ICGC Data Submission Manual

Document Version 0.5c
May 2011

Table of Contents

1. OVERVIEW OF DATA SUBMISSION PROCESS....2
2. SUBMISSION FILE FORMAT....2
3. SUBMISSION FILE VALIDATION....4
4. FILE SUBMISSION....5
5. SETUP BIOMART SERVER (OPTIONAL)....5
APPENDIX A: DCC DATA ELEMENT SPECIFICATION....7
APPENDIX B: VALUE CODES FOR DES WITH CONTROLLED VOCABULARY....38
APPENDIX C: EGA SEQUENCE DATA SUBMISSION GUIDE....49

1. Overview of data submission process

There are three major steps in the data submission process:

a. Submit raw sequence data to the European Genome-phenome Archive (**Appedix C**).

b. Prepare the ICGC submission files according to DCC specifications (**Section 2**).

c. Verify conformity of the submission files (**Section 3**).

d. Submit files to the DCC Secure FTP server (**Section 4**).

NB. All submitted data must be based on human reference genome assembly GRCh37 and Ensembl gene set version 61.

NB. Please include sample EGA accession in the ICGC Sample Data file (page 36, data element *sample_ega_accession*).

Please contact DCC (dcc-support@lists.oicr.on.ca) if you would like to set up an ICGC node or if you have any questions or comments about the data submission process.

2. Submission file format

- The submission data are kept in tab-limited text files. 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 Data Element specification (**Appendix A**).

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

```
# This is an example of a primary analysis file for simple somatic mutations.
# File name: ssm__ca__01__068__p__8__20090713.txt
#
tumour_sample_id    mutation_id     mutation_type   chromosome  …   note
m124                ssm_3396649     1               1           …   -999
m124                ssm_61023021    1               2           …   -999
m124                ssm_175270973   1               3           …   -999
m124                ssm_72390475    1               4           …   -999
```

- Types of experimental data being supported:
  - simple mutations/variations of ≤ 200 bp
  - copy number mutations/variations
  - structural mutations/variations
  - gene expression

- miRNA
- exon junctions
- DNA methylation

- File Format:

The input file formats are provided below. The files containing experiment results need to comply with the following naming convention (note the use of double underscores ('_') to separate components in the file name):

featureType__leadJurisdiction__tumourType__institution__fileType__platform__dateFileCreated.txt

The components of the file name are listed below:

Components	Description	Values
featureType	Simple somatic mutations including single base substitutions and indels of ≤200 bp	ssm
	Simple germline variations including single base substitutions and indels of ≤200 bp	sgv
	Copy number somatic mutations	cnsm
	Copy number germline variations	cngv
	Structural somatic mutations	stsm
	Structural germline variations	stgv
	Gene and exon expression	exp
	miRNA expression	mirna
	Exon junction	jcn
	Methylation	meth
leadJurisdiction	Jurisdiction leading the project	Appendix Table B1
tumourType	tumour type	Appendix Table B2
institution	Institution submitting the data	Appendix Table B3
fileType	Primary analysis file	p
	Secondary analysis file	s
	Metadata file	m
	Gene expression file	g
	Exon expression file	e
platform	Platform or technology used in the analysis	Appendix Table B5
dateFileCreated	The date on which the file is created	YYYYMMDD

The file names for donor, diagnosis and sample information follow the convention (note the use of double underscores ('_') to separate components in the file name):

leadJurisdiction__tumourType__institution__fileType__dateFileCreated.txt

The components of the file name are listed below:

Components	Description	Values
leadJurisdiction	Jurisdiction leading the project	Appendix Table B1
tumourType	tumour type	Appendix Table B2
institution	Institution submitting the data	Appendix Table B3
fileType	Donor information	donor
	Diagnosis information	diagnosis
	Sample information	sample
dateFileCreated	The date on which the file is created	YYYYMMDD

For examples of the file names see below:

Examples	Description
ssm__ca__01__068__p__8__20090713.txt	In pancreatic cancer project, the primary analysis file generated on July 13, 2009 by OICR (Canada) for simple somatic mutations analyzed on Affymetrix Genome-Wide Human SNP Array 6.0
cngv__ca__01__068__m__8__20090713.txt	In pancreatic cancer project, the metadata file generated on July 13, 2009 by OICR (Canada) for copy number germline variations analyzed on Affymetrix Genome-Wide Human SNP Array 6.0
ca__01__068__donor__20090713.txt	In pancreatic cancer project, donor information provided by OICR (Canada) on July 13, 2009
ca__01__068__sample__20090713.txt	In pancreatic cancer project, sample information provided by OICR (Canada) on July 13, 2009

3. Submission file validation

- For the purpose of validating the submission files, download MartLoader (software tool for processing ICGC data) from DCC's SVN server as below (you may change */home/software* to another local path):

```
cd /home/software
svn co https://code.oicr.on.ca/svn/dcc/martloader/branches/release-0_5_i4  martloader
```

- Create a work directory using a name of your choice (e.g. *workdir_testSept10*) for keeping all submission files:

```
cd /home/software/martloader
perl createWorkDir.pl workdir_testSept10
```

- Put all of the submission files into the appropriate subfolders under *'workdir_testSept10/input'* folder. The subfolders are listed as below:

    a. cnv (copy number variation)
    b. exp (expression)
    c. jcn (exon junction)
    d. meth (methylation)
    e. mirna (microRNA)
    f. sample (sample)
    g. snp (simple mutation/variation)
    h. sv (structural mutation/variation)

    A set of example input files can be found under the 'workdir_test/input' directory.

- Run data validation as below:

cd */home/software/*martloader/*workdir_testSept10*
perl runme.pl -c

- When validation finishes, please review the log files under */home/software/*martloader/*workdir_testSept10/logs*. Empty log files (0 bytes) can be safely ignored. Otherwise, review the messages in the log files. After making any necessary changes to the submission files, please rerun data validation.

4. File submission

- After the submission files have passed validation check, the files should be compressed and uploaded to DCC's Secure FTP server (data.dcc.icgc.org).

- Contact DCC if you need an SFTP account for file uploading or if you experience any difficulty with the SFTP server.

5. Setup BioMart server (optional)

As an alternative to submitting data to DCC, you can setup the data server on your own side by following the steps below:

- Install and configure MySQL database server and create necessary MySQL user account.

- Create a text file named 'dbuser' under /home/software/martloader/bin/, an example file is shown below:
    host=your_host.com
    port=3306

user=your_user_name
pass=password

Please note that this MySQL account needs to have permission to create databases.

- Run data loading as below:

cd */home/software/*martloader/*workdir_testSept10*
perl runme.pl -l

Important: with the above command, martloader will **delete** a MySQL database named *dcc_testSept10* if it exists, and it will create a new *dcc_testSept10* database and populate it with data transformed from submission files.

Once martloader finishes, please review the log files under */home/software/*martloader/*workdir_testSept10/logs*. Empty log files (0 bytes) can be safely ignored. Otherwise, review the messages in the log files. After making any necessary changes to the submission files, please rerun data loading again.

After data successfully loaded in the previous step, please consult the ***Preconfigure Portal Deployment*** section in the ***User Manual*** (available from http://www.biomart.org/rc6_documentation.pdf) of BioMart 0.8 release candidate 6, for configuring and setting up the BioMart server.

Appendix A: DCC Data Element Specification

Please do not leave any data elements empty in the submission files. Besides the possible values detailed in the tables below, values can also be one of the these codes:

-999 = data not supplied at this time
-888 = not applicable
-777 = data verified to be unknown

Legend: R = required, O = optional

1. Simple Somatic Mutations/Simple Germline Variations (SSM/SGV)

SSM and SGV include single base substitutions, multiple base substitutions (> 1bp and ≤200 bp) and short indels of ≤200 bp in length.

Simple Somatic Mutations (SSM) - Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	tumour_sample_id	Unique identifier for the tumour sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the control matched to the tumour sample		
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	

16	O	note	Optional field to leave notes	text	

Simple Somatic Mutations (SSM) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		
3	R	mutation_id	Unique identifier for the mutation		
4	R	mutation_type	Type of mutation	integer	1 = single base substitution 2 = insertion of <= 200 bp 3 = deletion of <= 200 bp 4 = multiple base substitution (>= 2bp and <= 200bp)
5	R	chromosome	Name of the chromosome containing the mutation/variation	integer	Appendix Table B6
6	R	chromosome_start	Start position of the mutation/variation on the chromosome	integer	
7	R	chromosome_end	End position of the mutation/variation on the chromosome	integer	
8	R	chromosome_strand	Chromosome strand	integer	1 = 1 -1 = -1
9	R	refsnp_allele	RefSNP alleles from dbSNP (use a dash for each missing base)	text	e.g. A/T, ---/AAA
10	O	refsnp_strand	Strand of RefSNP allele	integer	1 = 1 -1 = -1
11	R	reference_genome_allele	Allele in the reference genome (use a dash for each missing base)	text	
12	R	control_genotype	Genotype of the control sample (use a dash for each missing base)	text	
13	R	tumour_genotype	Genotype of the tumour sample (use a dash for each missing base)	text	
14	R	mutation	Mutation, e.g. C > G	text	
15	O	expressed_allele	The expressed allele(s) as revealed by RNA-seq, etc.	text	
16	O	quality_score	Average quality score for the mutation/variation call	integer	
17	O	probability	Probability of the mutation/variation call	decimal	
18	O	read_count	Average number of times the bases are covered by raw reads	decimal	
19	O	is_annotated	Indicate if the mutation/variation is annotated in dbSNP	integer	1 = annotated 2 = not annotated
20	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested

					3 = not valid
21	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
22	O	xref_ensembl_var_id	Cross-reference: Ensembl Variation ID	text	Variation ID in Ensembl Variation Database: e.g. rs12345; ENSSNP53189
23	O	note	Optional field to leave notes	text	

Simple Somatic Mutations (SSM) – Secondary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		
3	R	mutation_id	Unique identifier for the mutation		
4	R	consequence_type	Functional consequence of the SNP.	integer	Appendix Tables B7 & B8
5	O	aa_mutation	Changes at amino acid level. Indicate the reference aa, position and mutation aa.	text	e.g. P234W
6	O	cds_mutation	Changes in coding sequence. Indicate position, reference base and mutation base.	text	e.g. 12324T>G
7	O	protein_domain_affected	Protein domain containing the mutation/variation. Use Pfam accession.	text	
8	O	gene_affected	Gene(s) containing the mutation/variation.	text	
9	O	transcript_affected	Transcript(s) containing the mutation/variation. Use Ensembl transcript id.	text	
10	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	55
11	O	note	Optional field to leave notes	text	

Note: when a mutation affects more than one transcript, please use multiple rows to record the mutation consequence, one row per transcript.

Simple Germline Variations (SGV) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	control_sample_id	Unique identifier for the control sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the tumour matched to the control sample		
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table

					B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Simple Germline Variations (SGV) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	control_sample_id	Unique identifier for the control sample		
3	R	variation_id	Unique identifier for the variation		
4	R	variation_type	Type of variation	integer	1 = single base substitution 2 = insertion of <= 200 bp 3 = deletion of <= 200 bp 4 = multiple base substitution (>= 2bp and <= 200bp)
5	R	chromosome	Name of the chromosome containing the mutation/variation	integer	Appendix Table B6
6	R	chromosome_start	Start position of the mutation/variation on the chromosome	integer	
7	R	chromosome_end	End position of the mutation/variation on the chromosome	integer	
8	R	chromosome_strand	Chromosome strand	integer	1 = 1 -1 = -1
9	R	refsnp_allele	RefSNP alleles from dbSNP (use a dash for each missing base)	text	e.g. A/T, ---/AAA
10	O	refsnp_strand	Strand of RefSNP allele	integer	1 = 1

					-1 = -1
11	R	reference_genome_allele	Allele in the reference genome (use a dash for each missing base)	text	
12	R	control_genotype	Genotype of the control sample (use a dash for each missing base)	text	
13	R	tumour_genotype	Genotype of the tumour sample (use a dash for each missing base)	text	
14	O	expressed_allele	The expressed allele(s) as revealed by RNA-seq, etc.	text	
15	O	quality_score	Average quality score for the mutation/variation call	integer	
16	O	probability	Probability of the mutation/variation call	decimal	
17	O	read_count	Average number of times the bases are covered by raw reads	decimal	
18	O	is_annotated	Indicate if the mutation/variation is annotated in dbSNP	integer	1 = annotated 2 = not annotated
19	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested 3 = not valid
20	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
21	O	xref_ensembl_var_id	Cross-reference: Ensembl Variation ID	text	Variation ID in Ensembl Variation Database: e.g. rs12345; ENSSNP53189
22	O	note	Optional field to leave notes	text	

Further explanations for the following data elements in the Simple Mutation Dataset:

a. chromosome, chromosome_start, chromosome_end

- Nucleotide position in DNA sequence is expected to start from 1 for the first nucleotide of the forward strand, counting one by one up to the end.
- For any feature on the genome, chromosome_start is always less than or equal to chromosome_end.
- The size of a feature is calculated as: chromosome_end - chromosome_start + 1.
- For single nucleotide substitution, use the coordinate of the mutated nucleotide to report the mutation, e.g. chromosome:chr1, chromosome_start:12345, chromosome_end:12345.
- For multiple nucleotide substitution (≥2bp and ≤200bp), use the start and end coordinates of the mutated fragment, e.g. chromosome: chr1, chromosome_start: 12345, chromosome_end: 12355 for a 11bp substitution.
- For deletion, use the coordinates of the deleted fragment. e.g. chr1:12345-12355 is an 11 bp deletion from 12345 to 12355 on chromosome 1.
- For insertion, use the coordinate of the nucleotide that is immediately after the insertion point. e.g. an insertion at chr1:12345-12345 means that a fragment of DNA sequence is inserted immediately before position 12345 on chromosome 1.

b. chromosome_strand

- ‘chromosome_strand’ is used to record the reference genome strand on which the genotype alleles are located.
- For genotype detected using sequencing platforms, the forward strand sequence is used for genotypes, so chromosome_strand is always forward (i.e. 1).
- For genotype that is called using array based platforms, chromosome_strand can be either forward or reverse depending on what is reported by the assay.
- ‘chromosome_strand’ does not have anything to do with the strandness of the gene that contains the simple mutation.

c. mutation_type

- 1 = single base substitution
- 2 = insertion of <= 200 bp
- 3 = deletion of <= 200 bp
- 4 = multiple base substitution (>= 2bp and <= 200bp)

d. control_genotype, tumour_genotype

- Genotype is presented as nucleotide sequence all allele(s). For example, in a diploid genome at chr1:12345-12345, if one allele on the forward strand is A and the other is G, then the genotype is presented as A/G and ‘chromosome_strand’ being ‘1’ (i.e. forward strand). It may also be presented as T/C with ‘chromosome_strand’ being ‘-1’ (i.e. reverse strand).
- In the case that the genotype is hemizygous (e.g. G allele is missing), it can be presented as A/-.
- 'control_genotype' and 'tumour_genotype' are used to record genotype for the matched control sample and the primary tumour sample, respectively. Both genotypes must be presented using the same strand of the reference genome.
- Usually, genotypes in control samples are homozygous, and the nucleotides are the same as the reference genome. For example, at chr1:456789-456789, both alleles are A as in the reference genome, so the control genotype should be A/A.
- Due to aneuploidy and normal tissue contamination, it can be difficult to determine zygosity of tumour samples. In the previous example, the genotype of the tumour sample may be G/G but may appear as A/G when the sample is contaminated. If the tumour genotype can not be determined, please use -777 to indicate 'data verified to be unknown'.

e. mutation

- 'mutation' records the somatic mutation in the tumour sample.
- For mutation on a single allele, provide the control and tumour sample alleles separated by '>'. For example, 'A>G' indicates that one allele has an A to G mutation (single nucleotide substitution) in the tumour sample.
- In the case that both alleles are mutated, provide the control genotype and tumour genotype separated by '>', e.g. 'A/T>C/G'.
- For multiple nucleotide substitution (≥2bp and ≤200bp), provide the nucleotide sequences in the control and tumour sample alleles separated by '>', e.g. 'CTGAG>AGCCT'.

- For deletions, '-' is expected to represent each missing nucleotide, for example, at chr1:1234-1236, three nucleotides ATG are missing in the tumour sample, it is expressed as 'ATG>---'.
- For insertions, e.g. a DNA fragment 'CTGAG' inserted before nucleotide 'T' at chr1:12345-12345 can be presented as '->CTGAG'.

f. reference_genome_allele

- 'reference_genome_allele' is the forward strand nucleotide(s) at the corresponding location on the reference genome where the somatic mutation is detected in the tumour sample.

g. refsnp_alleles

- At the genomic location of a somatic mutation, if a refSNP entry is found in dbSNP database, the alleles described in that refSNP should be presented in 'refsnp_alleles'.
- When no refSNP is presented in dbSNP, use '-777' to indicate 'data verified to be unknown'.

h. refsnp_strand

- If a refSNP is presented, its strandness compared with reference genome assembly should be recorded in 'refsnp_strand'. For example, rs72466451 is located at chr2:198363487-198363487, the alleles are presented using reserve strand (i.e. "-1").

i. is_annotated

- 'is_annotated' indicates whether a SNP is known at the location of the reported mutation.
- If a SNP is present in dbSNP, please use 'annotated', otherwise use 'not annotated'.
- For mutation detected using array based platforms, the SNP should be 'annotated' since the microarray probes are designed from known SNPs.

2. Copy Number Somatic Mutations/Copy Number Germline Variations (CNSM/CNGV)

Copy Number Somatic Mutations (CNSM) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	tumour_sample_id	Unique identifier for the tumour sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the control matched to the tumour sample		
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table

					B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	O	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	O	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	O	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Copy Number Somatic Mutations (CNSM) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		
3	R	mutation_id	Unique identifier for the mutation		
4	R	mutation_type	Type of mutation	integer	1 = gain 2 = loss 3 = copy neutral LOH
5	R	chromosome	Name of the chromosome containing the mutation/variation (#)	integer	Appendix Table B6
6	R	chromosome_start	Start position of the mutation/variation on the chromosome	integer	
7	R	chromosome_end	End position of the mutation/variation on the chromosome	integer	
8	O	chromosome_start_range	Number of bases around chromosome_start that may contain the start position	integer	0 if start position is exactly at chromosome_start; positive integer for +/- number of bases around chromosome_start

9	O	chromosome_end_range	Number of bases around chromosome_end that may contain the end position	integer	0 if end position is exactly at chromosome_end ; positive integer for +/- number of bases around chromosome_end
10	O	start_probe_id	Probe id containing the chromosome_start if array platform was used	text	
11	O	end_probe_id	Probe id containing the chromosome_end if array platform was used	text	
12	O	copy_number	DNA copy number estimated	decimal	
13	O	quality_score	Quality score for the mutation/variation call	decimal	
14	O	probability	Probability of the mutation/variation call	decimal	
15	O	is_annotated	Indicate if the mutation/variation is annotated in the Database of Genomic Variants	integer	1 = annotated 2 = not annotated
16	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested 3 = not valid
17	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
18	O	note	Optional field to leave notes	text	

Copy Number Somatic Mutations (CNSM) – Secondary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		
3	R	mutation_id	Unique identifier for the mutation		
4	R	gene_affected	Gene(s) containing the mutation/variation. 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).	text	
5	O	transcript_affected	Transcript(s) containing the mutation/variation. Use Ensembl transcript id. Separate multiple transcripts from the same gene with commas, and separate transcripts from different genes with vertical bars. eg. transcriptA1,	text	

			transcriptA2|transcriptB1|transcriptC1,transcriptC2,transcriptC3. If no transcript is affected, use -888 (not applicable).		
6	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
7	O	note	Optional field to leave notes	text	

Copy Number Germline Variations (CNGV) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	control_sample_id	Unique identifier for the control sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the tumour matched to the control sample		
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	O	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	O	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	O	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Copy Number Germline Variations (CNGV) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of		

			samples		
2	R	control_sample_id	Unique identifier for the control sample		
3	R	variation_id	Unique identifier for the variation		
4	R	variation_type	Type of variation	integer	1 = gain 2 = loss 3 = copy neutral LOH
5	R	chromosome	Name of the chromosome containing the mutation/variation	integer	Appendix Table B6
6	R	chromosome_start	Start position of the mutation/variation on the chromosome	integer	
7	R	chromosome_end	End position of the mutation/variation on the chromosome	integer	
8	O	chromosome_start_range	Number of bases around chromosome_start that may contain the start position	integer	0 if start position is exactly at chromosome_start; positive integer for +/- number of bases around chromosome_start
9	O	chromosome_end_range	Number of bases around chromosome_end that may contain the end position	integer	0 if end position is exactly at chromosome_end; positive integer for +/- number of bases around chromosome_end
10	O	start_probe_id	Probe id containing the chromosome_start if array platform was used	text	
11	O	end_probe_id	Probe id containing the chromosome_end if array platform was used	text	
12	O	copy_number	DNA copy number estimated	decimal	
13	O	quality_score	Quality score for the mutation/variation call	decimal	
14	O	probability	Probability of the mutation/variation call	decimal	
15	O	is_annotated	Indicate if the mutation/variation is annotated in the Database of Genomic Variants	integer	1 = annotated 2 = not annotated
16	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested 3 = not valid
17	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
18	O	note	Optional field to leave notes	text	

3. Structural Somatic Mutations/Structural Germline Variations (StSM/StGV)

Structural Somatic Mutations (StSM) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor	text	
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor	text	
4	R	tumour_sample_id	Unique identifier for the tumour sample donated by the donor	text	
5	R	matched_sample_id	Unique identifier for the control matched to the tumour sample	text	
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Structural Somatic Mutations (StSM) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample	text	
3	R	sv_id	Unique variant id (institute wide). One id per single event	text	

4	R	placement	Ordering of breakpoint pairs within a single structural mutation/variation event	integer	
5	R	annotation	Annotation describing sequence mutation/variation based on breakpoint pairs	text	
6	O	interpreted_annotation	HGVS nomenclature for description of sequence mutation/variation. E.g. chr3:g.1234567-2345678inv.	text	
7	R	variant_type	Type of mutation/variation	integer	Appendix Table B9
8	R	chr_from	Name of the donor chromosome containing the mutation/variation	integer	Appendix Table B6
9	R	chr_from_bkpt	Breakpoint position of the mutation/variation on the donor chromosome	integer	
10	R	chr_from_strand	Donor chromosome strand	integer	1 = 1 -1 = -1
11	O	chr_from_range	Number of bases around chr_from_bkpt that may contain the real breakpoint	integer	
12	O	chr_from_flanking_seq	Flanking sequences that are 200bp upstream and 200bp downstream to the chr_from_bkpt position.	text	
13	R	chr_to	Name of the acceptor chromosome containing the mutation/variation	integer	Appendix Table B6
14	R	chr_to_bkpt	Breakpoint position of the mutation/variation on the acceptor chromosome	integer	
15	R	chr_to_strand	Acceptor chromosome strand	integer	1 = 1 -1 = -1
16	O	chr_to_range	Number of bases around chr_to_bkpt that may contain the real breakpoint	integer	
17	O	chr_to_flanking_seq	Flanking sequences that are 200bp upstream and 200bp downstream to the chr_to_bkpt position.	text	
18	O	microhomology_sequence	If a microhomology is inserted, provide sequence	text	
19	O	non_templated_sequence	If non-templated DNA is inserted, provide sequence	text	
20	O	evidence	Evidence supporting a structural mutation/variation	integer	1 = Copy number change 2 = FISH 3 = Flow-sorted chromosome evidence 4 = Paired sequence either side of breakpoint 5 = Partner breakpoint found 6 = PCR product

					across breakpoint 7 = Protein evidence 8 = Seen in multiple samples 9 = Sequence across breakpoint
21	O	quality_score	Quality score for the mutation/variation call	integer	
22	O	probability	Probability of the mutation/variation call	decimal	
23	O	zygosity	Zygosity	integer	1 = homozygous 2 = heterozygous 3 = hemizygous 4 = nullizygous
24	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested 3 = not valid
25	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
26	O	db_xref	Value code of cross-reference database:ID of the mutation in the cross-reference database. Separate multiple entries by commas.	text	
27	O	note	Optional field to leave notes	text	

Structural Somatic Mutatiosn (StSM) – Secondary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample	text	
3	R	sv_id	Unique variant id (institute wide). One id per single event	text	
4	R	placement	Ordering of breakpoint pairs within a single structural change event	integer	
5	O	bkpt_from_context	Contextual description of the first break location (Exonic, Intronic, Intergenic)	text	
6	O	gene_affected_by_bkpt_from	Gene(s) affected by the breakpoints. Use Ensembl gene id. Separate multiple genes with vertical bars in the form of geneA|geneB|geneC. If both breakpoints affect genes, then use "|" to separate them. If no gene is affected, use -888 (not applicable).	text	
7	O	transcript_affected_by_bkpt_from	Transcript(s) affected by the breakpoints. Use Ensembl transcript id. Separate multiple transcripts from the same gene with commas, and separate	text	

			transcripts from different genes with vertical bars. eg. transcriptA1, transcriptA2|transcriptB1|transcriptC1		
8	O	bkpt_to_context	Contextual description of the second break location (Exonic, Intronic, Intergenic)	text	
9	O	gene_affected_by_bkpt_to	Gene(s) affected by the breakpoints. Use Ensembl gene id. Separate multiple genes with vertical bars in the form of geneA|geneB|geneC. If both breakpoints affect genes, then use "|" to separate them. If no gene is affected, use -888 (not applicable).	text	
10	O	transcript_affected_by_bkpt_to	Transcript(s) affected by the breakpoints. Use Ensembl transcript id. Separate multiple transcripts from the same gene with commas, and separate transcripts from different genes with vertical bars. eg. transcriptA1, transcriptA2|transcriptB1|transcriptC1	text	
11	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
12	O	note	Optional field to leave notes	text	

Structural Germline Variations (StGV) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	control_sample_id	Unique identifier for the control sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the tumour matched to the control sample		
6	R	assembly_version	Version of reference genome assembly (#)	integer	Appendix Table B10
7	R	platform	Platform or technology used in detecting the mutation/variation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by		

			commas.		
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA 2 = dbSNP
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Structural Germline Variations (StGV) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	control_sample_id	Unique identifier for the control sample	text	
3	R	sv_id	Unique variant id (institute wide). One id per single event	text	
4	R	placement	Ordering of breakpoint pairs within a single structural mutation/variation event	integer	
5	R	annotation	Annotation describing sequence mutation/variation based on breakpoint pairs	text	
6	O	interpreted_annotation	HGVS nomenclature for description of sequence mutation/variation. E.g. chr3:g.1234567-2345678inv.	text	
7	R	variant_type	Type of mutation/variation	integer	Appendix Table B9
8	R	chr_from	Name of the donor chromosome containing the mutation/variation	integer	Appendix Table B6
9	R	chr_from_bkpt	Breakpoint position of the mutation/variation on the donor chromosome	integer	
10	R	chr_from_strand	Donor chromome strand	integer	1 = 1 -1 = -1
11	O	chr_from_range	Number of bases around chr_from_bkpt that may contain the real breakpoint	integer	
12	O	chr_from_flanking_seq	Flanking sequences that are 200bp upstream and 200bp downstream to the chr_from_bkpt position.	text	
13	R	chr_to	Name of the acceptor chromosome containing the mutation/variation	integer	Appendix Table B6
14	R	chr_to_bkpt	Breakpoint position of the mutation/variation on the acceptor chromosome	integer	
15	R	chr_to_strand	Acceptor chromome strand	integer	1 = 1 -1 = -1
16	O	chr_to_range	Number of bases around chr_to_bkpt that may contain the real breakpoint	integer	

17	O	chr_to_flanking_seq	Flanking sequences that are 200bp upstream and 200bp downstream to the chr_to_bkpt position.	text	
18	O	microhomology_sequence	If a microhomology is inserted, provide sequence	text	
19	O	non_templated_sequence	If non-templated DNA is inserted, provide sequence	text	
20	O	evidence	Evidence supporting a structural mutation/variation	integer	1 = Copy number change 2 = FISH 3 = Flow-sorted chromosome evidence 4 = Paired sequence either side of breakpoint 5 = Partner breakpoint found 6 = PCR product across breakpoint 7 = Protein evidence 8 = Seen in multiple samples 9 = Sequence across breakpoint
21	O	quality_score	Quality score for the mutation/variation call	integer	
22	O	probability	Probability of the mutation/variation call	decimal	
23	O	zygosity	Zygosity	integer	1 = homozygous 2 = heterozygous 3 = hemizygous 4 = nullizygous
25	R	validation_status	Indicate if the mutation/variation has been validated	integer	1 = validated 2 = not tested 3 = not valid
26	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
27	O	db_xref	Value code of cross-reference database:ID of the mutation in the cross-reference database. Separate multiple entries by commas.	text	
28	O	note	Optional field to leave notes	text	

4. Gene Expression

Expression – Metadata File

Order	O/R	Data element	Description	Data	Values

				type	
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	sample_id	Unique identifier for the sample being analyzed		
5	R	assembly_version	Version of reference genome assembly	integer	Appendix Table B10
6	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
7	R	platform	Platform or technology used in detecting the expression	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	normalization_algorithm	Name of normalization algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Expression – Gene File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	sample_id	Unique identifier for the sample being analyzed		
3	R	gene_stable_id	For annotated gene, use Ensembl gene ID. Otherwise, use assemblyBuild_chr_start_end where assemblyBuild is hg18 or hg19.	text	
4	R	gene_chromosome	Name of the chromosome containing the mRNA	integer	Appendix Table B6
5	R	gene_strand	Strand of the chromosome	integer	1 = 1 -1 = -1

6	R	gene_start	Start position of the gene on the chromosome	integer	
7	R	gene_end	End position of the transcript on the chromosome	integer	
8	R	normalized_read_count	Normalized count of sequencing reads if analyzed by sequencing platforms	decimal	
9	R	raw_read_count	Raw count of sequencing reads if analyzed by sequencing platforms	integer	
10	O	normalized_expression_level	Normalized value of expression level if analyzed by microarray platforms	decimal	
11	O	fold_change	Expressed fold change if differential expression is measured	decimal	
12	O	reference_sample	ID of the reference sample if differential expression is measured	text	
13	O	quality_score	Quality score for the expression call	integer	
14	O	probability	Probability of the expression call	decimal	
15	O	is_annotated	Indicate if the expressed fragment is annotated in Ensembl	integer	1 = annotated 2 = not annotated
16	R	validation_status	Indicate if the expressed fragment has been validated	integer	1 = validated 2 = not tested 3 = not valid
17	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
18	O	probeset_id	ID of the probeset used in microarray	test	
19	O	note	Optional field to leave notes	text	

5. miRNA

miRNA – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identifier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	sample_id	Unique identifier for the sample being analyzed		
5	R	assembly_version	Version of reference genome assembly (#)	integer	Appendix Table B10
6	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
7	R	platform	Platform or technology used in detecting the expression	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	

10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	normalization_algorithm	Name of normalization algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	O	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA
15	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

miRNA – Expression File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identifier for the analysis performed for a particular group of samples		
2	R	sample_id	Unique identifier for the sample being analyzed		
3	R	mirna_seq	Sequence of the miRNA	text	
4	R	normalized_read_count	Normalized count of sequencing reads if analyzed by sequencing platforms	decimal	
5	R	raw_read_count	Raw count of sequencing reads if analyzed by sequencing platforms	integer	
6	O	normalized_expression_level	Normalized value of expression level if analyzed by microarray platforms	decimal	
7	O	fold_change	Expressed fold change if differential expression is measured	decimal	
8	O	reference_sample	ID of the reference sample if differential expression is measured	text	
9	O	quality_score	Quality score for the call	integer	
10	O	probability	Probability of the call	decimal	
11	O	is_annotated	Indicate if the fragment is annotated	integer	1 = annotated 2 = not annotated
12	R	validation_status	Indicate if the fragment has been validated	integer	1 = validated 2 = not tested 3 = not valid
13	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
14	O	note	Optional field to leave notes	text	

miRNA – Mapping Information File

Order	O/R	Data element	Description	Data type	Values
1	R	mirna_seq	Sequence of the miRNA	text	

2	R	chromosome	Name of the chromosome expressing the fragment (#)	integer	Appendix Table B6
3	R	chromosome_start	Start position on the chromosome	integer	
4	R	chromosome_end	End position on the chromosome	integer	
5	O	chromosome_strand	Strand of the chromosome	integer	1 = 1 -1 = -1
6	O	xref_mirbase_id	Cross-reference to miRBase ID (e.g. has-let-7c) if available	text	
7	O	note	Optional field to leave notes	text	

6. Exon Junction

Exon Junction – Metadata File

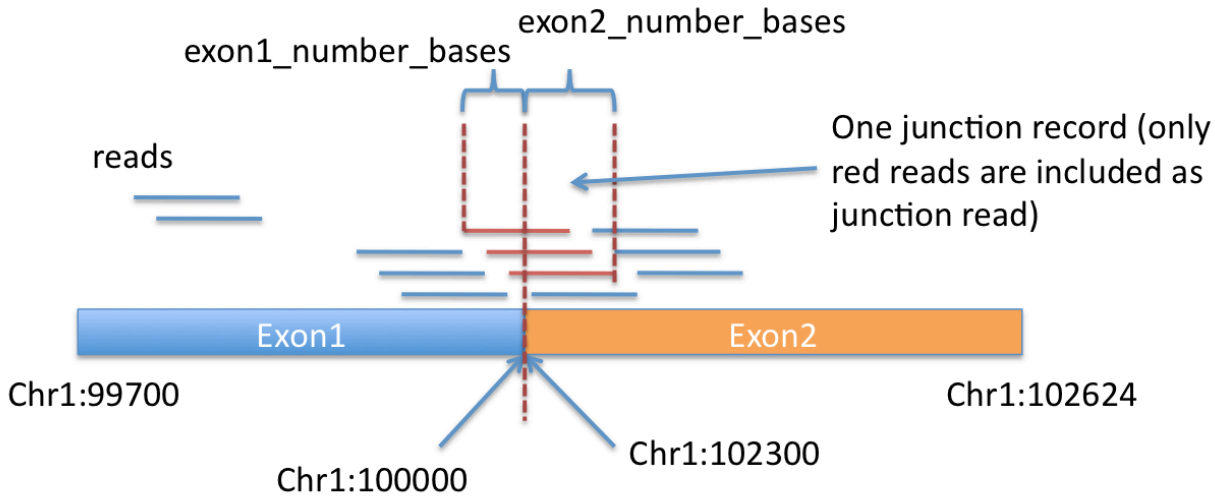
Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	sample_id	Unique identifier for the sample being analyzed		
5	R	assembly_version	Version of reference genome assembly (#)	integer	Appendix Table B10
6	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
7	R	platform	Platform or technology used in detecting the expression	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	
11	R	normalization_algorithm	Name of normalization algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.	text/url	
13	O	seq_coverage	Sequence coverage if analyzed by sequencing platforms	decimal	
14	R	raw_data_repository	Public repository where raw data is submitted (#)	integer	1 = EGA
15	R	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
16	O	note	Optional field to leave notes	text	

Exon Junction – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	sample_id	Unique identifier for the sample being analyzed		
3	R	junction_id	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 exone2chr 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.	text	
4	R	gene_stable_id	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.	text	
5	R	gene_chromosome	Name of the chromosome containing the above gene.	integer	Appendix Table B6
6	R	gene_strand	Strand of the chromosome	integer	1 = 1 -1 = -1
7	R	gene_start	Start position of the entire gene on the chromosome as annotated in Ensembl	integer	
8	R	gene_end	End position of the entire gene on the chromosome as annotated in Ensembl	integer	
9	O	second_gene_stable_id	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.	text	
10	R	exon1_chromosome	Name of the chromosome containing the 5' exon (#)	integer	Appendix Table B6
11	R	exon1_number_bases	Number of bases from 5' exon	integer	
12	R	exon1_end	End position of the 5' exon on the chromosome	integer	
13	O	exon1_strand	Chromsome strand of the 5' exon	integer	1 = 1 -1 = -1

14	R	exon2_chromosome	Name of the chromosome containing the 3' exon (#)	integer	Appendix Table B6
15	R	exon2_number_bases	Number of bases from 3' exon	integer	
16	R	exon2_start	Start position of the 3' exon on the chromosome	integer	
17	O	exon2_strand	Chromsome strand of the 3' exon	integer	1 = 1 -1 = -1
18	O	is_fusion_gene	Indicate if the function is the result of a fusion gene	integer	1 = yes 2 = no
19	O	is_novel_splice_form	Indicate if the splice form is novel	integer	1 = yes 2 = no
20	O	junction_seq	Provide junction sequence if either is_fusion_gene or is_novel_splice_form is true	text	
21	O	junction_type	Type of junction	integer	1 = canonical 2 = non-canonical 3 = U12
22	R	junction_read_count	Count of sequencing reads that span across exons	decimal	
23	O	quality_score	Quality score for the junction call	integer	
24	O	probability	Probability of the junction call	decimal	
25	R	validation_status	Indicate if the junction has been validated	integer	1 = validated 2 = not tested 3 = not valid
26	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
27	O	note	Optional field to leave notes	text	

The following diagram illustrates how junction_id is assigned, how junction_read_count, exon1_number_bases and exon2_number_bases are calculated:



- junction_id is: hg19_1_100000_1_102300
- junction read count is: 3

7. DNA Methylation

Methylation (METH) – Metadata File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	donor_id	Unique identifier for the donor		
3	R	diagnosis_id	Unique identifier for the diagnosis record for the donor		
4	R	tumour_sample_id	Unique identifier for the tumour sample donated by the donor		
5	R	matched_sample_id	Unique identifier for the control matched to the tumour sample		
6	R	assembly_version	Version of reference genome assembly	integer	Appendix Table B10
7	R	platform	Platform or technology used in detecting the methylation	integer	Appendix Table B5
8	O	experimental_protocol	Name of experimental protocol and URL to written protocol	text/url	
9	R	base_calling_algorithm	Name of base calling algorithm and URL to written protocol	text/url	
10	R	alignment_algorithm	Name of alignment algorithm and URL to written protocol	text/url	

11	R	variation_calling_algorithm	Name of variation calling algorithm and URL to written protocol	text/url	
12	O	other_analysis_algorithm	Names of other analysis algorithms. Separate multiple algorithms by commas.		
13	O	raw_data_repository	Public repository where raw data is submitted	integer	1 = EGA 2 = dbSNP
14	O	raw_data_accession	Accession and URL for referencing the raw data at the public repository	text/url	
15	O	note	Optional field to leave notes	text	

Methylation (METH) – Primary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		
3	R	methylated_fragment_id	Unique identifier for the methylated fragment		
4	R	chromosome	Name of the chromosome containing the methylation	integer	Appendix Table B6
5	R	chromosome_start	Start position of the methylation on the chromosome	integer	
6	R	chromosome_end	End position of the methylation on the chromosome	integer	
7	O	chromosome_strand	Chromosome strand	integer	1 = 1 -1 = 1
8	R	percent_methylation_1	Percent methylation or beta value for probe 1	decimal	
9	R	percent_methylation_2	Percent methylation or beta value for probe 2	decimal	
10	O	quality_score	Quality score for the methylation call	integer	
11	O	probability	Probability of the methylation call	decimal	
12	R	validation_status	Indicate if the methylation has been validated	integer	1 = validated 2 = not tested 3 = not valid
13	O	validation_platform	Platform or technology used in validation	integer	Appendix Table B5
14	O	note	Optional field to leave notes	text	

Methylation (METH) – Secondary Analysis File

Order	O/R	Data element	Description	Data type	Values
1	R	analysis_id	Unique identfier for the analysis performed for a particular group of samples		
2	R	tumour_sample_id	Unique identifier for the tumour sample		

3	R	methylated_fragment_id	Unique identifier for the methylation		
4	R	gene_affected	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).	text	
5	R	gene_build_version	Version of Ensembl gene build used for annotation	integer	
6	O	note	Optional field to leave notes	text	

8. Clinical and Sample Annotation

Donor Data File

Order	O/R	Data element	Description	Data type	Values
1	R	donor_id	Unique identifier for a donor. It should be a de-identified code that does not link explicitly to the particular individual.	text	
2	O	biobank_id	Unique identifier for a biobank	text	
3	R	gender	Gender of the donor (others: Turner syndrome, hermaphrodites, etc)	integer	1 = male 2 = female 3 = other
4	O	ethnicity	Ethnicity of the donor (others: khoisanid, australoid, etc)	text	1 = negroid 2 = mongoloid 3 = caucasoid 4 = other
5	O	country_of_residence	Country of residence	text	au = Australia ca = Canada cn = China fr = France de = Germany in = India jp = Japan es = Spain uk = UK us = USA
6	O	city_and_state_of_residence	City and state/province of residence	text	
7	O	vital_status	Indicate if the donor is alive or deceased. This element is updated every 6 months.	integer	1 = alive 2 = deceased
8	O	age_at_recruitment	Age of the donor at the time of recruitment (years)	integer	
9	R	age_at_last_follow_up	Age of the donor at last follow up (years)	integer	
10	O	age_at_death	Age of the donor at death (years)	integer	
11	O	age_at_relapse	Age of the donor at relapse (years)	integer	
12	O	relapse_type	Type of relapse	integer	1 = localized

					2 = distant
13	O	disease_outcome	Disease outcome	integer	1 = progression 2 = treatment free survival
14	O	post_diagnosis_survival	Number of months the donor survived after diagnosis (not applicable if the donor is still alive). This element is updated every 6 months	integer	
15	O	quality_of_life_karnofsky	Quality of life based on Karnofsky Performance Scale Index (0-100)	integer	
16	O	quality_of_life_ecog	Quality of life based on Eastern Cooperative Oncology Group (ECOG) Performance Status	integer	
17	O	family_history_of_cancer	Indicate if family history is available	integer	1 = Yes 2 = No
18	O	exposure_to_risk_factors	Indicate if donor has exposure to risk factors such as tobacco, alcohol and others	integer	1 = Yes 2 = No
19	O	tobacco	Indicate if there is tobacco use	integer	1 = Yes 2 = No 3 = in the past
20	O	alcohol	Indicate frequency of alcohol consumption	integer	1 = regular 2 = occasional 3 = none
21	O	environmental_exposure	Indicate if environmental exposure was recorded for the donor	integer	1 = Yes 2 = No
22	O	clinical_trial	Name of trial if donor is involved in clinical trials or cohort studies	text	
23	O	donor_record_release_date	Date of record released to DCC	date (YYYY MMDD)	
24	O	donor_record_created_date	Date of record created	date (YYYY MMDD)	
25	O	donor_record_last_update_date	Date of last update	date (YYYY MMDD)	
26	O	donor_record_notes	Optional field to leave notes	text	

Diagnosis Data File

Order	O/R	Data element	Description	Data type	Values
1	R	donor_id	Unique identifier for a donor. It should be a de-identified code that does not link explicitly to the particular individual.	text	
2	R	diagnosis_id	Unique identifier for a diagnosis record for the donor. It should be de-identified.	text	
3	O	consent	Indicate if consent was obtained	integer	1 = Yes

					2 = No
4	R	icd_10	Primary site of diagnosis (ICD-10 code)	text	Appendix Table B4
5	O	icd_o3	Morphology of cancer (ICD-O 3rd edition)	text	
6	O	therapy_type	Broad category of therapy received by the donor (*)	integer	1 = biologic response modifier 2 = chemotherapy - multiple agent 3 = chemotherapy - single agent 4 = cryotherapy 5 = hormonal therapy 6 = immuno therapy 7 = radiation - external 8 = radiation - internal 9 = surgical biopsy 10 = surgical resection - cancer directed 11 = surgical resection - non cancer directed 12 = other
7	O	therapy_response	Response of donor to the therapy (^). This element is updated every 6 months.	integer	1 = complete response 2 = partial response 3 = disease progression 4 = stable disease 5 = not evaluable
8	O	therapy_start_date	Start date of therapy	date (YYYY MMDD)	
9	O	therapy_end_date	End date of therapy	date (YYYY MMDD)	
10	O	date_of_examination	Date of examination	date (YYYY MMDD)	
11	R	date_of_diagnosis	Date of diagnosis	date (YYYY MMDD)	
12	R	age_at_diagnosis	Age of the donor at the time of diagnosis (years)	integer	
13	O	clinical_staging	Clinical staging using WHO system.	text	1 = I 2 = IA

					3 = IB 4 = IC 5 = II 6 = IIA 7 = IIB 8 = IIC 9 = III 10 = IIIA 11 = IIIB 12 = IIIC 13 = IV 14 = IVA 15 = IVB 16 = IVC
14	R	clinical_t	tumour status based on clinical examination	text	1 = T0 2 = T1 3 = T2 4 = T3 5 = T4 6 = TX 7 = Tis
15	R	clinical_n	Lymph node status based on clinical examination	text	1 = N0 2 = N1 3 = N2 4 = N3 5 = N4 6 = NX
16	R	clinical_m	Distant metastasis status based on clinical examination	text	1 = M0 2 = M1 3 = M2 4 = M3 5 = M4 6 = MX
17	O	tumour_staging_other	Alternative classification if TNM is not applicable (e.g. Binet/Rai for CLL, Ann Arbor for lymphomas, etc)		
18	O	tumour_progress	Indicate if tumour progress occurs	integer	1 = Yes 2 = No
19	O	concomitant_disease	Indicate if concomitant disease	integer	1 = Yes 2 = No
20	O	diagnosis_record_release_date	Date of record released to DCC	date (YYYYMMDD)	
21	O	diagnosis_record_created_date	Date of record created	date (YYYYMMDD)	
22	O	diagnosis_record_last_update_date	Date of last update	date (YYYYMMDD)	
23	O	diagnosis_record_notes	Optional field to leave notes	text	

Sample Data File

Order	O/R	Data element	Description	Data type	Values
1	R	donor_id	Unique identifier for a donor. It should be a de-identified code that does not link explicitly to the particular individual.	text	
2	R	diagnosis_id	Unique identifier for a diagnosis record for the donor. It should be de-identified.	text	
3	R	sample_id	Unique identifier for the sample as assigned by data provider	text	
4	O	sample_id_provided_by_central_repo	Unique identifier for the sample as provided by central repository such as biobank	text	
5	O	sample_name	Name of the sample	text	
6	O	sample_ega_accession	Sample EGA accession	text	
7	O	primary_secondary	Indicate if the tumour is primary or secondary	integer	1 = primary 2 = secondary
8	R	recurrent	Indicate if the tumour recurrent	integer	1 = Yes 2 = No
9	R	sample_type	Type of sample (#)	integer	1 = tumour tissue 2 = tumour xenograft 3 = matched control 4 = site-matched control 5 = blood 6 = buffy coat 7 = plasma 8 = serum 9 = saliva 10 = urine 11 = cell line 12 = cell line - tumour 13 = cell line - matched control
10	R	sample_collection_date	Date of sample collection or storage	date (YYYY MMDD)	
11	O	sample_collection_procedure	Procedure for collecting the sample	text	
12	R	sample_freezing_method	Method for freezing the sample	integer	1 = liquid nitrogen 2 = dry ice 3 = cyro-preservation 4 = others
13	R	tissue_fixation_protocol	Protocol for fixing the tissue	integer	1 = formalin 2 = formalin

					buffered 3 = embedding
14	R	time_between_tissue_removal_and_fixation_or_freezing	Time between tissue removal and fixation or cryo-preservation in hours and minutes (hhmm)	integer	
15	O	time_between_vascular_clamping_and_tissue_removal	Time between vascular clamping and tissue removal in hours and (hhmm)	integer	
16	O	duration_of_transport	Duration of transport in days, hours and minutes (ddhhmm)	integer	
17	O	temperature_during_transport	Temperature during transport (Celsius)	integer	
18	R	storage_method	Type of storage methods used for the sample	integer	1 = culture 2 = frozen 3 = liquid frozen 4 = parafin block 5 = RNA later frozen 6 = slide 7 = tissue array
19	O	initial_temperature_at_storage	Initial temperature at storage (Celsius)	integer	
20	O	temperature_during_storage	Temperature during storage (Celsius)	integer	
21	O	history_of_freezing_thawing	History of freezing/thawing	text	
22	R	quantity_on_hand	Amount of sample available (e.g. 3 aliquots, 5 mg, 2 tissue pieces)	text	
23	R	grading_system_used	Name of grading system used	text	
24	R	tumour_grading	Pathologist assigned grade	text	
25	R	digital_image_of_stained_section	Linkout to digital image of stained section	URL	
26	R	percent_intact_tumour_cells	Percentage of intact ("viable") tumour cells within sample	integer	
27	O	percent_necrotic_tissue	Percentage of necrotic tissue	integer	
28	O	percent_inflammatory_tissue	Percentage of inflammatory tissue	integer	
29	O	percent_debris	Percentage of debris	integer	
30	O	molecular_genetics_diagnostics	Flow cytometry charts as alternative	text	
31	O	name_of_pathologist	Initial and reference pathologist(s)	text	
32	O	sample_record_release_date	Date of record released to DCC	date (YYYYMMDD)	
33	R	sample_record_created_date	Date of record created	date (YYYYMMDD)	
34	R	sample_record_last_update_date	Date of last update	date (YYYYMMDD)	
35	O	sample_record_notes	Optional field to leave notes	text	

Appendix B: Value Codes for DEs with Controlled Vocabulary

Value codes or controlled vocabulary will be added as the projects evolve. Please contact DCC to provide suggestions.

Appendix Table B1. Lead Jurisdiction ID

B1

Lead Jurisdiction	ID
Australia	au
Canada	ca
China	cn
France	fr
Germany	de
India	in
Japan	jp
Spain	es
UK	uk
USA	us

Appendix Table B2. ID for Types of Primary Tumours

B2

Primary Tumour Type	ID
Pancreatic cancer	01
Breast cancer	02
Brain cancer	03
Colorectal cancer	04
Ovarian cancer	05
Gastric cancer	06
Liver cancer	07
Pediatric brain tumours	08
Oral cancer	09
Chronic lymphocytic leukemia	10
Lung cancer	11
Melanoma	12
Kidney renal clear cell cancinoma	13
Kidney renal papillary cell carcinoma	14
Acute Myeloid Leukemia	15
Head and Neck squamous cell carcinoma	16
Lung adenocaracinoma	17
Lung squamous cell carcinoma	18
Rectum adenocarcinoma	19
Stomach adenocarcinoma	20
Uterine Corpus Endometrioid Carcinoma	21

Colon adenocarcinoma	22

Appendix Table B3. Institute ID

B3

Institution	ID
Advanced Centre for Treatment, Research and Education in Cancer (Mumbai)	001
AMC Medical Research BV (Netherlands)	002
Applied Biosystems Inc.	003
Australian Pancreatic Cancer Network	004
Barcelona Supercomputer Center (BSC-Barcelona)	005
Baylor College of Medicine (Houston, TX)	006
BCCA (Canada)	007
Beijing Cancer Hospital/Insititute	008
Beijing Genome Institute/Shenzhen	009
Bioquant (Heidelberg)	010
British Columbia Cancer Agency (Vancouver, Canada)	011
Broad Institute (Cambridge, MA)	012
Catalan Institute of Oncology	013
Center for Cancer Research (CICSalamanca) and University Hospital	014
Center for Genomic Regulation (CRG) and Pompeu Fabra University (UPF)	015
Centre Leon Berard (Lyon, France)	016
Centre National de Génotypage (France)	017
Centre Val d'Aurelle (Montpellier, France)	019
Commissariat à l'Energie Atomique	020
CRUK (UK)	021
Dana-Farber Cancer Institute	022
DFCI (USA)	023
EMBL-EBI (Hinxton)	024
Erasmus (Netherlands)	025
European Molecular Biology Laboratory (EMBL), Heidelberg	026
Fondation Jean Dausset CEPH	027
Fondation Synergie-Lyon-Cancer	028
Garvan Institute of Medical Research	029
German Cancer Research Center (DKFZ), Heidelberg	030
Harvard Medical School and Brigham and Women's Hospital (Cambridge, MA)	031
Hiroshima University, Faculty of Medicine	032
Hospital Clinic, University of Barcelona	033
Hospital-University : AP-HP Paris (Beaujon, H. Mondor, A. Béclère and P. Brousse hospitals), Bordeaux, Rennes, Toulouse, Grenoble	034
HudsonAlpha Institute for Biotechnology (Huntsville, AL)	035
Human Genome Center, Institute of Medical Science, University of Tokyo	036
ICR (UK)	037
INCa (France)	038
Institut Curie (France)	039
Institut Génomique	041

Institut National de la Santé et de la Recherche Médicale	042
Institut National du Cancer (Boulogne-Billancourt, France)	044
Institut Paoli-Calmettes (Marseille, France)	045
Institute for Molecular Bioscience (Brisbane)	046
Institute for System Biology (Seattle, WA)	047
International Breast Cancer Genome Consortium (UK)	048
International Genome Consortium (Phoenix, AZ)	049
Johns Hopkins University (Baltimore, MD)	050
Lawrence Berkeley National Laboratory (Berkeley, CA)	051
Lund University (Sweden)	052
Massachusetts General Hospital	053
Max-Planck-Institut for Molecular Genetics (Berlin)	054
Mayo Clinic	055
Memorial Sloan-Kettering Cancer Center (New York, NY)	056
Mount Sinai Hospital (Toronto)	057
National Bioinformatics Institute	058
National Cancer Center	059
National Center for tumour Diseases (Heidelberg)	060
National DNA and tumour Bank Networks	061
National Institute of Biomedical Genomics (Kalyani)	062
National Institutes of Health; National Cancer Institute, National Human Genome Research Institute	063
National Sequencing Center (Barcelona)	064
NCI Bari (Italy)	065
Norwegian Radium Hospital (Norway)	066
Ontario Institute for Cancer Research	067
Osaka Medical Center for Cancer & Cardiovascular Diseases	068
Peking University School of Oncology	069
Peter MacCallum Cancer Centre	070
Queensland Centre for Medical Genomics	018
Queensland Institute of Medical Research	071
Radboud University (Netherlands)	072
Research Center for Advanced Science and Technology, University of Tokyo	073
RIKEN	074
Silicon Graphics Inc.	075
Singapore General Hospital (Hong Kong)	076
Spanish Cancer Research Network	077
Spanish National Cancer Research Centre (CNIO-Madrid)	078
UCSF	079
University Health Network (Toronto)	080
University of California (Santa Cruz, CA)	081
University of Cambridge (UK)	082
University of Deusto	083
University of Düsseldorf	084
University of Heidelberg	085
University of North Carolina (Chapel Hill, NC)	086

University of Oviedo	087
University of Queensland (Australia)	088
University of Southern California (Los Angeles, CA)	089
University of Tromsø (Norway)	090
University of Verona	091
Wakayama Medical University	092
Wellcome Trust Sanger Institute	093
Westmead Institute for Cancer Research	094
Washington University Genome Sequencing Center (St. Louis, MO)	095
The Cancer Genome Atlas	096

Please contact DCC if your institute is not listed or wish to modify the identifier

Appendix Table B4. ICD10 Codes for Disease Sites

B4

Disease Site	ICD10 Code
Pancreas	C25
Breast	C50
Brain	C71
Colon	C18
Rectum	C20
Ovary	C56
Liver	C22
Lung	C30-C39
Skin	C43-C44
Kidney	C64
Stomach	C16
Uterus	C54
Myeloid leukaemia	C92
Prostate	C61
Bladder	C67

Appendix Table B5: Value Codes for Platform or Validation Platform

B5

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

Affymetrix EMET Plus Premier Pack	13
Agilent Whole Human Genome Oligo Microarray Kit	14
Agilent Human Genome 244A	15
Agilent Human Genome 105A	16
Agilent Human CNV Association 2x105K	17
Agilent Human Genome 44K	18
Agilent Human CGH 1x1M	19
Agilent Human CGH 2x400K	20
Agilent Human CGH 4x180K	21
Agilent Human CGH 8x60K	22
Agilent Human CNV 2x400K	23
Agilent Human miRNA Microarray Kit (v2)	24
Agilent Human CpG Island Microarray Kit	25
Agilent Human Promoter ChIP-on-chip Microarray Set	26
Agilent Human SpliceArray	27
Illumina human1m-duo	28
Illumina human660w-quad	29
Illumina humancytosnp-12	30
Illumina human510s-duo	31
Illumina humanmethylation27	32
Illumina goldengate methylation	33
Illumina HumanHT-12 v4.0 beadchip	34
Illumina HumanWG-6 v3.0 beadchip	35
Illumina HumanRef-8 v3.0 beadchip	36
Illumina microRNA Expression Profiling Panel	37
Illumina humanht-16	38
Illumina humanht-17	39
Nimblegen Human CGH 3x720 Whole-Genome v3.0 Array	40
Nimblegen Human CGH 2.1M Whole-Genome v2.0D Array	41
Nimblegen Gene Expression 385K	42
Nimblegen Gene Expression 4x72K	43
Nimblegen Gene Expression 12x135K	44
Nimblegen Human Methylation 2.1M Whole-Genome sets	45
Nimblegen Human Methylation 385K Whole-Genome sets	46
Nimblegen CGS	47
Illumina Human1M OmniQuad chip	48
PCR and capillary sequencing	49
Custom-designed gene expression array	50
Affymetrix HT Human Genome U133A Array Plate Set	51
Agilent 244K Custom Gene Expression G4502A-07-1	52
Agilent 244K Custom Gene Expression G4502A-07-2	53
Agilent 244K Custom Gene Expression G4502A-07-3	54
Agilent Human Genome CGH Custom Microaary 2x415K	55
Affymetrix Human U133 Plus PM	56
Affymetrix Human U133 Plus 2.0	57

Affymetrix Human Exon 1.0 ST	58
Almac Human CRC	59
Illumina HiSeq	60
Affymetrix Human MIP 330K	61
Affymetrix Human Gene 1.0 ST	62
Illumina Human Omni1-Quad beadchip	63
Sequenom MassARRAY	64
Custom-designed cDNA array	65

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

Appendix Table B6. Chromosome Names for Reference Genomes NCB136 and GRCh37

B6

Chromosome Name	Values	Reference Genome	Gene Annotation
1	1	NCBI36 & GRCh37	Ensembl 53 & 55
2	2	NCBI36 & GRCh37	Ensembl 53 & 55
3	3	NCBI36 & GRCh37	Ensembl 53 & 55
4	4	NCBI36 & GRCh37	Ensembl 53 & 55
5	5	NCBI36 & GRCh37	Ensembl 53 & 55
6	6	NCBI36 & GRCh37	Ensembl 53 & 55
7	7	NCBI36 & GRCh37	Ensembl 53 & 55
8	8	NCBI36 & GRCh37	Ensembl 53 & 55
9	9	NCBI36 & GRCh37	Ensembl 53 & 55
10	10	NCBI36 & GRCh37	Ensembl 53 & 55
11	11	NCBI36 & GRCh37	Ensembl 53 & 55
12	12	NCBI36 & GRCh37	Ensembl 53 & 55
13	13	NCBI36 & GRCh37	Ensembl 53 & 55
14	14	NCBI36 & GRCh37	Ensembl 53 & 55
15	15	NCBI36 & GRCh37	Ensembl 53 & 55
16	16	NCBI36 & GRCh37	Ensembl 53 & 55
17	17	NCBI36 & GRCh37	Ensembl 53 & 55
18	18	NCBI36 & GRCh37	Ensembl 53 & 55
19	19	NCBI36 & GRCh37	Ensembl 53 & 55
20	20	NCBI36 & GRCh37	Ensembl 53 & 55
21	21	NCBI36 & GRCh37	Ensembl 53 & 55
22	22	NCBI36 & GRCh37	Ensembl 53 & 55
X	23	NCBI36 & GRCh37	Ensembl 53 & 55
Y	24	NCBI36 & GRCh37	Ensembl 53 & 55
MT	25	NCBI36 & GRCh37	Ensembl 53 & 55
c5_H2	26	NCBI36	Ensembl 53
c6_COX	27	NCBI36	Ensembl 53
c6_QBL	28	NCBI36	Ensembl 53
NT_113870	29	NCBI36	Ensembl 53
NT_113871	30	NCBI36	Ensembl 53
NT_113872	31	NCBI36	Ensembl 53
NT_113874	32	NCBI36	Ensembl 53
NT_113878	33	NCBI36	Ensembl 53
NT_113880	34	NCBI36	Ensembl 53
NT_113881	35	NCBI36	Ensembl 53
NT_113884	36	NCBI36	Ensembl 53
NT_113885	37	NCBI36	Ensembl 53
NT_113886	38	NCBI36	Ensembl 53

NT_113888	39	NCBI36	Ensembl 53
NT_113889	40	NCBI36	Ensembl 53
NT_113890	41	NCBI36	Ensembl 53
NT_113898	42	NCBI36	Ensembl 53
NT_113899	43	NCBI36	Ensembl 53
NT_113901	44	NCBI36	Ensembl 53
NT_113902	45	NCBI36	Ensembl 53
NT_113903	46	NCBI36	Ensembl 53
NT_113906	47	NCBI36	Ensembl 53
NT_113908	48	NCBI36	Ensembl 53
NT_113909	49	NCBI36	Ensembl 53
NT_113910	50	NCBI36	Ensembl 53
NT_113911	51	NCBI36	Ensembl 53
NT_113912	52	NCBI36	Ensembl 53
NT_113915	53	NCBI36	Ensembl 53
NT_113916	54	NCBI36	Ensembl 53
NT_113917	55	NCBI36	Ensembl 53
NT_113923	56	NCBI36	Ensembl 53
NT_113924	57	NCBI36	Ensembl 53
NT_113925	58	NCBI36	Ensembl 53
NT_113926	59	NCBI36	Ensembl 53
NT_113927	60	NCBI36	Ensembl 53
NT_113929	61	NCBI36	Ensembl 53
NT_113930	62	NCBI36	Ensembl 53
NT_113931	63	NCBI36	Ensembl 53
NT_113932	64	NCBI36	Ensembl 53
NT_113933	65	NCBI36	Ensembl 53
NT_113934	66	NCBI36	Ensembl 53
NT_113935	67	NCBI36	Ensembl 53
NT_113936	68	NCBI36	Ensembl 53
NT_113937	69	NCBI36	Ensembl 53
NT_113939	70	NCBI36	Ensembl 53
NT_113943	71	NCBI36	Ensembl 53
NT_113944	72	NCBI36	Ensembl 53
NT_113946	73	NCBI36	Ensembl 53
NT_113949	74	NCBI36	Ensembl 53
NT_113951	75	NCBI36	Ensembl 53
NT_113953	76	NCBI36	Ensembl 53
NT_113954	77	NCBI36	Ensembl 53
NT_113956	78	NCBI36	Ensembl 53
NT_113957	79	NCBI36	Ensembl 53
NT_113958	80	NCBI36	Ensembl 53
NT_113960	81	NCBI36	Ensembl 53
NT_113961	82	NCBI36	Ensembl 53
NT_113962	83	NCBI36	Ensembl 53
NT_113963	84	NCBI36	Ensembl 53
NT_113964	85	NCBI36	Ensembl 53
NT_113965	86	NCBI36	Ensembl 53
NT_113966	87	NCBI36	Ensembl 53
HSCHR17_1	88	GRCh37	Ensembl 55
HSCHR17_RANDOM_CTG2	89	GRCh37	Ensembl 55
HSCHR17_RANDOM_CTG3	90	GRCh37	Ensembl 55
HSCHR19_RANDOM_CTG2	91	GRCh37	Ensembl 55

HSCHR1_RANDOM_CTG12	92	GRCh37	Ensembl 55
HSCHR1_RANDOM_CTG5	93	GRCh37	Ensembl 55
HSCHR4_RANDOM_CTG2	94	GRCh37	Ensembl 55
HSCHR4_RANDOM_CTG3	95	GRCh37	Ensembl 55
HSCHR6_MHC_APD	96	GRCh37	Ensembl 55
HSCHR6_MHC_COX	97	GRCh37	Ensembl 55
HSCHR6_MHC_DBB	98	GRCh37	Ensembl 55
HSCHR6_MHC_MANN	99	GRCh37	Ensembl 55
HSCHR6_MHC_MCF	100	GRCh37	Ensembl 55
HSCHR6_MHC_QBL	101	GRCh37	Ensembl 55
HSCHR6_MHC_SSTO	102	GRCh37	Ensembl 55
HSCHR7_RANDOM_CTG1	103	GRCh37	Ensembl 55
HSCHR8_RANDOM_CTG1	104	GRCh37	Ensembl 55
HSCHR8_RANDOM_CTG4	105	GRCh37	Ensembl 55
HSCHR9_RANDOM_CTG2	106	GRCh37	Ensembl 55
HSCHR9_RANDOM_CTG4	107	GRCh37	Ensembl 55
HSCHR9_RANDOM_CTG5	108	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG1	109	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG10	110	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG11	111	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG13	112	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG14	113	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG15	114	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG16	115	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG17	116	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG2	117	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG20	118	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG21	119	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG22	120	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG23	121	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG26	122	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG29	123	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG3	124	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG30	125	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG31	126	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG32	127	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG33	128	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG34	129	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG35	130	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG36	131	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG4	132	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG40	133	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG5	134	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG6	135	GRCh37	Ensembl 55
HSCHRUN_RANDOM_CTG9	136	GRCh37	Ensembl 55

Appendix Table B7. Values for Consequences from SSM/SGV
Controlled vocabulary adopted from Ensembl Release 55

B7

Consequence	Value
3prime_utr	1
5prime_utr	2

upstream	3
downstream	4
essential_splice_site,3prime_utr	5
essential_splice_site,5prime_utr	6
essential_splice_site,intronic	7
essential_splice_site,non_synonymous_coding	8
essential_splice_site,stop_lost	9
essential_splice_site,synonymous_coding	10
frameshift_coding	11
frameshift_coding,splice_site	12
intergenic	13
intronic	14
non_synonymous_coding	15
non_synonymous_coding,splice_site	16
splice_site,3prime_utr	17
splice_site,5prime_utr	18
splice_site,intronic	19
splice_site,synonymous_coding	20
stop_gained	21
stop_gained,splice_site	22
stop_lost	23
stop_lost,splice_site	24
synonymous_coding	25
utr	26
splice_site	27
noncoding_rna	28
complex_indel	29
regulatory_region	30
inframe_indel	31
start_lost	32
ambiguous	33
complex_substitution	34

Appendix Table B8. Description of Consequences from SSM/SGV
Description adopted from Ensembl Release 55

B8

Consequence	Description
3' UTR	In 3' UTR
5' UTR	In 5' UTR
Upstream	Within 5 kb upstream of the 5'-end of a transcript
Splice site	1-3 bps into an exon or 3-8 bps into an intron
Downstream	Within 5 kb downstream of the 3'-end of a transcript
Essential splice site	In the first 2 or the last 2 basepairs of an intron
Frameshift	In coding sequence, resulting in a frameshift
Intronic	In intron
Non-synonymous	In coding sequence, resulting in an aa change

Synonymous	In coding sequence, not resulting in an aa change
Start lost	In coding sequence, resulting in the loss of a start codon
Stop lost	In coding sequence, resulting in the loss of a stop codon
Stop gained	In coding sequence, resulting in the gain of a stop codon
Regulatory region	In regulatory region annotated by Ensembl
Intergenic	More than 5 kb away from a transcript
Ambiguous	In coding sequence, resulting in unpredictable effect on amino acid due to ambiguous nucleotide change
Complex InDel	Insertion or deletion that spans an exon/intron border or a coding sequence/UTR border.
Complex substitution	Substitution that is 2bps or longer

Appendix Table B9. Values for Types of StSM/StGV
Controlled vocabulary adpted from ICGC DCM WG

B9

Type of StSM/StGV	Subtype	Value
intrachromosomal rearrangement	deletion	1
	tandem duplication	2
	inversion	3
	inverted duplication - head-to-head	4
	inverted duplication - tail-to-tail	5
	insertion	6
	intrachromosomal rearrangement with inverted orientation	7
	intrachromosomal rearrangement with non-inverted orientation	8
	fold-back inversion	9
	complex intrachromosomal rearrangement	10
interchromosomal rearrangement	reciprocal translocation	11
	unbalanced translocation	12
	interchromosomal insertion	13
	interchromosomal rearrangement - unknown type	14
	complex interchromosomal rearrangement	15
rearrangements involving amplicons	intrachromosomal amplicon-to-amplicon	16
	intrachromosomal amplicon-to-nonamplified dna	17
	interchromosomal amplicon-to-amplicon	18
	interchromosomal amplicon-to-nonamplified dna	19
	extrachromosomal	20

Appendix Table B10. Value Codes for Reference Genome Assembly Version

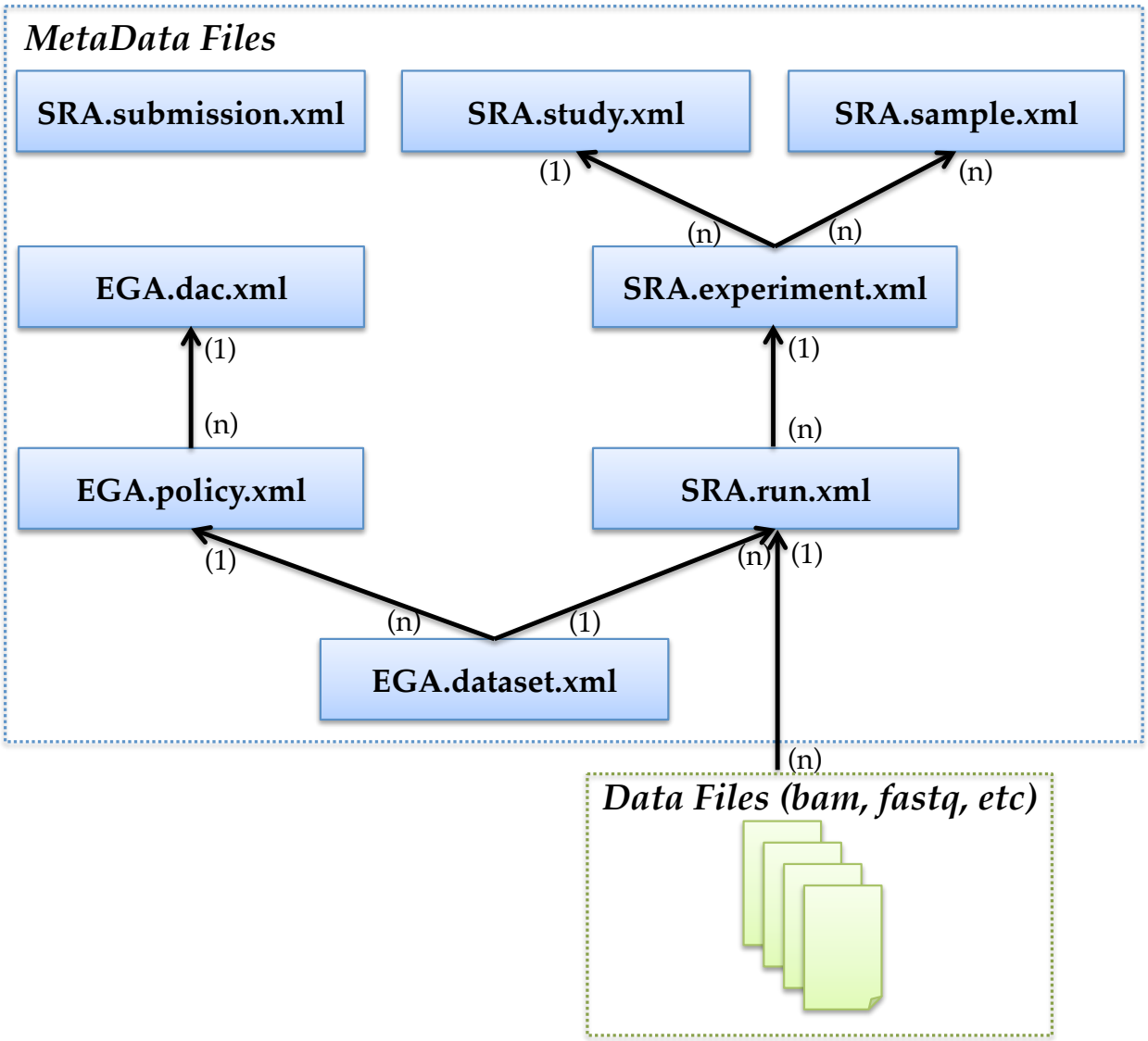
B10

Reference Genome Assembly Version	Values
GRCh37	1
NCBI36	2
GRCh37.p1	3
GRCh37.p2	4

Appendix C: EGA sequence data submission guide

The following instructions are meant to provide ICGC members with guidance on submitting raw sequence data to the European Genome-Phenome Archive (EGA). ICGC members are encouraged to consult the EGA guidelines prior to data submission. Detailed instructions on data submission are available EGA website at http://www.ebi.ac.uk/ega/page.php?page=data_submission.

C1. Overview of the SRA metadata xml files



C2. Examples of EGA submission templates

Below are example template files for DAC, POLICY, STUDY, DATASET, SAMPLE, EXPERIMENT, and RUN metadata. Further information regarding XML preparation can be found at: http://www.ebi.ac.uk/ena/about/sra_preparing_metadata

C2.1 DAC and POLICY xml files

The DAC and POLICY xml files are written as per EGA's specifications and can be used by all ICGC members in their data submission without any further modifications.

- **DAC xml file**

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<DAC_SET>
<DAC alias="ICGC Cancer Genome Projects" center_name="ICGC"
broker_name="">
 <TITLE>ICGC Data Access Compliance Office</TITLE>
 <CONTACTS>
   <CONTACT name="helpdesk" email="info@icgc.org" organisation="ICGC"
telephone_number=""/>
 </CONTACTS>
</DAC>
</DAC_SET>
```

- **POLICY xml file**

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<POLICY_SET>
<POLICY alias="ICGC Data Access Agreements" center_name="ICGC"
broker_name="">
 <TITLE>ICGC Data Access</TITLE>
 <DAC_REF refname="ICGC Cancer Genome Projects" refcenter="ICGC"/>
 <POLICY_TEXT>Please use the ICGC website for applying access to the
data</POLICY_TEXT>
 <POLICY_LINKS>
   <POLICY_LINK>
     <URL_LINK>
       <LABEL>ICGC Data Access Agreements</LABEL>
       <URL>http://www.icgc.org </URL>
     </URL_LINK>
   </POLICY_LINK>
 </POLICY_LINKS>
</POLICY>
</POLICY_SET>
```

C2.2 DATASET and STUDY xml files

The following examples of DATASET and STUDY xml files are written as per EGA's specifications with key items required for all ICGC submissions highlighted in yellow.

- **DATASET xml files**

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<DATASETS>
<DATASET alias="EGAS00010000006-ega-20110311" center_name="OICR"
broker_name="">
 <TITLE>Pancreatic Cancer Genome Sequencing</TITLE>
 <RUN_REF refname="SC_RUN_4050_1"/>
 <RUN_REF refname="SC_RUN_4000_2"/>
 <POLICY_REF refname="ICGC Data Access Agreements" refcenter="ICGC"/>
 <DATASET_LINKS>
  <DATASET_LINK>
   <URL_LINK>
    <LABEL>ICGC Data Portal</LABEL>
    <URL>http://dcc.icgc.org</URL>
   </URL_LINK>
  </DATASET_LINK>
 </DATASET_LINKS>
</DATASET>
</DATASETS>
```

- **STUDY xml file**

```
<?xml version="1.0" encoding="UTF-8"?>
<STUDY_SET>
<STUDY alias="Pancreatic Cancer Genome Sequencing" center_name="OICR">
 <DESCRIPTOR>
  <STUDY_TITLE>Title of publication</STUDY_TITLE>
  <STUDY_TYPE existing_study_type="Whole Genome Sequencing"/>
  <STUDY_ABSTRACT> STUDY ABSTRACT AS IT COULD APPEAR IN A
PUBLICATION</STUDY_ABSTRACT>
  <CENTER_PROJECT_NAME>Pancreatic Cancer Sequencing
Initiative</CENTER_PROJECT_NAME>
 </DESCRIPTOR>
 <STUDY_ATTRIBUTES>
  <STUDY_ATTRIBUTE>
   <TAG>Consortium</TAG>
   <VALUE>ICGC</VALUE>
  </STUDY_ATTRIBUTE>
  <STUDY_ATTRIBUTE>
   <TAG>Consortium Project</TAG>
   <VALUE>ICGC Cancer Genome Projects</VALUE>
  </STUDY_ATTRIBUTE>
 </STUDY_ATTRIBUTES>
</STUDY>
```

</STUDY_SET>

C2.3 SAMPLE, EXPERIMENT and RUN xml files

For SAMPLE, EXPERIMENT and RUN metadata, only fragments of the xml files are provided to illustrate how certain IDs, shown in red, are referenced among those files. Key items required for all ICGC submissions are highlighted in yellow.

- **Fragment of the SAMPLE xml file**

```
<SAMPLE alias="CLLS0123" ….>
  <SAMPLE_ATTRIBUTES>
    <SAMPLE_ATTRIBUTE>
      <TAG>Sample ID</TAG>
      <VALUE>CLLS0123</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
      <TAG>Donor ID</TAG>
      <VALUE>CLLD0015</VALUE>
    </SAMPLE_ATTRIBUTE>
  <SAMPLE_ATTRIBUTES>
</SAMPLE>
```

- **Fragment of the EXPERIMENT xml file**

```
<EXPERIMENT alias="EXP12345" ….. >
<STUDY_REF refname="Pancreatic Cancer Genome Sequencing"/>
<SAMPLE_DESCRIPTOR refname="CLLS0123"/>
```

- **Fragment of the RUN xml file**

```
<EXPERIMENT_REF refname="EXP12345"/>
<DATA_BLOCK member_name="CLLS0123">
<!-- member_name should be the name (usually sample alias) given in the experiment xml of a
pooled experiment. -->
          <FILES>
             <FILE filename="CLLS0001.bam" filetype="bam"/>
        </FILES>
</DATA_BLOCK>
```