithm	VARCHAR(512)	R	Name of variation calling algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
seq_coverage	FLOAT(5,2)	O	Sequence coverage if analyzed by sequencing platforms
raw_data_repository	VARCHAR(512)	O	Public repository where raw data is submitted (#) *(See CV Table A.3)*
raw_data_accession	VARCHAR(512)	O	Accession and URL for referencing the raw data at the public repository
note	TEXT	O	Optional field to leave notes

1.2.2 Simple Germline Variations - Primary Analysis File

Simple Germline Variations [sgv] – Primary Analysis File [p]

Table 1.4: Simple Germline Variations - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	TEXT	R	Unique identifier for the analysis performed for a particular set of samples
control_sample_id	TEXT	R	Unique identifier for the analyzed control sample

Continued on next page

Table 1.4 – continued from previous page

Name	Type	R/O	Description / Values
variation_id	VARCHAR(128)	R	Unique identifier for the variation
variation_type	VARCHAR(64)	R	Type of variation **1** = single base substitution **2** = insertion of <=200bp **3** = deletion of <=200bp **4** = multiple base substitution (>=2bp and <=200bp)
chromosome	VARCHAR(64)	R	Name of the chromosome containing the mutation/variation *(See CV Table A.4)*
chromosome_start	INTEGER	R	Start position of the mutation/variation on the chromosome
chromosome_end	INTEGER	R	End position of the mutation/variation on the chromosome
chromosome_strand	INTEGER	R	Chromosome strand **1** = 1 **-1** = -1
refsnp_allele	VARCHAR(512)	R	RefSNP alleles from dbSNP (use a dash for each missing base) *e.g.: A/T, —/AAA*
refsnp_strand	INTEGER	O	Strand of RefSNP allele **1** = 1 **-1** = -1
reference_genome_allele	VARCHAR(512)	R	Allele in the reference genome (use a dash for each missing base)
control_genotype	VARCHAR(512)	R	Genotype of the control sample (use a dash for each missing base)
tumour_genotype	VARCHAR(512)	R	Genotype of the tumour sample (use a dash for each missing base)
expressed_allele	VARCHAR(512)	O	The expressed allele(s) as revealed by RNA-seq, etc.
quality_score	INTEGER	O	Average quality score for the mutation/variation call
probability	FLOAT(3,2)	O	Probability of the mutation/variation call
read_count	FLOAT(5,2)	O	Average number of times the bases are covered by raw reads
is_annotated	VARCHAR(64)	O	Indicate if the mutation/variation is annotated in dbSNP **1** = annotated **2** = not annotated

Continued on next page

Table 1.4 – concluded from previous page

Name	Type	R/O	Description / Values
validation_status	VARCHAR(64)	R	Indicate if the mutation/variation has been validated **1** = validated **2** = not tested **3** = not valid
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
xref_ensembl_var_id	VARCHAR(128)	O	Cross-reference: Ensembl Variation ID in Ensembl Variation database. *e.g.: rs12345; ENSSNP53189*
note	TEXT	O	Optional field to leave notes

1.3 Exon Junction

The following diagram (from ICGC DCC manual) illustrates how junction_id is assigned, how junction_read_count, exon1_number_bases and exon2_number_bases are calculated:

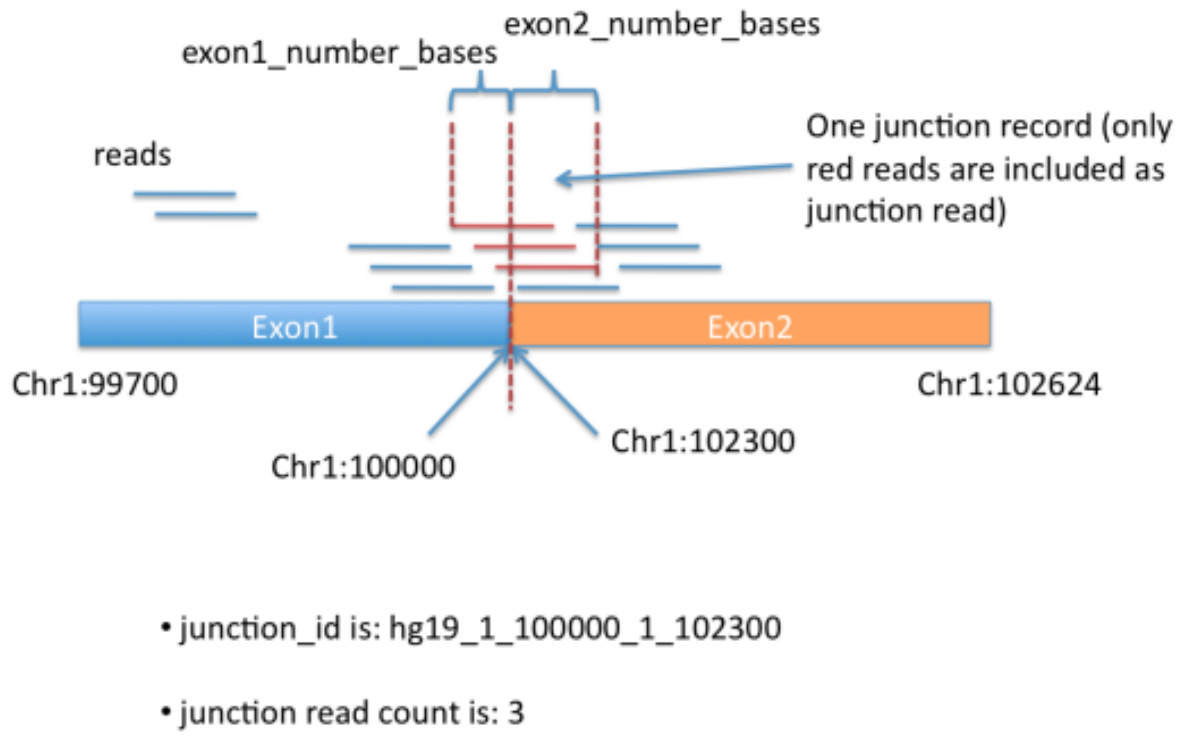

Figure 1.3: Junction Read Count explanation

1.3.1 Exon Junction - Metadata File

Exon Junction [jcn] – Metadata File [m]

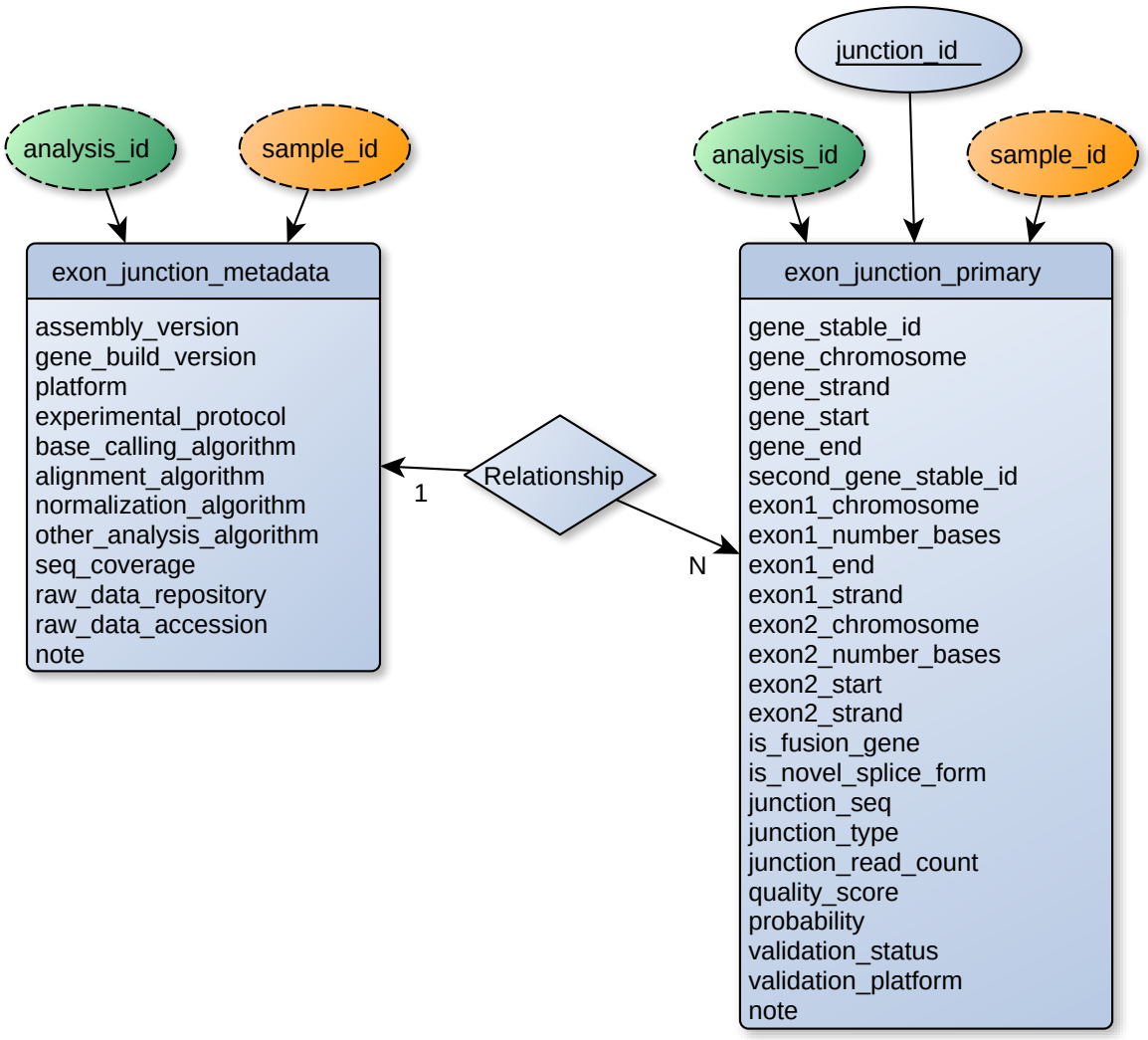

Figure 1.4: Test

Table 1.5: Exon Junction - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples
sample_id	TEXT	R	Unique identifier for the analyzed sample

Continued on next page

Table 1.5 – concluded from previous page

Name	Type	R/O	Description / Values
assembly_version	VARCHAR(64)	R	Version of reference genome assembly (#) *(See CV Table A.1)*
gene_build_version	INTEGER	R	Version of Ensembl gene build used for annotation
platform	VARCHAR(512)	R	Platform or technology used in detecting the expression *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
normalization_algorithm	VARCHAR(512)	R	Name of normalization algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
seq_coverage	FLOAT(5,2)	O	Sequence coverage if analyzed by sequencing platforms **1** = EGA **2** = dbSNP
raw_data_repository	VARCHAR(128)	R	Public repository where raw data is submitted (#) *(See CV Table A.3)*
raw_data_accession	VARCHAR(128)	R	Accession and URL for referencing the raw data at the public repository
note	TEXT	O	Optional field to leave notes

1.3.2 Exon Junction - Primary Analysis File

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

Table 1.6: Exon Junction - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular group of samples
sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample
junction_id	VARCHAR(256)	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.
gene_stable_id	VARCHAR(64)	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.
gene_chromosome	VARCHAR(64)	R	Name of the chromosome containing the above gene. *(See CV Table A.4)*
gene_strand	INTEGER	R	Strand of the chromosome **1** = 1 **-1** = -1
gene_start	INTEGER	R	Start position of the entire gene on the chromosome as annotated in Ensembl
gene_end	INTEGER	R	End position of the entire gene on the chromosome as annotated in Ensembl
second_gene_stable_id	VARCHAR(64)	O	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.
exon1_chromosome	VARCHAR(64)	R	Name of the chromosome containing the 5' exon (#) *(See CV Table A.4)*
exon1_number_bases	INTEGER	R	Number of bases from 5' exon
exon1_end	INTEGER	R	End position of the 5' exon on the chromosome

Continued on next page

Table 1.6 – concluded from previous page

Name	Type	R/O	Description / Values
exon1_strand	INTEGER	O	Chromosome strand of the 5' exon **1** = 1 **-1** = -1
exon2_chromosome	VARCHAR(64)	R	Name of the chromosome containing the 3' exon (#) *(See CV Table A.4)*
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	O	Chromsome strand of the 3' exon **1** = 1 **-1** = -1
is_fusion_gene	VARCHAR(16)	O	Indicate if the function is the result of a fusion gene **1** = yes **2** = no
is_novel_splice_form	VARCHAR(16)	O	Indicate if the splice form is novel **1** = yes **2** = no
junction_seq	TEXT	O	Provide junction sequence if either is_fusion_gene or is_novel_splice_form is true
junction_type	VARCHAR(64)	O	Type of junction **1** = canonical **2** = non-canonical **3** = U12
junction_read_count	FLOAT(5,2)	R	Count of sequencing reads that span across exons
quality_score	INTEGER	O	Quality score for the junction call
probability	FLOAT(3,2)	O	Probability of the junction call
validation_status	VARCHAR(64)	R	Indicate if the junction has been validated **1** = validated **2** = not tested **3** = not valid
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
note	TEXT	O	Optional field to leave notes

1.4 DNA Methylation and Hydroxy-Methylation

1.4.1 Methylation - Secondary Analysis File

Methylation [meth] – Secondary Analysis File [s]

Table 1.7: Methylation - Secondary Analysis File

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

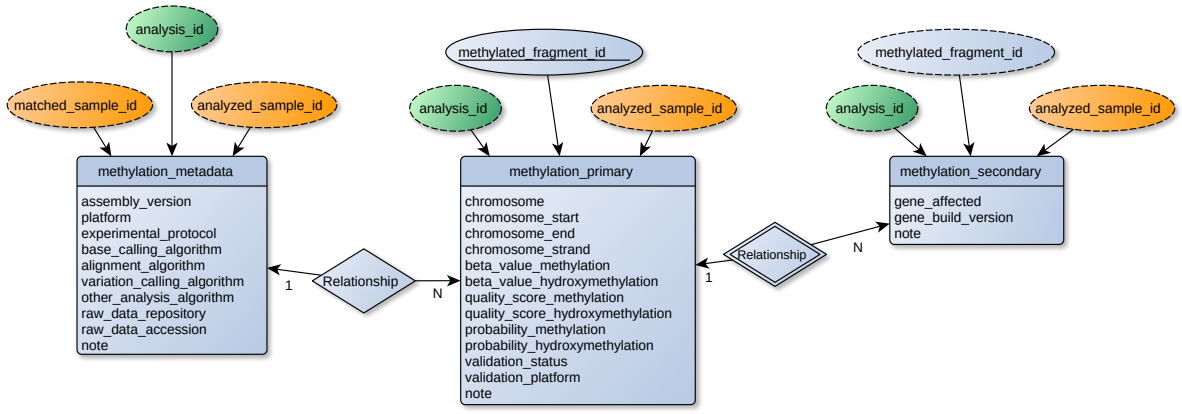

Figure 1.5: Test

Table 1.7 – concluded from previous page

Name	Type	R/O	Description / Values
analysis_id	TEXT	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
methylated_fragment_id	TEXT	R	Unique identifier for the methylation
gene_affected	VARCHAR(128)	R	Gene(s) containing the methylation. Use Ensembl gene id. Separate multiple genes with vertical bars in the form of geneA—geneB—geneC. If no gene is affected, use -888 (not applicable).
gene_build_version	INTEGER	R	Version of Ensembl gene build used for annotation
note	TEXT	O	Optional field to leave notes

1.4.2 Methylation - Metadata File

Methylation [meth] – Metadata File [m]

Table 1.8: Methylation - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample

Continued on next page

Table 1.8 – concluded from previous page

Name	Type	R/O	Description / Values
matched_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed control sample
assembly_version	VARCHAR(64)	R	Version of reference genome assembly *(See CV Table A.1)*
platform	VARCHAR(512)	R	Platform or technology used in detecting the methylation *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
variation_calling_algorithm	VARCHAR(512)	R	Name of variation calling algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
raw_data_repository	VARCHAR(128)	O	Public repository where raw data is submitted *(See CV Table A.3)*
raw_data_accession	VARCHAR(128)	O	Accession and URL for referencing the raw data at the public repository
note	TEXT	O	Optional field to leave notes

1.4.3 Methylation - Primary Analysis File

Methylation [meth] – Primary Analysis File [p]

Table 1.9: Methylation - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	TEXT	R	Unique identifier for the analysis performed for a particular group of samples
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
methylated_fragment_id	VARCHAR(128)	R	Unique identifier for the methylated fragment
chromosome	VARCHAR(64)	R	Name of the chromosome containing the methylation *(See CV Table A.4)*

Continued on next page

Table 1.9 – concluded from previous page

Name	Type	R/O	Description / Values
chromosome_start	INTEGER	R	Start position of the methylation on the chromosome
chromosome_end	INTEGER	R	End position of the methylation on the chromosome
chromosome_strand	INTEGER	O	Chromosome strand **1** = 1 **-1** = -1
beta_value_methylation	FLOAT(5,2)	O	Methylation Beta value for interrogated site
beta_value_hydroxymethylation	FLOAT(5,2)	O	Hydroxymethylation Beta value for interrogated site
quality_score_methylation	INTEGER	O	Quality score for the methylation call
quality_score_hydroxymethylation	INTEGER	O	Quality score for the hydroxymethylation call
probability_methylation	FLOAT(3,2)	O	Probability of the methylation call
probability_hydroxymethylation	FLOAT(3,2)	O	Probability of the hydroxymethylation call
validation_status	VARCHAR(64)	R	Indicate if the methylation has been validated **1** = validated **2** = not tested **3** = not valid
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
note	TEXT	O	Optional field to leave notes

1.5 Protein-DNA interactions

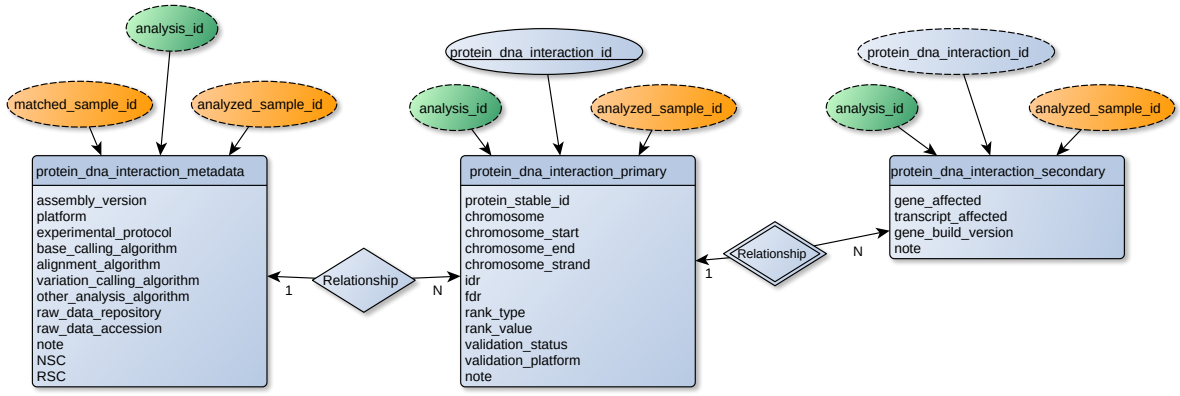

Figure 1.6: Test

1.5.1 Protein-DNA interaction - Metadata File

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

Table 1.10: Protein-DNA interaction - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample
matched_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed control sample
assembly_version	VARCHAR(64)	R	Version of reference genome assembly *(See CV Table A.1)*
platform	VARCHAR(512)	R	Platform or technology used in detecting the protein-DNA interaction *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
variation_calling_algorithm	VARCHAR(512)	R	Name of variation calling algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
raw_data_repository	VARCHAR(128)	O	Public repository where raw data is submitted *(See CV Table A.3)*
raw_data_accession	VARCHAR(128)	O	Accession and URL for referencing the raw data at the public repository
NSC	FLOAT(5,2)	O	Normalized strand-cross correlation of the analysis
RSC	FLOAT(5,2)	O	Relative strand-cross correlation of the analysis
note	TEXT	O	Optional field to leave notes

1.5.2 Protein-DNA interaction - Primary Analysis File

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

Table 1.11: Protein-DNA interaction - Primary Analysis File

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

Continued on next page

Table 1.11 – concluded from previous page

Name	Type	R/O	Description / Values
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
protein_dna_interaction_id	VARCHAR(128)	R	Unique identifier for the protein-DNA interaction
protein_stable_id	VARCHAR(128)	R	Ensembl protein stable id of the interacting protein
chromosome	VARCHAR(64)	R	Name of the chromosome where the protein-DNA interaction happened *(See CV Table A.4)*
chromosome_start	INTEGER	R	Start position where the interaction happened on the chromosome
chromosome_end	INTEGER	R	End position the interaction happened on the chromosome
chromosome_strand	INTEGER	O	Chromosome strand **1** = 1 **-1** = -1
idr	FLOAT(5,2)	R	Irreproducible discovery rate
fdr	FLOAT(5,2)	O	False discovery rate
rank_type	VARCHAR(64)	O	Kind of used ranking
rank_value	FLOAT(5,2)	O	Rank value
validation_status	VARCHAR(64)	R	Indicate if the detected protein-DNA interaction has been validated **1** = validated **2** = not tested **3** = not valid
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
note	TEXT	O	Optional field to leave notes

1.5.3 Protein-DNA interaction - Secondary Analysis File

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

Table 1.12: Protein-DNA interaction - Secondary Analysis File

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

Continued on next page

Table 1.12 – concluded from previous page

Name	Type	R/O	Description / Values
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
protein_dna_interaction_id	TEXT	R	Unique identifier for the protein-DNA interaction
gene_affected	VARCHAR(128)	R	Gene on the protein-DNA interaction area. Use Ensembl gene id. If no gene is affected, use -888 (not applicable).
transcript_affected	VARCHAR(128)	R	Transcript on the protein-DNA interaction area. Use Ensembl transcript id. Separate multiple transcripts with vertical bars in the form of trasncriptA—trasncriptB—trasncriptC. If no trasncript is affected, use -888 (not applicable).
gene_build_version	INTEGER	R	Version of Ensembl gene build used for annotation
note	TEXT	O	Optional field to leave notes

1.6 Clinical Data Submission File Specifications

Overview

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

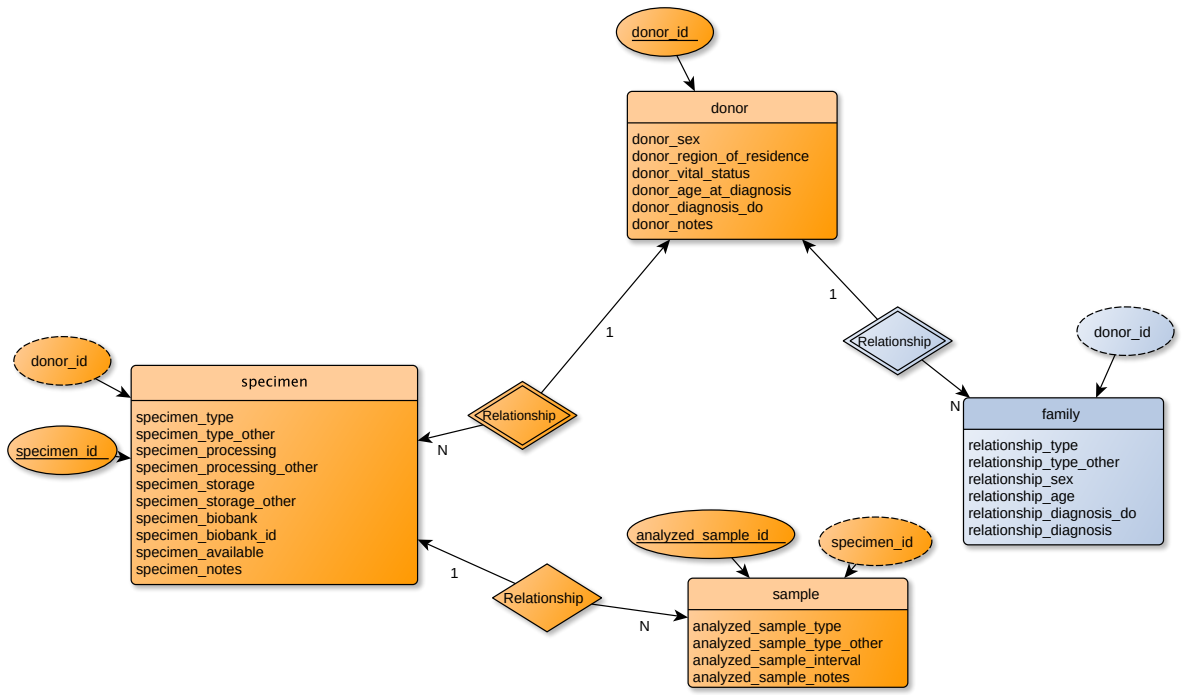

Figure 1.7: Test

1.6.1 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 1.13: Analyzed Sample Data File

Name	Type	R/O	Description / Values
analyzed_sample_id	VARCHAR(64)	R	*Unique identifier* for the sample assigned by data provider
specimen_id	VARCHAR(64)	R	*Unique identifier* for the specimen assigned by data provider. The corresponding specimen id **must** appear in the specimen data submission file
analyzed_sample_type	VARCHAR(128)	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)
analyzed_sample_type_other	VARCHAR(64)	O	Free text description of site of sample if "other" was specified in *sample_type* field
analyzed_sample_interval	INTEGER	O	Interval from specimen acquisition to sample use in an analytic procedure (e.g. DNA extraction), in days
analyzed_sample_notes	TEXT	O	Freetext notes about sample allowed

1.6.2 Donor Data File

Donor Data File [donor] (required)

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

Table 1.14: Donor Data File

Name	Type	R/O	Description / Values
donor_id	VARCHAR(64)	R	*Unique identifier* for the donor; assigned by data provider.

Continued on next page

Table 1.14 – concluded from previous page

Name	Type	R/O	Description / Values
donor_sex	VARCHAR(128)	R	Donor biological sex. *"Other" has been removed from the controlled vocabulary due to identifiability concerns.* **1** = male **2** = female
donor_region_of_residence	VARCHAR(64)	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_vital_status	VARCHAR(128)	R	Donor's last known vital status **1** = alive **2** = deceased
donor_age_at_diagnosis	INTEGER	R	Age at primary diagnosis *Use "90" for patients >=90*
donor_diagnosis_do	VARCHAR(64)	R	Disease Ontology code *Disease Ontology code (http://diseaseontology.sourceforge.net/)*
donor_notes	TEXT	O	Free text notes concerning donor *Any additional non-identifying information can be included here.*

1.6.3 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 1.15: Specimen Data File

Name	Type	R/O	Description / Values
donor_id	VARCHAR(64)	R	Unique identifier for the donor; assigned by data provider. It must be coded, and correspond to a donor ID listed in the donor data file.
specimen_id	VARCHAR(64)	R	Unique identifier for the specimen assigned by data provider.
specimen_type	VARCHAR(128)	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** = tumour (other)
specimen_type_other	VARCHAR(64)	R	Free text description of site of specimen if "normal control (other)" or "tumour (other)" was specified in specimen_type field.
specimen_processing	VARCHAR(128)	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 1.15 – concluded from previous page

Name	Type	R/O	Description / Values
specimen_processing_other	VARCHAR(64)	R	If "other" specified for specimen_processing, may indicate technique here.
specimen_storage	VARCHAR(128)	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	VARCHAR(64)	R	If "other" specified for specimen_storage, may indicate technique here.
specimen_biobank	VARCHAR(64)	R	If the specimen was obtained from a biobank, provide the biobank name here
specimen_biobank_id	VARCHAR(64)	R	If the specimen was obtained from a biobank, provide the biobank accession number here.
specimen_available	VARCHAR(128)	R	Whether additional tissue is available for followup studies. **1** = no **2** = yes
specimen_notes	TEXT	O	Free text notes allowed *Any additional non-identifying information can be included here.*

1.6.4 Donor Family History

Donor Family History [family] (optional)

This file describes the family history of the donor.

Table 1.16: Donor Family History

Name	Type	R/O	Description / Values
donor_id	TEXT	R	Unique identifier for the donor; assigned by data provider. It must be coded, and correspond to a donor ID listed in the donor data file.
relationship_type	VARCHAR(128)	R	Relationship to the donor **1** = sibling **2** = parent **3** = grandparent **4** = uncle/aunt **5** = cousin **6** = other

Continued on next page

Table 1.16 – concluded from previous page

Name	Type	R/O	Description / Values
relationship_type_other	TEXT	R	If "other" answered in previous column, specify the relationship type here
relationship_sex	VARCHAR(128)	R	Biological sex of related individual **1** = male **2** = female
relationship_age	INTEGER	R	Age of relative at primary diagnosis (years) *Use 90 for ages >= 90 years.*
relationship_diagnosis_do	TEXT	R	Disease Ontology code for the relative's diagnosis status
relationship_diagnosis	TEXT	R	Diagnosis (disease or healthy status) *e.g. "breast cancer"*

1.7 Regulatory Regions

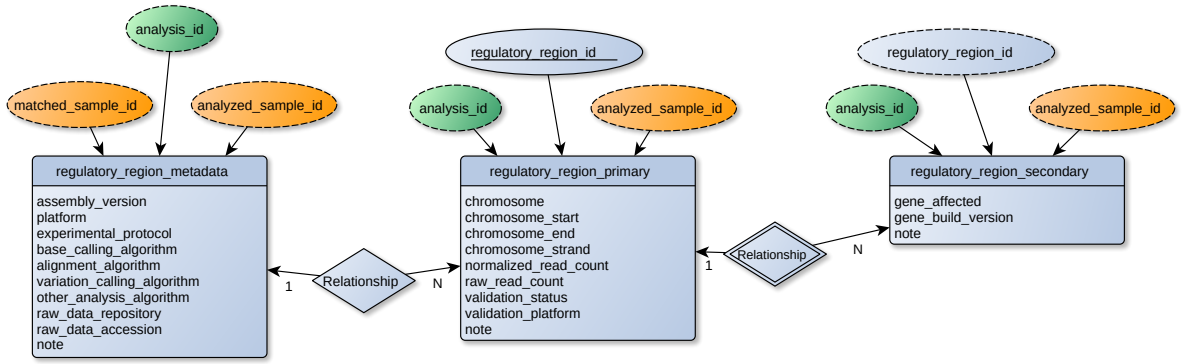

Figure 1.8: Test

1.7.1 Regulatory regions - Primary Analysis File

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

Table 1.17: Regulatory regions - Primary Analysis File

Name	Type	R/O	Description / Values
analysis_id	TEXT	R	Unique identifier for the analysis performed for a particular group of samples
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
regulatory_region_id	VARCHAR(128)	R	Unique identifier for the identified regulatory region

Continued on next page

Table 1.17 – concluded from previous page

Name	Type	R/O	Description / Values
chromosome	VARCHAR(64)	R	Name of the chromosome containing the regulatory region *(See CV Table A.4)*
chromosome_start	INTEGER	R	Start position of the regulatory region on the chromosome
chromosome_end	INTEGER	R	End position of the regulatory region on the chromosome
chromosome_strand	INTEGER	O	Chromosome strand **1** = 1 **-1** = -1
normalized_read_count	FLOAT(5,2)	R	Normalized count of sequencing reads if analyzed by sequencing platforms
raw_read_count	INTEGER	R	Raw count of sequencing reads if analyzed by sequencing platforms
validation_status	VARCHAR(64)	R	Indicate if the regulatory region has been validated **1** = validated **2** = not tested **3** = not valid
validation_platform	VARCHAR(512)	O	Platform or technology used in validation *(See CV Table A.2)*
note	TEXT	O	Optional field to leave notes

1.7.2 Regulatory regions - Secondary Analysis File

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

Table 1.18: Regulatory regions - Secondary Analysis File

Name	Type	R/O	Description / Values
analysis_id	TEXT	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	TEXT	R	Unique identifier for the analyzed sample
regulatory_region_id	TEXT	R	Unique identifier for the identified regulatory region
gene_affected	VARCHAR(128)	R	Gene(s) related to the regulatory region. Use Ensembl gene id. Separate multiple genes with vertical bars in the form of geneA—geneB—geneC. If no gene is affected, use -888 (not applicable).

Continued on next page

Table 1.18 – concluded from previous page

Name	Type	R/O	Description / Values
gene_build_version	INTEGER	R	Version of Ensembl gene build used for annotation
note	TEXT	O	Optional field to leave notes

1.7.3 Regulatory regions - Metadata File

Regulatory regions [rreg] – Metadata File [m]

Table 1.19: Regulatory regions - Metadata File

Name	Type	R/O	Description / Values
analysis_id	VARCHAR(64)	R	Unique identifier for the analysis performed for a particular set of samples
analyzed_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed sample
matched_sample_id	VARCHAR(64)	R	Unique identifier for the analyzed control sample
assembly_version	VARCHAR(64)	R	Version of reference genome assembly *(See CV Table A.1)*
platform	VARCHAR(512)	R	Platform or technology used in detecting the regulatory region *(See CV Table A.2)*
experimental_protocol	VARCHAR(512)	O	Name of experimental protocol and URL to written protocol
base_calling_algorithm	VARCHAR(512)	R	Name of base calling algorithm and URL to written protocol
alignment_algorithm	VARCHAR(512)	R	Name of alignment algorithm and URL to written protocol
variation_calling_algorithm	VARCHAR(512)	R	Name of variation calling algorithm and URL to written protocol
other_analysis_algorithm	VARCHAR(512)	O	Names of other analysis algorithms. Separate multiple algorithms by commas.
raw_data_repository	VARCHAR(128)	O	Public repository where raw data is submitted *(See CV Table A.3)*
raw_data_accession	VARCHAR(128)	O	Accession and URL for referencing the raw data at the public repository

Continued on next page

Table 1.19 – concluded from previous page

Name	Type	R/O	Description / Values
note	TEXT	O	Optional field to leave notes

Appendix A

CV Tables

A.1 CV Table appendix_B10.tsv

Table A.1: Test

Key	Description
1	GRCh37
2	NCBI36
3	GRCh37.p1
4	GRCh37.p2
5	GRCh37.p3
6	GRCh37.p4
7	GRCh37.p5

A.2 CV Table appendix_B5.tsv

Table A.2: Test

Key	Description
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
14	Agilent Whole Human Genome Oligo Microarray Kit

Continued on next page

Table A.2 – continued from previous page

Key	Description
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	Description
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 CV Table appendix_B12.tsv

Table A.3: Test

Key	Description
1	EGA
2	dbSNP
3	TCGA
4	CGHub
5	GEO

A.4 CV Table appendix_B6.tsv

Table A.4: Test

Key	Description
1	1
2	2
3	3

Continued on next page

Table A.4 – continued from previous page

Key	Description
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
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

Continued on next page

Table A.4 – continued from previous page

Key	Description
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
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

Continued on next page

Table A.4 – continued from previous page

Key	Description
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
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

Continued on next page

Table A.4 – concluded from previous page

Key	Description
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