DOWNLOAD FILE CHANGE LOG

This log details changes made to the download files. The details relate to the structure, content and naming of the files produced for v99 (November-2023) in relation to the archive versions as produced for v97 (November-2022)

KEY BENEFITS

- Increased interoperability between data sets: → IDs assigned for genes, mutations, samples and more
- Increased findability of data by gene, classification or samples: → Stable IDs between versions

CHANGES SUMMARY

Main changes in new download files:

- Download file package naming convention is now **[Project]_[Filename]_[ReleaseVersion]_GRCh[assembly].[format].gz**

PREVIOUS FORMAT	CURRENT FORMAT
CosmicCompleteCNA.GRCh37.99.tsv.gz	Cosmic_CompleteCNA_v99_GRCh37.tsv.gz
CosmicCLP_CompleteCNA.GRCh37.99.tsv.gz	CellLinesProject_CompleteCNA_v99_GRCh37.tsv.gz

- Final download file is now a tar file containing both the download file and its associated README

FILE PACKAGE (CURRENT)	FILE CONTENTS (CURRENT)
Cosmic_CompleteCNA_Tsv_v99_GRCh37.tar	Cosmic_CompleteCNA_v99_GRCh37.tsv.gz README_CompleteCNA_v99_GRCh37.txt

- **Cosmic_Fusion**: Added 5' and 3' gene symbol and Transcript accession for negatives. Also for positive data without coordinates.
- **CosmicMutantExport** has been deprecated. This is replaced by **Cosmic_GenomeScreensMutant** and **Cosmic_CompleteTargetedScreensMutant** (excluding negative data, meaning no genomic data in the MUTATION_GENOME_POSITION, mutation_cds, mutation_aa columns)
- **Cosmic_MutationTracking** now contains **all** legacy_mutation_ids. This file only contains mutations linked to released samples and studies to be consistent with other mutation files.
- **COSMIC_Genes** file now has formatted FASTA file using the Python Bio.SeqIO library

- COSO ID replaces classification data which connect via the classification file
- All files are now tsv.gz format with the exception of the .vcf
- **COSMICNCV.tsv.gz** is renamed to **Cosmic_NonCodingVariants_v99_GRCh37.tsv.gz**
- **Cosmic_NonCodingVariants_v99_GRCh37.vcf** now includes Complex - compound substitution (id_mut_type=29)
- **CosmicCodingMuts.vcf** is split into two files: **Cosmic_GenomeScreensMutant_v99_GRCh37.vcf** and **Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.vcf** Mutations with samples in both targeted and genome screens have been added to the genome screens file only to avoid duplication
- **CosmicHGNC** is replaced with **Cosmic_Gene**
- CLP files match COSMIC files (same column and naming formats)
- Chromosome 23,24 are now X,Y in all download files
- Chromosome 25 is replaced with MT in line with 23,24 becoming X,Y

KNOWN ISSUES WITH v99

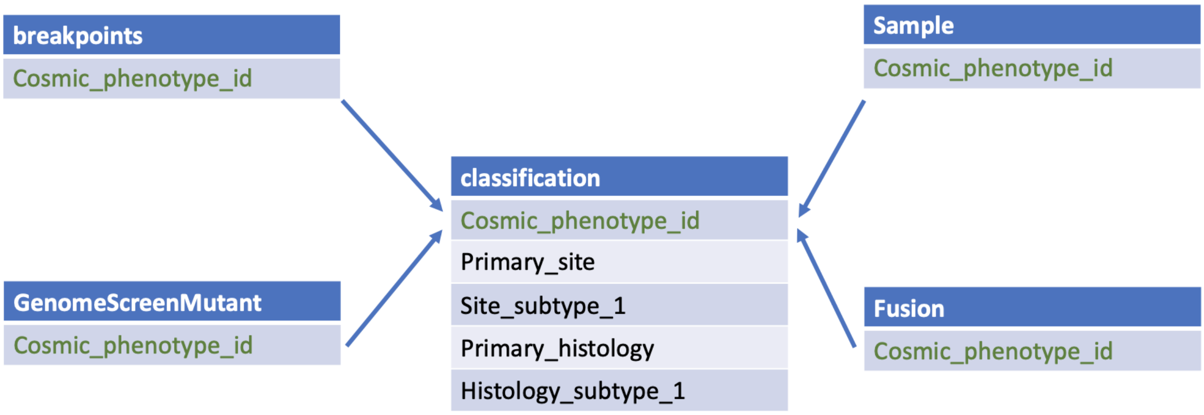
- Cosmic_Genes_v99 FASTA files contains the gene coordinates for each transcript instead of the associate transcript coordinates

DATA NOT PRESENT IN v99 DOWNLOADS

Current data omissions:

- **Institute, Institute_Address and Catalogue_Number** columns are not present in:
  CellLinesProject_CompleteTargetedScreensMutant_v99_GRCh37.tsv
  CellLinesProject_GenomeScreensMutant_v99_GRCh37.tsv
- Original paper classifications are currently missing from the classification download file

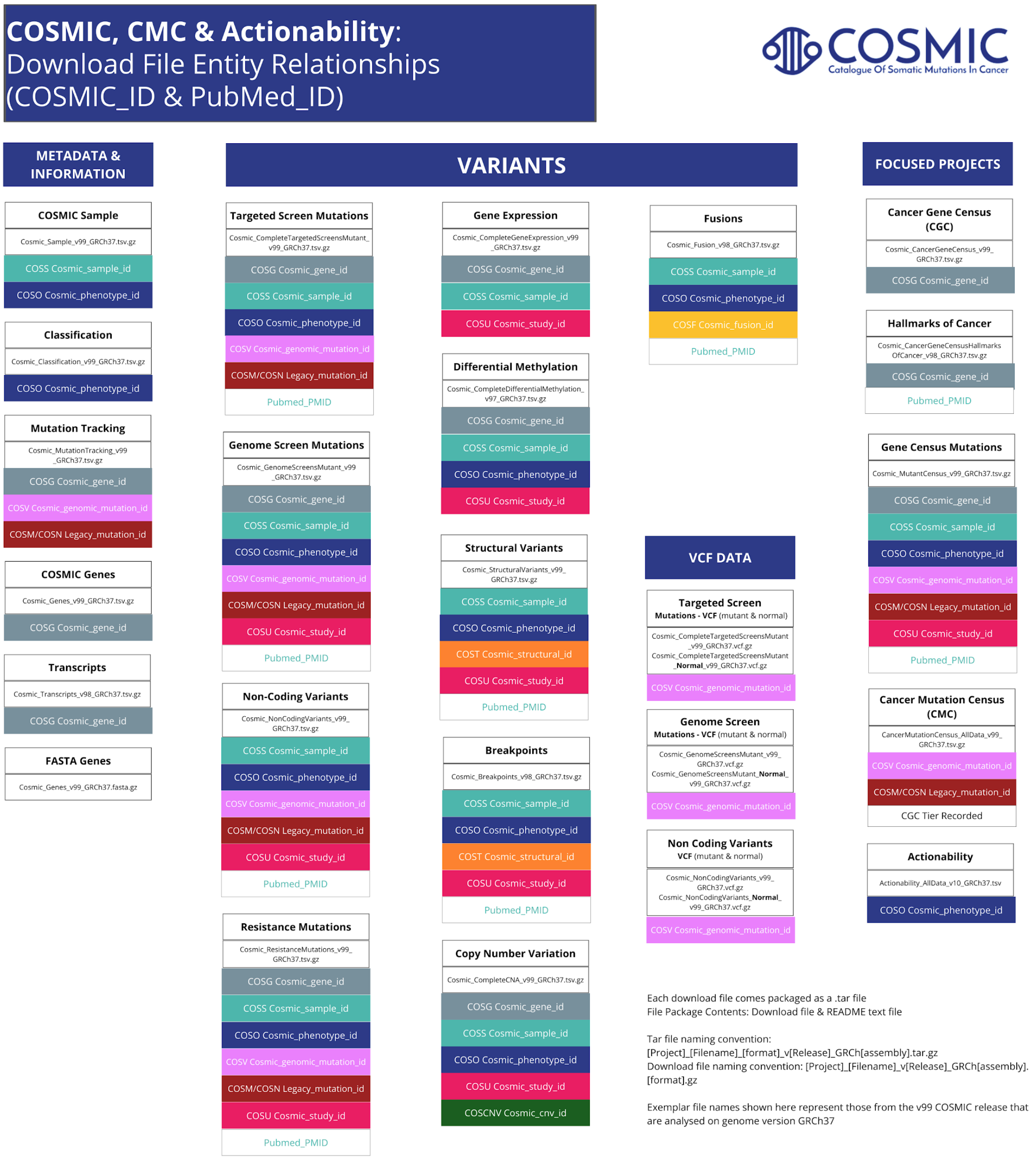
CONNECTING FILES WITH NEW IDENTIFIERS



LIST OF NEW IDENTIFIERS

Cosmic_phenotype_id COSO123	Cosmic_gene_id COSG123	Cosmic_sample_id COSS123
Cosmic_structural_id COST123	Cosmic_cnv_id COSCNV123	Cosmic_fusion_id COSF123
Cosmic_ncv_id COSN123	Cosmic_paper_id COSP123	Cosmic_study_id COSU123

COSMIC IDENTIFIER ENTITY RELATIONSHIP DIAGRAM



COSMIC, CMC & Actionability:
Download File Entity Relationships
(COSMIC_ID & PubMed_ID)

COSMIC Catalogue Of Somatic Mutations In Cancer

METADATA & INFORMATION

COSMIC Sample
Cosmic_Sample_v99_GRCh37.tsv.gz
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id

Classification
Cosmic_Classification_v99_GRCh37.tsv.gz
COSO Cosmic_phenotype_id

Mutation Tracking
Cosmic_MutationTracking_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id

COSMIC Genes
Cosmic_Genes_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id

Transcripts
Cosmic_Transcripts_v98_GRCh37.tsv.gz
COSG Cosmic_gene_id

FASTA Genes
Cosmic_Genes_v99_GRCh37.fasta.gz

VARIANTS

Targeted Screen Mutations
Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
Pubmed_PMID

Genome Screen Mutations
Cosmic_GenomeScreensMutant_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
COSU Cosmic_study_id
Pubmed_PMID

Non-Coding Variants
Cosmic_NonCodingVariants_v99_GRCh37.tsv.gz
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
COSU Cosmic_study_id
Pubmed_PMID

Resistance Mutations
Cosmic_ResistanceMutations_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
COSU Cosmic_study_id
Pubmed_PMID

Gene Expression
Cosmic_CompleteGeneExpression_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSU Cosmic_study_id

Differential Methylation
Cosmic_CompleteDifferentialMethylation_v97_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSU Cosmic_study_id

Structural Variants
Cosmic_StructuralVariants_v99_GRCh37.tsv.gz
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COST Cosmic_structural_id
COSU Cosmic_study_id
Pubmed_PMID

Breakpoints
Cosmic_Breakpoints_v98_GRCh37.tsv.gz
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COST Cosmic_structural_id
COSU Cosmic_study_id
Pubmed_PMID

Copy Number Variation
Cosmic_CompleteCNA_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSU Cosmic_study_id
COSCNV Cosmic_cnv_id

Fusions
Cosmic_Fusion_v98_GRCh37.tsv.gz
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSF Cosmic_fusion_id
Pubmed_PMID

VCF DATA

Targeted Screen
Mutations - VCF (mutant & normal)
Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.vcf.gz
Cosmic_CompleteTargetedScreensMutant_**Normal**_v99_GRCh37.vcf.gz
COSV Cosmic_genomic_mutation_id

Genome Screen
Mutations - VCF (mutant & normal)
Cosmic_GenomeScreensMutant_v99_GRCh37.vcf.gz
Cosmic_GenomeScreensMutant_**Normal**_v99_GRCh37.vcf.gz
COSV Cosmic_genomic_mutation_id

Non Coding Variants
VCF (mutant & normal)
Cosmic_NonCodingVariants_v99_GRCh37.vcf.gz
Cosmic_NonCodingVariants_**Normal**_v99_GRCh37.vcf.gz
COSV Cosmic_genomic_mutation_id

FOCUSED PROJECTS

Cancer Gene Census (CGC)
Cosmic_CancerGeneCensus_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id

Hallmarks of Cancer
Cosmic_CancerGeneCensusHallmarksOfCancer_v98_GRCh37.tsv.gz
COSG Cosmic_gene_id
Pubmed_PMID

Gene Census Mutations
Cosmic_MutantCensus_v99_GRCh37.tsv.gz
COSG Cosmic_gene_id
COSS Cosmic_sample_id
COSO Cosmic_phenotype_id
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
COSU Cosmic_study_id
Pubmed_PMID

Cancer Mutation Census (CMC)
CancerMutationCensus_AllData_v99_GRCh37.tsv.gz
COSV Cosmic_genomic_mutation_id
COSM/COSN Legacy_mutation_id
CGC Tier Recorded

Actionability
Actionability_AllData_v10_GRCh37.tsv
COSO Cosmic_phenotype_id

Each download file comes packaged as a .tar file
File Package Contents: Download file & README text file

Tar file naming convention:
[Project]_[Filename]_[format]_v[Release]_GRCh[assembly].tar.gz
Download file naming convention: [Project]_[Filename]_v[Release]_GRCh[assembly].[format].gz

Exemplar file names shown here represent those from the v99 COSMIC release that are analysed on genome version GRCh37

Published November-2023

NEW DOWNLOAD FILE CONTENT AND CHANGES

1) Cosmic_Classification_v99_GRCh37.tsv

File package

Cosmic_Classification_Tsv_v99_GRCh37.tar contains:

- Cosmic_Classification_v99_GRCh37.tsv.gz
- README_Cosmic_Classification_v99_GRCh37.txt

Main changes

- Format changed from CSV (comma separated) to TSV (tab separated)
- Paper original classification has been removed from the file to be in sync with website and other download files, the file size has consequently reduced

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
COSMIC_PHENOTYPE_ID	COSMIC_PHENOTYPE_ID	COSO id (tum_class_link.id_site_class + tum_class_link.id_hist_class)	COSO36286727
SITE_PRIMARY	PRIMARY_SITE		thyroid
SITE_SUBTYPE1	SITE_SUBTYPE_1		NS
SITE_SUBTYPE2	SITE_SUBTYPE_2		NS
SITE_SUBTYPE3	SITE_SUBTYPE_3		carcinoma
HISTOLOGY	PRIMARY_HISTOLOGY		papillary_carcinoma
HIST_SUBTYPE1	HISTOLOGY_SUBTYPE_1		papillary_carcinoma
HIST_SUBTYPE2	HISTOLOGY_SUBTYPE_2		follicular_variant
HIST_SUBTYPE3	HISTOLOGY_SUBTYPE_3		NS
SITE_PRIMARY_COSMIC			
SITE_SUBTYPE1_COSMIC			
SITE_SUBTYPE2_COSMIC			
SITE_SUBTYPE3_COSMIC			
HISTOLOGY_COSMIC			
HIST_SUBTYPE1_COSMIC			
HIST_SUBTYPE2_COSMIC			
HIST_SUBTYPE3_COSMIC			
NCI_CODE	NCI_CODE		C7381

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
COSMIC_PHENOTYPE_ID	COSMIC_PHENOTYPE_ID	COSO id (tum_class_link.id_site_class + tum_class_link.id_hist_class)	COSO36286727
SITE_PRIMARY	PRIMARY_SITE		thyroid
SITE_SUBTYPE1	SITE_SUBTYPE_1		NS
SITE_SUBTYPE2	SITE_SUBTYPE_2		NS
SITE_SUBTYPE3	SITE_SUBTYPE_3		carcinoma
EFO	EFO		http://www.ebi.ac.uk/efo/EFO_1000261

README file

```
---------------------------------
COSMIC Classification Information
---------------------------------
COSMIC cancer classification information in a tab separated file.  [ Cosmic_Classification_v99_GRCh37.tsv.gz ]

File Description
[column number:label] Heading                    Description
--------------------------------------------------------------------------------------------------------
[1:A]         COSMIC_PHENOTYPE_ID         A unique COSMIC identifier (COSO) for the classification. Other download
files can be linked to this file using this identifier
[2:B]         PRIMARY_SITE               Primary tissue specified in COSMIC
[3:C]         SITE_SUBTYPE_1              Sub tissue specified in COSMIC
[4:D]         SITE_SUBTYPE_2              Sub tissue specified in COSMIC
[5:E]         SITE_SUBTYPE_3              Sub tissue specified in COSMIC
[6:F]         PRIMARY_HISTOLOGY           Primary histology specified in COSMIC
[7:G]         HISTOLOGY_SUBTYPE_1         Sub histology specified in COSMIC
[8:H]         HISTOLOGY_SUBTYPE_2         Sub histology specified in COSMIC
[9:I]         HISTOLOGY_SUBTYPE_3         Sub histology specified in COSMIC.
[10:J]        NCI_CODE                    NCI thesaurus code for tumour histological classification. For details see
https://ncit.nci.nih.gov
[11:Q]        EFO                     Experimental Factor Ontology (EFO), for details see https://www.ebi.ac.uk/efo/
```

2) Cosmic_Transcripts_v99_GRCh37.tsv

File package

Cosmic_Transcripts_Tsv_v99_GRCh37.tar contains:

- Cosmic_Transcripts_v99_GRCh37.tsv.gz
- README_Cosmic_Transcripts_v99_GRCh37.txt

Main changes

- File now contains all the transcripts and a canonical and biotype flag
- File can be connected to the new CosmicGenes file using COSMIC_GENE_ID

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_ID			
TRANSCRIPT_ID	TRANSCRIPT_ACCESSION	Transcript accession+version	ENST00000394810.2
	COSMIC_GENE_ID	COSG+id_gene	COSG11842
GENE_NAME		Now in gene file	
STRAND	STRAND		-1
	BIOTYPE		protein_coding
	IS_CANONICAL		y

README file

COSMIC Transcripts

All transcript data in COSMIC is represented by a unique Ensembl transcript accession from the current release in a tab separated file. Transcripts are associated with a unique COSMIC gene id, strand, biotype and canonical flag. [Cosmic_Transcripts_v99_GRCh37.tsv.gz]

File Description
[column number:label] Heading Description

[1:A] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html
[2:B] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file
[3:C] STRAND Positive or negative (+1/-1)
[4:D] BIOTYPE Classification of genes and transcripts (protein coding, pseudogene, processed pseudogene, miRNA, rRNA, scRNA, snoRNA, snRNA.). More information: https://www.ensembl.org/Help/Faq?id=468
[5:E] IS_CANONICAL The Ensembl Canonical transcript is a single, representative transcript identified at every locus. For details see: https://www.ensembl.org/info/genome/genebuild/canonical.html

3) Cosmic_Genes_v99_GRCh37.tsv (replaces CosmicHGNC)

File package

Cosmic_Genes_Tsv_v99_GRCh37.tar contains:

- Cosmic_Genes_v99_GRCh37.tsv.gz
- README_Cosmic_Genes_v99_GRCh37.txt

Main changes

- Contains all the ENSG genes
- File previously called CosmicHGNC.tsv

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
COSMIC_ID	COSMIC_GENE_ID	COSG+id_gene	COSG42652
COSMIC_GENE_NAME	GENE_SYMBOL		BRCA2
	GENE_ACCESSION	ENSG + version	ENSG00000139618.10
ENTREZ_ID	ENTREZ_ID		675
HGNC_ID	HGNC_ID		1101
Mutated?		Value was always 'y'	
Cancer_census?	IN_CANCER_CENSUS		y
Expert_Curated?	IS_EXPERT_CURATED		y

README file

```
------------
COSMIC Genes
------------
 All the COSMIC gene data from the current release in a tab separated file. Genes are associated with COSMIC unique
gene identifier, gene symbol, Ensembl gene identifier, Entrez and HGNC mapping. [ Cosmic_Genes_v99_GRCh37.tsv.gz ]

 File Description
[column number:label] Heading                 Description
-----------------------------------------------------------------------------------------------------
[1:A]          COSMIC_GENE_ID          A unique COSMIC gene identifier (COSG) is used to identify a gene within the file.
Other download files can be linked to this file using this identifier.
[2:B]          GENE_SYMBOL             The gene name for which the data has been curated in COSMIC. In most cases this
is the accepted HGNC identifier.
[3:C]          GENE_ACCESSION          Unique Ensembl gene identifier (ENSG). For details see:
https://www.ensembl.org/info/genome/stable_ids/index.html
[4:D]          ENTREZ_ID               Entrez ID mapping
[5:E]          HGNC_ID                 HGNC mapping
[6:F]          IN_CANCER_CENSUS          is this gene part of the cancer census (y/n)
[7:G]          IS_EXPERT_CURATED         Has the gene been manually curated by the team of expert curators (y/n)
```

4) Cosmic_Genes_v99_GRCh37.fasta

File package

Cosmic_Genes_Fasta_v99_GRCh37.tar contains:

- Cosmic_Genes_v99_GRCh37.fasta.gz
- README_Cosmic_Genes_v99_GRCh37.txt

Main changes

- Now using Python Bio.SeqIO library
- Sequence is now uppercase
- Transcript accession now contains the version

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES
GENE_NAME	GENE_SYMBOL	
TRANSCRIPT_ID	TRANSCRIPT_ACCESSION	Transcript accession+version
CHROMOSOME	CHROMOSOME	
CHR_START	CHR_START	
CHR_END	CHR_END	
STRAND	STRAND	
TRANSCRIPT_CDS_SEQUENCE	TRANSCRIPT_CDS_SEQUENCE	

Example data:

```
>OR4F5 ENST00000335137.3 1:69091-70008(+)
ATGGTGACTGAATTCATTTTTCTGGGTCTCTCTGATTCTCAGGAACTCCAGACCTTCCTA
TTTATGTTGTTTTTTGTATTCTATGGAGGAATCGTGTTTGGAAACCTTCTTATTGTCATA
```

README file

```
------------------------
COSMIC Fasta File (genes)
------------------------
 CDS sequence for all the coding genes in COSMIC. [ Cosmic_Genes_v99_GRCh37.fasta ]

FASTA SEQUENCE HEADER
----------------------------------------------------------------------------------------------------
>GENE_SYMBOL TRANSCRIPT_ACCESSION CHROMOSOME:GENOME_START-GENOME_STOP(STRAND)
SEQUENCE
```

5) Cosmic_CancerGeneCensusHallmarksOfCancer_v99_GRCh37.tsv

File package

Cosmic_CancerGeneCensusHallmarksOfCancer_Tsv_v99_GRCh37.tar contains:

- Cosmic_CancerGeneCensusHallmarksOfCancer_v99_GRCh37.tsv.gz
- README_Cosmic_CancerGeneCensusHallmarksOfCancer_v99_GRCh37.txt

Main changes

- RenameD gene name column for consistency
- Added cosmic_gene_id to link to the Gene file

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_NAME	GENE_SYMBOL		ABI1
	COSMIC_GENE_ID	COSG+id_gene	COSG5120
CELL_TYPE	CELL_TYPE		hepatocellular carcinoma
PUBMED_PMID	PUBMED_PMID		28339046
HALLMARK	HALLMARK		role in cancer
IMPACT	IMPACT		oncogene
DESCRIPTION	DESCRIPTION		oncogene
CELL_LINE	CELL_LINE		HepG2 and MHCC97H

README file

```
---------------------------------------------
COSMIC Cancer Gene Census Hallmarks Of Cancer
---------------------------------------------
 A tab separated table listing the hallmarks of cancer for a subset of cancer census genes. [
Cosmic_CancerGeneCensusHallmarksOfCancer_v99_GRCh37.tsv.gz ]

  File Description

[column number:label] Heading                    Description
-----------------------------------------------------------------------------------------------------
[1:A]        GENE_SYMBOL          The gene name for which the data has been curated in COSMIC. In most cases this
is the accepted HGNC identifier
[2:B]        COSMIC_GENE_ID        A unique COSMIC gene identifier (COSG) is used to identify a gene within the file.
This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file
[3:C]        CELL_TYPE            Tissue or cancer for which the Hallmark is described
[4:D]        PUBMED_PMID          The PUBMED ID for the paper that the Hallmark was noted in
[5:E]        HALLMARK           Name of the biological process that when dysregulated, may promote cancer or
other data category describing the role of a gene in cancer
[6:F]        IMPACT           Describes how the gene activity impacts the hallmarks of cancer i.e.
promotes/suppresses or characterises the role of a gene in carcinogenesis i.e. Oncogene/Tumour suppressor
Gene/Fusion
[7:G]        DESCRIPTION          A brief functional summary of how gene's activity impacts a hallmark of cancer
[8:H]        CELL_LINE          For evidence obtained from experiments on cell lines, the name of the cell lines are
provided here.
```

6) Cosmic_Breakpoints_v99_GRCh37.tsv

File package

Cosmic_Breakpoints_Tsv_v99_GRCh37.tar contains:

- Cosmic_Breakpoints_v99_GRCh37.tsv.gz
- README_Cosmic_Breakpoints_v99_GRCh37.txt

Main changes

- Added new identifier ids to connect to sample, mutation, classification and study files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_NAME	SAMPLE_NAME		PD4107a
ID_SAMPLE	COSMIC_SAMPLE_ID	COSS + id_sample	COSS1317049
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28395278
MUTATION_ID	COSMIC_STRUCTURAL_ID	COST[ID_STRUCT_MUT]	COST25748
MUTATION_TYPE	MUTATION_TYPE		intrachromosomal tandem duplication
ID_TUMOUR			
PRIMARY_SITE		In classification file	
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
BREAKPOINT_ORDER		empty column	
GRCH		Removed since it's now in the file name	
CHROM_FROM	CHROM_FROM		22
LOCATION_FROM_MIN	LOCATION_FROM_MIN		29815139
LOCATION_FROM_MAX	LOCATION_FROM_MAX		29815139
STRAND_FROM	STRAND_FROM		-
CHROM_TO	CHROM_TO		22
LOCATION_TO_MIN	LOCATION_TO_MIN		30698769
LOCATION_TO_MAX	LOCATION_TO_MAX		30698769

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
STRAND_TO	STRAND_TO		-
NON_TEMPLATED_INS_SEQ	NON_TEMPLATED_INS_SEQ		CAG
PUBMED_PMID	PUBMED_PMID		22722201
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	COSU385

README file

COSMIC Breakpoints

All breakpoint data from the current release in a tab separated table. [Cosmic_Breakpoints_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] SAMPLE_NAME A sample is an instance of a portion of a tumour being examined for mutations. The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process. A number of samples can be taken from a single tumour and a number of tumours can be obtained from one individual. There can be multiple ids, if the same sample has been entered into the database multiple times from different papers.
[2:B] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional sample information from the Cosmic_Sample file.
[3:C] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file
[4:D] COSMIC_STRUCTURAL_ID A COSMIC structural identifier (COST). This identifier can be used to retrieve structural variants from the Cosmic_StructuralVariants file
[5:E] MUTATION_TYPE Type of mutation : Intra/Inter (chromosomal), tandem duplication, deletion, inversion, complex substitutions, complex amplicons.
[6:F] CHROM_FROM The chromosome where the first variant/breakpoint occurs.
[7:G] LOCATION_FROM_MIN The first position in breakpoint range.
[8:H] LOCATION_FROM_MAX The last position in breakpoint range.
[9:I] STRAND_FROM Positive or negative (+1/-1).
[10:J] CHROM_TO The chromosome where the last variant/breakpoint occurs.
[11:K] LOCATION_TO_MIN The first position in breakpoint range.
[12:L] LOCATION_TO_MAX The last position in breakpoint range.
[13:M] STRAND_TO Positive or negative (+1/-1).
[14:N] NON_TEMPLATED_INS_SEQ Non Templated Sequence (if any) which is inserted at the breakpoint. The sequence is not encoded.
[15:O] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in
[16:P] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this structural mutation

7) Cosmic_CancerGeneCensus_v99_GRCh37.tsv

File package

Cosmic_CancerGeneCensus_Tsv_v99_GRCh37.tar contains:

- Cosmic_CancerGeneCensus_v99_GRCh37.tsv.gz
- README_Cosmic_CancerGeneCensus_v99_GRCh37.txt

Main changes

- File format changed from CSV to TSV
- Added new cosmic_gene_id to be able to connect to the gene file
- Replaces yes/null with y/n for consistency

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_SYMBOL	GENE_SYMBOL		A1CF
NAME	NAME		APOBEC1 complementation factor
	COSMIC_GENE_ID	COSG+id_gene	COSG46891
ENTREZ_GENE_ID			
GENOME_LOCATION		Replaced with individual columns	
	CHROMOSOME		10
	GENOME_START		52559169
	GENOME_STOP		52645435
Hallmark		Can be retrieved from the Hallmarks file using COSG id.	Yes
CHR_BAND	CHR_BAND		11.23
SOMATIC	SOMATIC		y
GERMLINE	GERMLINE		n
TUMOUR_TYPES_SOMATIC	TUMOUR_TYPES_SOMATIC		melanoma
TUMOUR_TYPES_GERMLINE	TUMOUR_TYPES_GERMLINE		
CANCER_SYNDROME	CANCER_SYNDROME		
TISSUE_TYPE	TISSUE_TYPE		E
MOLECULAR_GENETICS	MOLECULAR_GENETICS		
ROLE_IN_CANCER	ROLE_IN_CANCER		Oncogene
MUTATION_TYPES	MUTATION_TYPES		Mis
TRANSLOCATION_PARTNER	TRANSLOCATION_PARTNER		
OTHER_GERMLINE_MUT	OTHER_GERMLINE_MUT		n
OTHER_SYNDROME	OTHER_SYNDROME		
COSMIC ID			
TIER	TIER		2
COSMIC_GENE_NAME			
SYNONYMS	SYNONYMS		A1CF,ENSG00000148584.10,Q9NQ94,29974,ACF,ACF64,ACF65,APOBEC1CF,ASP

README file

COSMIC Cancer Gene Census

A list of all cancer census genes from the current release in a comma separated table. The census table is exported from https://cancer.sanger.ac.uk/census and the format is the same. [Cosmic_CancerGeneCensus_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.

[2:B] NAME Gene descriptive name.

[3:C] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.

[4:D] CHROMOSOME The chromosome location of a given mutation census (1-22, X, Y or MT).

[5:E] GENOME_START The start coordinate of a given mutation census.

[6:F] GENOME_STOP The end coordinate of a given mutation census.

[7:G] CHR_BAND Chromosome (1-22, X, Y or MT), arm (p or q) and cytogenetic band.

[8:H] SOMATIC Somatic mutations have been detected (y/n).

[9:I] GERMLINE Germline mutations have been detected (y/n).

[10:J] TUMOUR_TYPES_SOMATIC Somatic mutations in the gene are associated with the following diseases (see abbreviations tab for details: https://cancer.sanger.ac.uk/cosmic/help/census#abbrev).

[11:K] TUMOUR_TYPES_GERMLINE Germline mutations in the gene are associated with the following diseases (see abbreviations tab for details: https://cancer.sanger.ac.uk/cosmic/help/census#abbrev)

[12:L] CANCER_SYNDROME Syndrome associated with germline mutation.

[13:M] TISSUE_TYPE Type of tissue, see abbreviations tab for details: https://cancer.sanger.ac.uk/cosmic/help/census#abbrev.

[14:N] MOLECULAR_GENETICS See abbreviations tab for details: https://cancer.sanger.ac.uk/cosmic/help/census#abbrev.

[15:O] ROLE_IN_CANCER Role in Cancer: oncogene: hyperactivity of the gene drives the transformation; TSG: loss of gene function drives the transformation. Some genes can play either of these roles depending on cancer type. Fusion: the gene is known to be involved in oncogenic fusions.

[16:P] MUTATION_TYPES Types of mutation: See abbreviations tab for details: https://cancer.sanger.ac.uk/cosmic/help/census#abbrev

[17:Q] TRANSLOCATION_PARTNER Gene symbol of fusion partner

[18:R] OTHER_GERMLINE_MUT Other germline mutations not implicated in cancer

[19:S] OTHER_SYNDROME Other non-cancerous syndrome

[20:T] TIER Indicates to which tier of the Cancer Gene Census the gene belongs (1/2)

[21:U] SYNONYMS Gene alternative names

8) Cosmic_CompleteCNA_v99_GRCh37.tsv

File package

Cosmic_CompleteCNA_Tsv_v99_GRCh37.tar contains:

- Cosmic_CompleteCNA_v99_GRCh37.tsv.gz
- README_Cosmic_CompleteCNA_v99_GRCh37.txt

Main changes

- Data directly linked to gene instead of Transcript, file is now smaller as a result
- Added new identifier ids to connect to sample, Gene, classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
CNV_ID	COSMIC_CNV_ID	COSCNV[CNA_ID]	COSCNV2372777
	COSMIC_GENE_ID	COSG+id_gene	COSG17603
	GENE_SYMBOL		SOX13
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS + ctso.id_sample	COSS1337807
SAMPLE_NAME	SAMPLE_NAME		TCGA-02-2470-01
ID_TUMOUR		Data in classification file	
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
	COSMIC_PHENOTYPE	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28245232
TOTAL_CN	TOTAL_CN		19
MINOR_ALLELE	MINOR_ALLELE		1
MUT_TYPE	MUT_TYPE		gain
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	COSU329
GRCH		Remove GRCh since it's in the file name	
CHROMOSOME_G_START_G_STOP		Remove concatenation	
	CHROMOSOME		1
	GENOME_START		203921668
	GENOME_STOP		205128958
TRANSCRIPT_ACCESSION			

README file

COSMIC Copy Number Variants

All copy number variants from the current release in a tab separated table. For more information on copy number data, please see https://cancer.sanger.ac.uk/cosmic/help/cnv/overview. [Cosmic_CompleteCNA_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_CNV_ID A Copy number variant identifier (COSCNV) is used to identify the copy number variants within the file.
[2:B] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.
[3:C] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[4:D] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[5:E] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process..
[6:F] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[7:G] TOTAL_CN The sum of the major and minor allele counts e.g. if ABB, total copy number = 3.
[8:H] MINOR_ALLELE The number of copies of the least frequent allele e.g. if ABB, minor allele = A (1 copy) and major allele = B (2 copies).
[9:I] MUT_TYPE Defined as Gain or Loss. For ICGC samples; as defined in the original data. For TCGA samples reanalysed with ASCAT -
* LOSS = average genome ploidy <= 2.7 AND total copy number = 0 OR average genome ploidy > 2.7 AND total copy number < (average genome ploidy - 2.7)
* GAIN = average genome ploidy <= 2.7 AND total copy number >= 5 OR average genome ploidy > 2.7 AND total copy number >= 9
[10:J] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this copy number variation.
[11:K] CHROMOSOME The chromosome location of a given copy number variant (1-22, X, Y or MT)
[12:L] GENOME_START The start coordinate of a given copy number variant
[13:M] GENOME_STOP The end coordinate of a given copy number variant

9) Cosmic_CompleteDifferentialMethylation_v99_GRCh37.tsv

File package

Cosmic_CompleteDifferentialMethylation_Tsv_v99_GRCh37.tar contains:

- Cosmic_CompleteDifferentialMethylation_v99_GRCh37.tsv.gz
- README_Cosmic_CompleteDifferentialMethylation_v99_GRCh37.txt

Main changes

- Data directly linked to gene instead of Transcript, file is now smaller as a result
- Added new identifier ids to connect to sample, Gene, study, classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
	COSMIC_GENE_ID	COSG+id_gene	COSG15191
STUDY_ID	COSMIC_STUDY_ID	COSU + study_id	COSU376
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS1651254
SAMPLE_NAME	SAMPLE_NAME		TCGA-D5-6536-01
ID_TUMOUR			
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
FRAGMENT_ID	FRAGMENT_ID		cg07802401
GENOME_VERSION		Now in file name	
CHROMOSOME	CHROMOSOME		11
POSITION	POSITION		26354057
STRAND	STRAND		-1
GENE_NAME	GENE_SYMBOL		ANO3
METHYLATION	METHYLATION		L
AVG_BETA_VALUE_NORMAL	AVG_BETA_VALUE_NORMAL		0.682
BETA_VALUE	BETA_VALUE		0.159
TWO_SIDED_P_VALUE	TWO_SIDED_P_VALUE		0.0000000218275988395
ACCESSION_NUMBER			
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28694826

README file

COSMIC Methylation

TCGA Level 3 methylation data from the ICGC portal for the current release in a tab separated table. More information on the methylation data is available from https://cancer.sanger.ac.uk/cosmic/analyses. [Cosmic_CompleteDifferentialMethylation_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this methylation data.

[2:B] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.

[3:C] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[4:D] FRAGMENT_ID The unique probe Id for a specific CpG.
[5:E] CHROMOSOME The chromosome location of the probe (1-22, X or Y).
[6:F] POSITION The genome location of the CpG targeted by the probe (1-based coordinates).
[7:G] STRAND positive or negative (+1/-1).
[8:H] GENE_SYMBOL The gene name (if the probe falls within the coding region of a COSMIC gene) or the probe annotation as described by Illumina.
[9:I] METHYLATION The methylation level; H (High, beta-value >0.8) or L (Low, beta-value < 0.2).
[10:J] AVG_BETA_VALUE_NORMAL The average beta-value across the normal population. The beta-value of the tumour must differ from this value by >0.5 to be considered a variant.
[11:K] BETA_VALUE The beta-value for the probe in the tumour sample. Only values >0.8 (High) or <0.2 (Low) are included.
[12:L] TWO_SIDED_P_VALUE The two sided p-value.
[13:M] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.

10) Cosmic_CompleteGeneExpression_v99_GRCh37.tsv

File package

Cosmic_CompleteGeneExpression_Tsv_v99_GRCh37.tar contains:

- Cosmic_CompleteGeneExpression_v99_GRCh37.tsv.gz
- README_Cosmic_CompleteGeneExpression_v99_GRCh37.txt

Main changes

- Data directly linked to gene instead of Transcript, file is now smaller as a result
- Better gene name coverage
- Fixed wrong gene name mapping
- Added new identifier ids to connect to sample, Gene, study files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_ID	COSMIC_SAMPLE_ID	COSS + id_sample	COSS1337808
SAMPLE_NAME	SAMPLE_NAME		TCGA-02-2483-01
	COSMIC_GENE_ID	COSG+id_gene	COSG483
GENE_NAME	GENE_SYMBOL		ALG14
REGULATION	REGULATION		normal
Z_SCORE	Z_SCORE		0.282
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	COSU329
ACCESSION_NUMBER			

README file

COSMIC Complete Gene Expression

All gene expression level 3 data from the TCGA portal for the current release in a tab separated table. Please note : The platform codes currently used to produce the COSMIC gene expression values are: IlluminaGA_RNASeqV2, IlluminaHiSeq_RNASeqV2, AgilentG4502A_07_2, AgilentG4502A_07_3. For more information on the gene expression data, please see https://cancer.sanger.ac.uk/cosmic/analyses. [Cosmic_CompleteGeneExpression_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[2:B] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[3:C] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.
[4:D] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[5:E] REGULATION The regulation can be over or under depending on the scores from different platforms if they are above or below the threshold.
[6:F] Z_SCORE z_score serves as an indicative score taken from the gene_expression from different platforms in order of preference: IlluminaHiSeq_RNASeqV2, IlluminaGA_RNASeqV2, AgilentG4502A_07_3.
[7:G] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this gene expression data.

11) Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.tsv

File package

Cosmic_CompleteTargetedScreensMutant_Tsv_v99_GRCh37.tar contains:

- Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.tsv.gz
- README_Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.txt

Main changes

- File similar in size and content to current file
- Added new identifier ids to connect to sample, Gene, study, mutation tracking and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_NAME	GENE_SYMBOL		GEN1

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
	COSMIC_GENE_ID	COSG+id_gene	COSG47494
ACCESSION_NUMBER	ACCESSION_NUMBER		ENST00000317402.7
GENE_CDS_LENGTH			
HGNC_ID		In gene file	
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS1235084
SAMPLE_NAME	SAMPLE_NAME		HCC2157
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28395278
ID_TUMOUR			
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
GENOME_WIDE_SCREEN		no point. All genome_wide_screen are y	
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV58058865
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSM33318
MUTATION_ID	MUTATION_ID		26016977
MUTATION_CDS	MUTATION_CDS		c.824G>T
MUTATION_AA	MUTATION_AA		p.R275L
MUTATION_DESCRIPTION	MUTATION_DESCRIPTION		missense_variant
MUTATION_ZYGOSITY	MUTATION_ZYGOSITY		het
LOH	LOH		
GRCH			
MUTATION_GENOME_POSITION		Replace with genome_start, genome_end, chromosome	
CHROMOSOME	CHROMOSOME		2
	GENOME_START		17953922
	GENOME_END		17953922
MUTATION_STRAND	STRAND		+
SNP		Gnomad score? – new column from cmc?	
RESISTANCE_MUTATION			
FATHMM_PREDICTION			
FATHMM_SCORE			

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
MUTATION_SOMATIC_STATUS			
PUBMED_PMID	PUBMED_PMID		16959974
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	
SAMPLE_TYPE			
TUMOUR_ORIGIN			
AGE			
HGVSP	HGVSP		ENSP00000318977.7:p.Arg275Leu
HGVSC	HGVSC		ENST00000317402.7:c.824G>T
HGVSG	HGVSG		2:g.17953922G>T
	GENOMIC_WT_ALLELE		G
	GENOMIC_MUT_ALLELE		T

README file

COSMIC Complete Mutation Data (Targeted Screens)

A tab separated table of the complete curated COSMIC dataset (targeted screens) from the current release. It includes all coding point mutations, and the negative data set. [Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.tsv.gz]
The Cosmic_Mutant file can be re-created by linking the Cosmic_GenomeScreensMutant with the positive data (data with mutation ids) from this file Cosmic_CompleteTargetedScreensMutant

File Description

[column number:label] Heading Description

[1:A] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[2:B] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.
[3:C] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.
[4:D] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[5:E] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[6:F] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[7:G] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[8:H] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.

[9:I] MUTATION_ID An internal mutation identifier to uniquely represent each mutation on a specific transcript on a given assembly build. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.

[10:J] MUTATION_CDS The change that has occurred in the nucleotide sequence. Formatting is identical to the method used for the peptide sequence.

[11:K] Mutation_AA The change that has occurred in the peptide sequence. Formatting is based on the recommendations made by the Human Genome Variation Society. The description of each type can be found by following the link to the Mutation Overview page.

[12:L] MUTATION_DESCRIPTION Type of mutation at the amino acid level (substitution, deletion, insertion, complex, fusion, unknown etc.).

[13:M] MUTATION_ZYGOSITY Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample.

[14:N] LOH LOH Information on whether the gene was reported to have loss of heterozygosity in the sample: yes, no or unknown.

[15:O] CHROMOSOME The chromosome location of a given targeted screen (1-22, X, Y or MT).

[16:P] GENOME_START The start coordinate of a given targeted screen.

[17:Q] GENOME_STOP The end coordinate of a given targeted screen.

[18:R] STRAND Positive or negative (+/-).

[19:S] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication.

[20:T] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this sample.

[21:U] HGVSP Human Genome Variation Society peptide syntax.

[22:V] HGVSC Human Genome Variation Society coding dna sequence syntax (CDS).

[23:W] HGVSG Human Genome Variation Society genomic syntax (3' shifted).

[24:X] GENOMIC_WT_ALLELE Genomic Wild type allele sequence.

[25:Y] GENOMIC_MUT_ALLELE Genomic mutation allele sequence.

[26:Z] MUTATION_SOMATIC_STATUS Information on whether the sample was reported to be Confirmed somatic variant, Reported in another cancer sample as somatic or Variant of unknown origin:

* Reported in another cancer sample as somatic = when the mutation has been reported as somatic previously but not in current paper

* Confirmed somatic variant = if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient

* Variant of unknown origin = When the tumour has been sequenced without a matched normal tissue from the same individual, the somatic status of the variant cannot be assessed

12) Cosmic_Fusion_v99_GRCh37.tsv

File package

Cosmic_Fusion_Tsv_v99_GRCh37.tar contains:

- Cosmic_Fusion_v99_GRCh37.tsv.gz
- README_Cosmic_Fusion_v99_GRCh37.txt

Main changes

- Added negative fusion data.These are samples tested but where no fusion was detected. Users have to cross reference with the complete mutation file (targeted) to find out what the negative samples were tested against. The mutation file lists one gene tested rather than the gene pair
- Added Gene symbol and Transcript accession for 5'/3' gene pair for negative data and positive data without coordinates
- Added new identifier ids to connect to sample and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_ID	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS1000017
SAMPLE_NAME	SAMPLE_NAME		1000017
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO36286727
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
FUSION_ID	COSMIC_FUSION_ID	include prefix COSF	COSF1271
TRANSLOCATION_NAME	FUSION_SYNTAX		ENST00000263102.6(CCDC6):r.1_535::ENST00000355710.3(RET):r.2369_5659
5'_CHROMOSOME	FIVE_PRIME_CHROMOSOME	Removed special chars to make file processing easier	10
5'_STRAND	FIVE_PRIME_STRAND		-
5'_GENE_ID	FIVE_PRIME_TRANSCRIPT_ACCESSION		ENST00000263102.6
5'_GENE_NAME	FIVE_PRIME_GENE_SYMBOL		CCDC6
5'_LAST_OBSERVED_EXON	FIVE_PRIME_LAST_OBSERVE_EXON		1
5'_GENOME_START_FROM	FIVE_PRIME_GENOME_START_FROM		61665880
5'_GENOME_START_TO	FIVE_PRIME_GENOME_START_TO		61665880
5'_GENOME_STOP_FROM	FIVE_PRIME_GENOME_STOP_FROM		61666414
5'_GENOME_STOP_TO	FIVE_PRIME_GENOME_STOP_TO		61666414
3'_CHROMOSOME	THREE_PRIME_CHROMOSOME		10
3'_STRAND	THREE_PRIME_STRAND		+
3'_GENE_ID	THREE_PRIME_TRANSCRIPT_ACCESSION		ENST00000355710.3
3'_GENE_NAME	THREE_PRIME_GENE_SYMBOL		RET
3'_FIRST_OBSERVED_EXON	THREE_PRIME_FIRST_OBSERVE_EXON		12
3'_GENOME_START_FROM	THREE_PRIME_GENOME_START_FROM		43612032
3'_GENOME_START_TO	THREE_PRIME_GENOME_START_TO		43612032

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
3'_GENOME_STOP_FROM	THREE_PRIME_GENOME_STOP_FROM		43625799
3'_GENOME_STOP_TO	THREE_PRIME_GENOME_STOP_TO		43625799
FUSION_TYPE	FUSION_TYPE		Observed mRNA
PUBMED_PMID	PUBMED_PMID		16784981

README file

COSMIC Fusion

All gene fusion mutation data from the current release in a tab separated table. This file includes all the tested samples, with and without fusion detected. [Cosmic_Fusion_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[2:B] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[3:C] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[4:D] COSMIC_FUSION_ID A fusion mutation identifier (COSF). This identifier can be null for samples tested but where no fusion was detected.
[5:E] FUSION_SYNTAX Syntax describing the portions of mRNA present (in HGVS 'r.' format) from each gene (allows representation of UTR sequences).
[6:F] FIVE_PRIME_CHROMOSOME Chromosome of 5’ gene.
[7:G] FIVE_PRIME_STRAND Positive or negative of the 5’ gene (+/-).
[8:H] FIVE_PRIME_TRANSCRIPT_ID The Ensembl Transcript identifier (ENST) of the 5' gene. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.
[9:I] FIVE_PRIME_GENE_SYMBOL Gene symbol for the 5' gene fusion partner for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[10:J] FIVE_PRIME_LAST_OBSERVE_EXON Last observed exon number of the 5' gene fusion partner.
[11:K] FIVE_PRIME_GENOME_START_FROM The genomic coordinate of the start (+ strand)/breakpoint (- strand) of the 5’ fusion gene as described in the fusion syntax.
[12:L] FIVE_PRIME_GENOME_START_TO The range of genomic coordinates of the start (+ strand)/breakpoint (- strand) of the 5’ fusion gene if it is an unknown base position.
[13:M] FIVE_PRIME_GENOME_STOP_FROM The genomic coordinate of the breakpoint (+ strand)/start (- strand) of the 5’ fusion gene as described in the Translocation Name.
[14:N] FIVE_PRIME_GENOME_STOP_TO The range of genomic coordinates of the breakpoint (+ strand)/start (- strand) of the 5’ fusion gene if it is an unknown base position.
[15:O] THREE_PRIME_CHROMOSOME Chromosome of 3’ gene.
[16:P] THREE_PRIME_STRAND Positive or negative of the 3’ gene (+/-).
[17:Q] THREE_PRIME_TRANSCRIPT_ID The Ensembl Transcript identifier (ENST) of the 3' gene. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.
[18:R] THREE_PRIME_GENE_SYMBOL Gene symbol for the 3' gene fusion partner for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[19:S] THREE_PRIME_FIRST_OBSERVE_EXON First observed exon number of the 3' gene fusion partner.
[20:T] THREE_PRIME_GENOME_START_FROM The genomic coordinate of the breakpoint (+ strand)/stop (- strand) of the 3’ fusion gene as described in the Translocation Name.

[21:U] THREE_PRIME_GENOME_START_TO The range of genomic coordinates of the breakpoint (+ strand)/stop (- strand) of the 3' fusion gene if it is an unknown base position.
[22:V] THREE_PRIME_GENOME_STOP_FROM The genomic coordinate of the stop (+ strand)/breakpoint (- strand) of the 3' fusion gene as described in the Translocation Name.
[23:W] THREE_PRIME_GENOME_STOP_TO The range of genomic coordinates of the stop (+ strand)/breakpoint (- strand) of the 3' fusion gene if it is an unknown base position.
[24:X] FUSION_TYPE Type of mutation.
[25:Y] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in.

13) Cosmic_GenomeScreensMutant_v99_GRCh37.tsv

File package

Cosmic_GenomeScreensMutant_Tsv_v99_GRCh37.tar contains:

- Cosmic_GenomeScreensMutant_v99_GRCh37.tsv.gz
- README_Cosmic_GenomeScreensMutant_v99_GRCh37.txt

Main changes

- Added new identifier ids to connect to sample, Gene, study, mutation tracking and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_NAME	GENE_SYMBOL		ZSCAN22
	COSMIC_GENE_ID	COSG+id_gene	COSG40135
ACCESSION_NUMBER	TRANSCRIPT_ACCESSION	Added version	ENST00000329665.4
GENE_CDS_LENGTH			
HGNC_ID			
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS1651625
SAMPLE_NAME	SAMPLE_NAME		TCGA-EI-6882-01
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28664826
ID_TUMOUR			
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
GENOME_WIDE_SCREEN		All genome_wide_screen are y	
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV61639233
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSM3423316
MUTATION_ID	MUTATION_ID		25675684
MUTATION_CDS	MUTATION_CDS		c.102C>T
MUTATION_AA	MUTATION_AA		p.G34=
MUTATION_DESCRIPTION	MUTATION_DESCRIPTION		synonymous_variant
MUTATION_ZYGOSITY	MUTATION_ZYGOSITY		
LOH	LOH		
GRCH		Now in the file name	
MUTATION_GENOME_POSITION		Replace with genome_start, genome_end, chromosome	
CHROMOSOME	CHROMOSOME		19
	GENOME_START		58846270
	GENOME_END		58846270
MUTATION_STRAND	STRAND		+
SNP		Gnomad score? – new column from cmc?	
RESISTANCE_MUTATION			
FATHMM_PREDICTION			
FATHMM_SCORE			
MUTATION_SOMATIC_STATUS			
PUBMED_PMID	PUBMED_PMID		
ID_STUDY	COSMIC_STUDY_ID	COSU+ study_id	COSU375
SAMPLE_TYPE			
TUMOUR_ORIGIN			
AGE			
HGVSP	HGVSP		ENSP00000332433.3:p.Gly34=
HGVSC	HGVSC		ENST00000329665.4:c.102C>T
HGVSG	HGVSG		19:g.58846270C>T
	GENOMIC_WT_SEQ		C
	GENOMIC_MUT_SEQ		T

README file

COSMIC Mutation Data (Genome Screens)

A tab separated table of coding point mutations from genome wide screens (including whole exome sequencing) from the current release. [Cosmic_GenomeScreensMutant_v99_GRCh37.tsv.gz]
The Cosmic_Mutant file can be re-created by linking this file Cosmic_GenomeScreensMutant with the positive data (data with mutation ids) from the Cosmic_CompleteTargetedScreensMutant file

File Description

[column number:label] Heading Description

[1:A] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[2:B] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.
[3:C] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.
[4:D] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[5:E] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[6:F] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[7:G] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[8:H] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.
[9:I] MUTATION_ID An internal mutation identifier to uniquely represent each mutation on a specific transcript on a given assembly build. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[10:J] MUTATION_CDS The change that has occurred in the nucleotide sequence. Formatting is identical to the method used for the peptide sequence.
[11:K] Mutation_AA The change that has occurred in the peptide sequence. Formatting is based on the recommendations made by the Human Genome Variation Society. The description of each type can be found by following the link to the Mutation Overview page.
[12:L] MUTATION_DESCRIPTION Type of mutation at the amino acid level (substitution, deletion, insertion, complex, fusion, unknown etc.).
[13:M] MUTATION_ZYGOSITY Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample.
[14:N] LOH LOH Information on whether the gene was reported to have loss of heterozygosity in the sample: yes, no or unknown.
[15:O] CHROMOSOME The chromosome location of a given genome screen (1-22, X, Y or MT).
[16:P] GENOME_START The start coordinate of a given genome screen.
[17:Q] GENOME_STOP The end coordinate of a given genome screen.
[18:R] STRAND Positive or negative (+/-).
[19:S] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication.
[20:T] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this sample.
[21:U] HGVSP Human Genome Variation Society peptide syntax.
[22:V] HGVSC Human Genome Variation Society coding dna sequence syntax (CDS).
[23:W] HGVSG Human Genome Variation Society genomic syntax (3' shifted).
[24:X] GENOMIC_WT_ALLELE Genomic Wild type allele sequence.
[25:Y] GENOMIC_MUT_ALLELE Genomic mutation allele sequence.

[26:Z] MUTATION_SOMATIC_STATUS Information on whether the sample was reported to be Confirmed somatic variant, Reported in another cancer sample as somatic or Variant of unknown origin:
* Reported in another cancer sample as somatic = when the mutation has been reported as somatic previously but not in current paper
* Confirmed somatic variant = if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient
* Variant of unknown origin = When the tumour has been sequenced without a matched normal tissue from the same individual, the somatic status of the variant cannot be assessed

14) Cosmic_MutantCensus_v99_GRCh37.tsv

File package

Cosmic_MutantCensus_Tsv_v99_GRCh37.tar contains:

- Cosmic_MutantCensus_v99_GRCh37.tsv.gz
- README_Cosmic_MutantCensus_v99_GRCh37.txt

Main changes

- File similar in size and content to current file
- Added new identifier ids to connect to gene, sample and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_NAME	GENE_SYMBOL		ALDH2
	COSMIC_GENE_ID	COSG+id_gene	COSG55681
ACCESSION_NUMBER	TRANSCRIPT_ACCESSION	Added version	ENST00000261733.2
GENE_CDS_LENGTH			
HGNC_ID			
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS2658236
SAMPLE_NAME	SAMPLE_NAME		T207430
ID_TUMOUR			
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO28864826
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
HISTOLOGY_SUBTYPE_3			
GENOME_WIDE_SCREEN		removed, connect to sample	
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV55665914
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSM6598241
MUTATION_ID	MUTATION_ID		20830979
MUTATION_CDS	MUTATION_CDS		c.1366G>A
MUTATION_AA	MUTATION_AA		p.A456T
MUTATION_DESCRIPTION	MUTATION_DESCRIPTION		missense_variant
MUTATION_ZYGOSITY	MUTATION_ZYGOSITY		
LOH	LOH		
TIER		can be fetched from census file	
GRCH		In file name	
MUTATION_GENOME_POSITION		remove concatenation	
CHROMOSOME	CHROMOSOME		12
	GENOME_START		112237827
	GENOME_END		112237827
MUTATION_STRAND	STRAND		+
SNP			
RESISTANCE_MUTATION		it can be part of the clinical phase 2/3 as this info doesn't exist in the curation database, also this can be fetched from resistance mut file	
FATHMM_PREDICTION			
FATHMM_SCORE			
MUTATION_SOMATIC_STATUS		always null	
PUBMED_PMID	PUBMED_PMID		27149842
ID_STUDY	COSMIC_STUDY_ID	COSU + id_Study	
SAMPLE_TYPE		remove it as this data can be fetched it from sample file	
TUMOUR_ORIGIN		remove it as this data can be fetched it from sample file	
AGE			
HGVSP	HGVSP		ENSP00000261733.2:p.Ala456Thr
HGVSC	HGVSC		ENST00000261733.2:c.1366G>A
HGVSG	HGVSG		12:g.112237827G>A
	GENOMIC_WT_ALLELE		G
	GENOMIC_MUT_ALLELE		A

README file

COSMIC Mutations Census Genes

All coding mutations in genes listed in the Cancer Gene Census (https://cancer.sanger.ac.uk/census) in a tab separated table. [Cosmic_MutantCensus_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.

[2:B] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.

[3:C] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.

[4:D] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.

[5:E] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.

[6:F] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.

[7:G] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.

[8:H] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.

[9:I] MUTATION_ID An internal mutation identifier to uniquely represent each mutation on a specific transcript on a given assembly build. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.

[10:J] MUTATION_CDS The change that has occurred in the nucleotide sequence. Formatting is identical to the method used for the peptide sequence.

[11:K] Mutation_AA The change that has occurred in the peptide sequence. Formatting is based on the recommendations made by the Human Genome Variation Society. The description of each type can be found by following the link to the Mutation Overview page.

[12:L] MUTATION_DESCRIPTION Type of mutation at the amino acid level (substitution, deletion, insertion, complex, fusion, unknown etc.).

[13:M] MUTATION_ZYGOSITY Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample.

[14:N] LOH LOH Information on whether the gene was reported to have loss of heterozygosity in the sample: yes, no or unknown.

[15:O] CHROMOSOME The chromosome location of a given mutation census (1-22, X, Y or MT).

[16:P] GENOME_START The start coordinate of a given mutation census.

[17:Q] GENOME_STOP The end coordinate of a given mutation census.

[18:R] STRAND Positive or negative (+/-).

[19:S] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication.

[20:T] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this sample.

[21:U] HGVSP Human Genome Variation Society peptide syntax.

[22:V] HGVSC Human Genome Variation Society coding dna sequence syntax (CDS).

[23:W] HGVSG Human Genome Variation Society genomic syntax (3' shifted).

[24:X] GENOMIC_WT_ALLELE Genomic Wild type allele sequence.

[25:Y] GENOMIC_MUT_ALLELE Genomic mutation allele sequence.

[26:Z] MUTATION_SOMATIC_STATUS Information on whether the sample was reported to be Confirmed somatic variant, Reported in another cancer sample as somatic or Variant of unknown origin:

* Reported in another cancer sample as somatic = when the mutation has been reported as somatic previously but not in current paper

* Confirmed somatic variant = if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient

* Variant of unknown origin = When the tumour has been sequenced without a matched normal tissue from the same individual, the somatic status of the variant cannot be assessed

15) Cosmic_MutationTracking_v99_GRCh37.tsv

File package

Cosmic_MutationTracking_Tsv_v99_GRCh37.tar contains:

- Cosmic_MutationTracking_v99_GRCh37.tsv.gz
- README_Cosmic_MutationTracking_v99_GRCh37.txt

Main changes

- CosmicMutationTracking now contains all the legacy_mutation_id instead of a representative (minimum ids between multiple) and also non-coding mutations. File also only contains mutations linked to released samples and studies to be consistent with other mutation files.
- Added new identifier ids to connect to Gene and mutation files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
GENE_NAME	GENE_SYMBOL		FMNL2
	COSMIC_GENE_ID	COSG+id_gene	COSG36014
ACCESSION_NUMBER	TRANSCRIPT_ACCESSION	Added version	ENST00000288670.9
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV56497166
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSN9069337/COSM
MUTATION_ID	MUTATION_ID		22807785
	MUTATION_NC_ID		62482415
GRCH		Remove GRCh since it's in the file name	
MUTATION_TYPE	MUTATION_TYPE	extend to non-coding	coding
IS_CANONICAL	IS_CANONICAL	**y/n/NULL**	y

README file

------------------------

COSMIC Mutation Tracking

A tab separated table listing the mapping of all COSMIC's legacy mutations (COSMs or COSNs) to the new genomic identifiers (COSVs). This file also helps to identify the transcripts and the accession numbers on which the current mutation is annotated on, along with the mutation type. [Cosmic_MutationTracking_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.

[2:B] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.

[3:C] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release.

[4:D] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.

[5:E] MUTATION_ID An internal mutation identifier to uniquely represent each mutation on a specific transcript on a given assembly build.

[6:F] MUTATION_NC_ID An internal mutation identifier to uniquely represent each non-coding mutation on a specific transcript on a given assembly build.

[7:G] MUTATION_TYPE Type of mutation (coding or non-coding)

[8:H] IS_CANONICAL The Ensembl Canonical transcript is a single, representative transcript identified at every locus. For details see: https://www.ensembl.org/info/genome/genebuild/canonical.html

16) Cosmic_NonCodingVariants_v99_GRCh37.tsv

File package

Cosmic_NonCodingVariants_Tsv_v99_GRCh37.tar contains:

- Cosmic_NonCodingVariants_v99_GRCh37.tsv.gz
- README_Cosmic_NonCodingVariants_v99_GRCh37.txt

Main changes

- Higher number of rows because multiple NCV_IDs can have the same sample, genomic_mutation_id and legacy_mutation_id
- Renamed COSMICNCV.tsv.gz to COSMICNonCodingVariants.tsv.gz for consistency
- Added new identifier ids to connect to sample, study, mutation tracking and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
	MUTATION_NC_ID		51864116
SAMPLE_NAME	SAMPLE_NAME		1192-01-02TD

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
ID_SAMPLE	COSMIC_SAMPLE_ID	COSS + id_sample	COSS2456388
ID_TUMOUR			
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO27985045
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV63265199
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSN19086933
ZYGOSITY	ZYGOSITY		Unknown
GRCH			
GENOME_POSITION			
	CHROMOSOME		6
	GENOME_START		77520142
	GENOME_END		77520142
MUTATION_SOMATIC_STATUS			
WT_SEQ		replace with genomic	
MUT_SEQ		replace with genomic	
	GENOMIC_WT_SEQ		C
	GENOMIC_MUT_SEQ		T
SNP		Always "y'	
FATHMM_MKL_NON_CODING_SCORE			
FATHMM_MKL_NON_CODING_GROUPS		always null	
FATHMM_MKL_CODING_SCORE			
FATHMM_MKL_CODING_GROUPS		always null	
WHOLE_GENOME_RESEQ			
WHOLE_EXOME			
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	COSU340
PUBMED_PMID	PUBMED_PMID		21642962
HGVSG	HGVSG		6:g.77520142C>T

README file

COSMIC Non coding variants

A tab separated table of all non-coding mutations from the current release. [Cosmic_NonCodingVariants_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] MUTATION_NC_ID An internal mutation identifier to uniquely represent each non-coding mutation on a specific transcript on a given assembly build.
[2:B] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[3:C] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[4:D] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[5:E] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[6:F] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.
[7:G] ZYGOSITY Information on whether the mutation was reported to be homozygous, heterozygous or unknown within the sample.
[8:H] CHROMOSOME The chromosome location of a given non coding variant (1-22, X, Y or MT).
[9:I] GENOME_START The start coordinate of a given non coding variant.
[10:J] GENOME_STOP The end coordinate of a given non coding variant.
[11:K] GENOMIC_WT_ALLELE Genomic Wild type allele sequence.
[12:L] GENOMIC_MUT_ALLELE Genomic mutation allele sequence.
[13:M] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this sample.
[14:N] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication.
[15:O] HGVSG Human Genome Variation Society genomic syntax (3' shifted).
[16:P] MUTATION_SOMATIC_STATUS Information on whether the sample was reported to be Confirmed somatic variant, Reported in another cancer sample as somatic or Variant of unknown origin:

* Reported in another cancer sample as somatic = when the mutation has been reported as somatic previously but not in current paper
* Confirmed somatic variant = if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient
* Variant of unknown origin = When the tumour has been sequenced without a matched normal tissue from the same individual, the somatic status of the variant cannot be assessed

17) Cosmic_ResistanceMutations_v99_GRCh37.tsv

File package

Cosmic_ResistanceMutations_Tsv_v99_GRCh37.tar contains:

- Cosmic_ResistanceMutations_v99_GRCh37.tsv.gz
- README_Cosmic_ResistanceMutations_v99_GRCh37.txt

Main changes

- Same size as current file
- Added new identifier ids to connect to sample, Gene, study, mutation tracking and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_NAME	SAMPLE_NAME		1000815
SAMPLE_ID	COSMIC_SAMPLE_ID	COSS + id_sample	COSS1000815
GENE_NAME	GENE_SYMBOL		EGFR
	COSMIC_GENE_ID	COSG + gene_id	COSG35617
TRANSCRIPT	TRANSCRIPT_ACCESSION	Added version	ENST00000275493.2
TIER		1 or null	
CENSUS_GENE	CENSUS_GENE		Yes
DRUG_NAME	DRUG_NAME		Gefitinib
GENOMIC_MUTATION_ID	GENOMIC_MUTATION_ID		COSV51765492
LEGACY_MUTATION_ID	LEGACY_MUTATION_ID		COSM6240
MUTATION_ID	MUTATION_ID		22182846
AA_MUTATION	AA_MUTATION		p.T790M
CDS_MUTATION	CDS_MUTATION		c.2369C>T
	GENOMIC_WT_SEQ		C
	GENOMIC_MUT_SEQ		T
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO29974826
PRIMARY_TISSUE			
TISSUE_SUBTYPE_1			
TISSUE_SUBTYPE_2			
HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
PUBMED_ID	PUBMED_ID		16983123
CGP_STUDY		no data	
	COSMIC_STUDY_ID	COSU + study_id	
SOMATIC_STATUS		Not currently available in curation	
SAMPLE_TYPE		Can be fetched from sample	
MUTATION_ZYGOSITY	MUTATION_ZYGOSITY		
Genome Coordinates (GRCh38)			
	CHROMOSOME		7
	GENOME_START		55249071
	GENOME_END		55249071
	STRAND		+
HGVSP	HGVSP		ENSP00000275493.2:p.Thr790Met

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
HGVSC	HGVSC		ENST00000275493.2:c.2369C>T
HGVSG	HGVSG		7:g.55249071C>T

README file

COSMIC Resistance Mutations

A tab separated table listing the details of all mutations in COSMIC which are known to confer drug resistance. [Cosmic_ResistanceMutations_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.
[2:B] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[3:C] GENE_SYMBOL The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier.
[4:D] COSMIC_GENE_ID A unique COSMIC gene identifier (COSG) is used to identify a gene within the file. This identifier can be used to retrieve additional Gene information from the Cosmic_Genes file.
[5:E] TRANSCRIPT_ACCESSION Unique Ensembl Transcript identifier (ENST). For details see: https://www.ensembl.org/info/genome/stable_ids/index.html. This identifier can be used to retrieve additional Transcript information from the Cosmic_Transcripts file.
[6:F] CENSUS_GENE Is the gene in the Cancer Gene Census (Yes/No).
[7:G] DRUG_NAME The name of the drug which the mutation confers resistance to.
[8:H] GENOMIC_MUTATION_ID Genomic mutation identifier (COSV) to indicate the definitive position of the variant on the genome. This identifier is trackable and stable between different versions of the release. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[9:I] LEGACY_MUTATION_ID Legacy mutation identifier (COSM) or (COSN) that will represent existing COSM or COSN mutation identifiers.
[10:J] MUTATION_ID An internal mutation identifier to uniquely represent each mutation on a specific transcript on a given assembly build. This identifier can be used to retrieve additional legacy mutation ids from the Cosmic_MutationTracking file.
[11:K] MUTATION_CDS The change that has occurred in the nucleotide sequence. Formatting is identical to the method used for the peptide sequence.
[12:L] Mutation_AA The change that has occurred in the peptide sequence. Formatting is based on the recommendations made by the Human Genome Variation Society. The description of each type can be found by following the link to the Mutation Overview page.
[13:M] GENOMIC_WT_ALLELE Genomic Wild type allele sequence.
[14:N] GENOMIC_MUT_ALLELE Genomic mutation allele sequence.
[15:O] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[16:P] PUBMED_PMID The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication.
[17:Q] COSMIC_STUDY_ID A unique COSMIC study identifier (COSU) is used to identify a study that have involved this sample.
[18:R] MUTATION_ZYGOSITY Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample.

[19:S] CHROMOSOME The chromosome location of a given resistance mutation (1-22, X, Y or MT).
[20:T] GENOME_START The start coordinate of a given resistance mutation.
[21:U] GENOME_STOP The end coordinate of a given resistance mutation.
[22:V] STRAND Positive or negative (+/-).
[23:W] HGVSP Human Genome Variation Society peptide syntax.
[24:X] HGVSC Human Genome Variation Society coding dna sequence syntax (CDS).
[25:Y] HGVSG Human Genome Variation Society genomic syntax (3' shifted).
[26:Z] MUTATION_SOMATIC_STATUS Information on whether the sample was reported to be Confirmed somatic variant, Reported in another cancer sample as somatic or Variant of unknown origin:
* Reported in another cancer sample as somatic = when the mutation has been reported as somatic previously but not in current paper
* Confirmed somatic variant = if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient
* Variant of unknown origin = When the tumour has been sequenced without a matched normal tissue from the same individual, the somatic status of the variant cannot be assessed

18) Cosmic_Sample_v99_GRCh37.tsv

File package

Cosmic_Sample_Tsv_v99_GRCh37.tar contains:

- Cosmic_Sample_v99_GRCh37.tsv.gz
- README_Cosmic_Sample_v99_GRCh37.txt

Main changes

- Same size as current file
- Added new identifier ids to connect to sample and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_ID	COSMIC_SAMPLE_ID	include the prefix COSS + id_sample	COSS2367783
SAMPLE_NAME	SAMPLE_NAME		2367783
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO37914801
ID_TUMOUR	ID_TUMOUR		2230621
SAMPLE_TYPE	SAMPLE_TYPE		surgery - NOS
ID_INDIVIDUAL	ID_INDIVIDUAL		2081267
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
THERAPY_RELATIONSHIP			
SAMPLE_DIFFERENTIATOR			
MUTATION_ALLELE_SPECIFICATION			
MSI			
AVERAGE_PLOIDY			
WHOLE_GENOME_SCREEN	WHOLE_GENOME_SCREEN		n
WHOLE_EXOME_SCREEN	WHOLE_EXOME_SCREEN		n
	TARGETED_SCREEN		y
	RNASEQ_SCREEN		n
	REARRANGEMENT_SCREEN		n
SAMPLE_REMARK			
DRUG_RESPONSE			
GRADE			
AGE_AT_TUMOUR_RECURRENCE			
STAGE			
CYTOGENETICS			
METASTATIC_SITE			
TUMOUR_SOURCE			NS
TUMOUR_REMARK			
AGE			
ETHNICITY			
ENVIRONMENTAL_VARIABLES			
GERMLINE_MUTATION			
THERAPY			
FAMILY			
NORMAL_TISSUE_TESTED	NORMAL_TISSUE_TESTED		y
GENDER	GENDER		u
INDIVIDUAL_REMARK			
NCI_CODE		Already in the classification file	

README file

COSMIC Sample

All the COSMIC sample data without the features from the current release in a tab separated file. [Cosmic_Sample_v99_GRCh37.tsv.gz]

File Description

[column number:label] Heading Description

[1:A] COSMIC_SAMPLE_ID A unique COSMIC sample identifier (COSS) is used to identify a sample. Other download files can be linked to this file using this identifier.

[2:B] SAMPLE_NAME The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process.

[3:C] COSMIC_PHENOTYPE_ID A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.

[4:D] TUMOUR_ID A number of samples can be taken from a single tumour and a number of tumours can be obtained from one individual.

[5:E] SAMPLE_TYPE Describes where the sample originated from.

[6:F] INDIVIDUAL_ID A unique id to identify an individual

[7:G] WHOLE_GENOME_SCREEN Was the sample whole genome screened (y/n).

[8:H] WHOLE_EXOME_SCREEN Was the sample whole exome sequenced (y/n).

[9:I] TARGETED_SCREEN Was the sample targeted screened (y/n).

[10:J] RNASEQ_SCREEN Was the sample RNASeq screened (y/n).

[11:K] REARRANGEMENT_SCREEN Was the sample rearrangement screened (y/n)

[12:L] TUMOUR_SOURCE Source of tumour tissue sample e.g. primary, metastasis.

[13:M] NORMAL_TISSUE_TESTED If normal tissue from the same individual has been screened for mutations.

[14:N] GENDER Sex of individual.

[15:O] AGE Age (in years) of individual at diagnosis or at the earliest tumour presentation.

[16:P] THERAPY_RELATIONSHIP Relates the time-point of tissue sampling to the drug therapy used to treat the tumour.

[17:Q] SAMPLE_DIFFERENTIATOR Gives additional information if more than one sample (e.g. carcinomatous and sarcomatous components) from a tumour has been screened for mutations or if samples from a tumour were taken at different time points.

[18:R] MUTATION_ALLELE_SPECIFICATION Where a publication has information on more than one mutation for one gene in a sample and reports whether or not the mutations occurred on the same or different chromosomes.

[19:S] MSI If microsatellite instability data is given in the publication per sample then High, Low, Stable/Low, MSI or Stable is reported in COSMIC. Unknown is the default.

[20:T] AVERAGE_PLOIDY The average ploidy of the sample, calculated from copy number data (where available).

[21:U] SAMPLE_REMARK Any additional sample information e.g. % mutant allele burden.

[22:V] DRUG_RESPONSE Clinical and in vitro responses to drugs (particularly targeted drugs). Phrasing based on RECIST guidelines. Note that in COSMIC, SD (stable disease) and PD (progressive disease) = clinical primary non response.

[23:W] GRADE Grade of tumour. The phrase 'Some Grade data are given in publication' is used when publication reports grade data or when data hasn't been given per sample. More detailed data follow commonly used grading systems in tumours.

[24:X] AGE_AT_TUMOUR_RECURRENCE Where both primary and recurrent tumour samples from an individual have been screened for mutations and the age (in years or months) of the patient at the time of the recurrence is different to that at diagnosis.

[25:Y] STAGE Stage of tumour. The phrase 'Some Stage data are given in publication' is used when publication reports stage data or when data hasn't been given per sample. More detailed data follow commonly used staging systems in tumours.

[26:Z] CYTOGENETICS Karyotype of the tumour.

[27:AA] METASTATIC_SITE Tissue site of any metastases identified in an individual.

[28:AB] TUMOUR_REMARK Any additional tumour information e.g. metachronous tumour.

[29:AC] ETHNICITY Ethnicity (e.g. Caucasian) of individual.

[30:AD] ENVIRONMENTAL_VARIABLES Environmental variables to which an individual has been exposed (e.g. viral exposure, smoking status).
[31:AE] GERMLINE_MUTATION Gene name/mutation if a germline mutation as well as a somatic mutation has been detected in the same gene in the same tumour sample.
[32:AF] THERAPY Any significant treatment an individual has received prior to mutation screening.
[33:AG] FAMILY Any familial cancer history for an individual or familial relationships of individuals screened for mutations in the same publication.
[34:AH] INDIVIDUAL_REMARK Any additional individual information (e.g. age group, hereditary syndromes).

19) Cosmic_StructuralVariants_v99_GRCh37.tsv

File package

Cosmic_StructuralVariants_Tsv_v99_GRCh37.tar contains:

- Cosmic_StructuralVariants_v99_GRCh37.tsv.gz
- README_Cosmic_Struct_v99_GRCh37.txt

Main changes

- Same size as current file
- Added new identifier ids to connect to sample, study and classification files

List of column changes

Blue: new column

Black: no change

Green: column renamed

Orange: column changed

Red: column removed

PREVIOUS CONTENT	CURRENT CONTENT	CHANGES	EXAMPLE
SAMPLE_NAME	SAMPLE_NAME		A21A-0096_CRUK_PC_0096_M1_DNA
ID_SAMPLE	COSMIC_SAMPLE_ID	include the prefix COSS	COSS2340984
	COSMIC_PHENOTYPE_ID	COSO + tum_class_link.id_site_class + tum_class_link.id_hist_class	COSO32054826
ID_TUMOUR			
PRIMARY_SITE			
SITE_SUBTYPE_1			
SITE_SUBTYPE_2			
SITE_SUBTYPE_3			
PRIMARY_HISTOLOGY			
HISTOLOGY_SUBTYPE_1			
HISTOLOGY_SUBTYPE_2			
HISTOLOGY_SUBTYPE_3			
MUTATION_ID	COSMIC_STRUCTURAL_ID	COST[ID_STRUCT_MUT]	COST188305
MUTATION_TYPE	MUTATION_TYPE		intrachromosomal inversion

GRCH			
DESCRIPTION	DESCRIPTION		chr8:g.51293657_52888676inv
PUBMED_PMID	PUBMED_PMID		
ID_STUDY	COSMIC_STUDY_ID	COSU + study_id	COSU538
	ID_STRUC_GEN		86403
	CHROMOSOME_FROM		8
	CHROMOSOME_TO		8
	LOCATION_FROM_MIN		51293657
	LOCATION_FROM_MAX		51293657
	LOCATION_TO_MIN		52888676
	LOCATION_TO_MAX		52888676
	STRAND_FROM		-
	STRAND_TO		+

README file

```
--------------------
COSMIC Structural Variants
--------------------
 All structural variants from the current release in a tab separated table. [ Cosmic_StructuralVariants_v99_GRCh37.tsv.gz ]

  File Description

[column number:label] Heading                    Description
----------------------------------------------------------------------------------------------------
[1:A]           SAMPLE_NAME                    The sample name can be derived from a number of sources. In many cases it originates from the cell line name. Other sources include names assigned by the annotators, or an incremented number assigned during an anonymization process..
[2:B]           COSMIC_SAMPLE_ID                A unique COSMIC sample identifier (COSS) is used to identify a sample. This identifier can be used to retrieve additional Sample information from the Cosmic_Sample file.
[3:C]           COSMIC_PHENOTYPE_ID             A unique COSMIC identifier (COSO) for the classification. This identifier can be used to retrieve tissue and histology information from the classification file.
[4:D]           COSMIC_STRUCTURAL_ID            A COSMIC structural identifier (COST). This identifier can be used to retrieve structural variants from the Cosmic_StructuralVariants file
[5:E]           MUTATION_TYPE                  Type of mutation : Intra/Inter (chromosomal), tandem duplication, deletion, inversion, complex substitutions, complex amplicons.
[6:F]           DESCRIPTION                  A syntax which describes the structural variant, based on HGVS recommendations.
[7:G]           PUBMED_PMID                    The PUBMED ID for the paper that the sample was noted in.
[8:H]           COSMIC_STUDY_ID                  A unique COSMIC study identifier (COSU) is used to identify a study that have involved this structural variant.
[9:I]          ID_STRUC_GEN                 A id representing structural genomic.
[10:J]          CHROMOSOME_FROM                 The chromosome where the first structural variant occurs.
[11:K]           CHROMOSOME_TO                The chromosome where the last structural variant occurs.
[12:L]          LOCATION_FROM_MIN               The first position in structural variant range.
[13:M]           LOCATION_FROM_MAX               The last position in structural variant range.
[14:N]          LOCATION_TO_MIN               The first position in structural variant range.
[15:O]          LOCATION_TO_MAX               The last position in structural variant range.
[16:P]          STRAND_FROM                  Positive or negative (+1/-1) where the first structural variant occurs.
[17:Q]           STRAND_TO                 Positive or negative (+1/-1) where the last structural variant occurs.
```

20) Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.vcf

File packages

Cosmic_CompleteTargetedScreensMutant_Vcf_v99_GRCh37.tar contains:

- Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.vcf.gz
- README_Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.txt

Cosmic_CompleteTargetedScreensMutant_VcfNormal_v99_GRCh37.tar contains:

- Cosmic_CompleteTargetedScreensMutant_Normal_v99_GRCh37.vcf.gz
- README_Cosmic_CompleteTargetedScreensMutant_Normal_v99_GRCh37.txt

Main changes

- CosmicCodingMuts.vcf splitted into Targeted and Genome to match TSV files

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       869556  COSV59704645    A       G       .       .
GENE=SAMD11;TRANSCRIPT=ENST00000342066.3;STRAND=+;LEGACY_ID=COSN15657006;CDS=c
.306-1596A>G;AA=p.?;HGVSC=ENST00000342066.3:c.306-1596A>G;HGVSG=1:g.869556A>G;SAMP
LE_COUNT=1;IS_CANONICAL=y;SO_TERM=SNV;
1
```

README file

```
-------------------
COSMIC Coding Mutations (Targeted Screens) VCF
-------------------
 VCF file of the complete curated COSMIC dataset (targeted screens) from the current release. [
Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.vcf.gz ]

  File Description

##fileformat=VCFv4.1
##source=COSMICv99
##reference=GRCh37
##fileDate=20210917
##comment="Missing nucleotide details indicate ambiguity during curation process"
##comment="URL stub for ID field (use the whole COSV
identifier)='https://cancer.sanger.ac.uk/cosmic/search?genome=37&q='"
##comment="REF and ALT sequences are both forward strand
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene name">
##INFO=<ID=TRANSCRIPT,Number=1,Type=String,Description="Transcript accession">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
##INFO=<ID=LEGACY_ID,Number=1,Type=String,Description="Legacy Mutation ID">
##INFO=<ID=CDS,Number=1,Type=String,Description="CDS annotation">
##INFO=<ID=AA,Number=1,Type=String,Description="Peptide annotation">
##INFO=<ID=HGVSC,Number=1,Type=String,Description="HGVS cds syntax">
##INFO=<ID=HGVSP,Number=1,Type=String,Description="HGVS peptide syntax">
##INFO=<ID=HGVSG,Number=1,Type=String,Description="HGVS genomic syntax">
##INFO=<ID=SAMPLE_COUNT,Number=1,Type=Integer,Description="How many samples have this mutation">
##INFO=<ID=IS_CANONICAL,Number=1,Type=String,Description="The Ensembl Canonical transcript is a single,
representative transcript identified at every locus. For details see:
https://www.ensembl.org/info/genome/genebuild/canonical.html">
##INFO=<ID=TIER,Number=1,Type=String,Description="Indicates to which tier of the Cancer Gene Census the gene
belongs (1/2)">
```

```
##INFO=<ID=SO_TERM,Number=1,Type=String,Description="SO term for this mutation">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
```

A tab separated table of the complete curated COSMIC dataset (targeted screens) from the current release. It includes all coding point mutations, and the negative data set. [Cosmic_CompleteTargetedScreensMutant_v99_GRCh37.tsv.gz]
The CosmicMutantExport file can be re-created by linking the Cosmic_GenomeScreensMutant with the positive data from this file Cosmic_CompleteTargetedScreensMutant

21) Cosmic_GenomeScreensMutant_v99_GRCh37.vcf

File packages

Cosmic_GenomeScreensMutant_Vcf_v99_GRCh37.tar contains:

- Cosmic_GenomeScreensMutant_v99_GRCh37.vcf.gz
- README_Cosmic_GenomeScreensMutant_v99_GRCh37.txt

Cosmic_GenomeScreensMutant_VcfNormal_v99_GRCh37.tar contains:

- Cosmic_GenomeScreensMutant_Normal_v99_GRCh37.vcf.gz
- README_Cosmic_GenomeScreensMutant_Normal_v99_GRCh37.txt

Main changes

- CosmicCodingMuts.vcf splitted into Targeted and Genome to match TSV files

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       69224   COSV58737130    A       C       .       .
GENE=OR4F5;TRANSCRIPT=ENST00000335137.3;STRAND=+;LEGACY_ID=COSM3677745;CDS=c.13
4A>C;AA=p.D45A;HGVSC=ENST00000335137.3:c.134A>C;HGVSP=ENSP00000334393.3:p.Asp45Ala
;HGVSG=1:g.69224A>C;SAMPLE_COUNT=1;IS_CANONICAL=y;SO_TERM=SNV;
```

README file

COSMIC Coding Mutations (Genome Screens) VCF

VCF file of coding point mutations from genome wide screens (including whole exome sequencing) from the current release. [Cosmic_GenomeScreensMutant_v99_GRCh37.vcf.gz]

File Description

```
##fileformat=VCFv4.1
##source=COSMICv99
##reference=GRCh37
##fileDate=20210917
##comment="Missing nucleotide details indicate ambiguity during curation process"
##comment="URL stub for ID field (use the whole COSV
identifier)='https://cancer.sanger.ac.uk/cosmic/search?genome=37&q='"
##comment="REF and ALT sequences are both forward strand
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene name">
##INFO=<ID=TRANSCRIPT,Number=1,Type=String,Description="Transcript accession">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
##INFO=<ID=LEGACY_ID,Number=1,Type=String,Description="Legacy Mutation ID">
##INFO=<ID=CDS,Number=1,Type=String,Description="CDS annotation">
##INFO=<ID=AA,Number=1,Type=String,Description="Peptide annotation">
##INFO=<ID=HGVSC,Number=1,Type=String,Description="HGVS cds syntax">
```

```
##INFO=<ID=HGVSP,Number=1,Type=String,Description="HGVS peptide syntax">
##INFO=<ID=HGVSG,Number=1,Type=String,Description="HGVS genomic syntax">
##INFO=<ID=SAMPLE_COUNT,Number=1,Type=Integer,Description="How many samples have this mutation">
##INFO=<ID=IS_CANONICAL,Number=1,Type=String,Description="The Ensembl Canonical transcript is a single,
representative transcript identified at every locus. For details see:
https://www.ensembl.org/info/genome/genebuild/canonical.html">
##INFO=<ID=TIER,Number=1,Type=String,Description="Indicates to which tier of the Cancer Gene Census the gene
belongs (1/2)">
##INFO=<ID=SO_TERM,Number=1,Type=String,Description="SO term for this mutation">
#CHROM  POS    ID    REF    ALT    QUAL    FILTER  INFO
```

22) Cosmic_NonCodingVariants_v99_GRCh37.vcf

File packages

Cosmic_NonCodingVariants_Vcf_v99_GRCh37.tar contains:

- Cosmic_NonCodingVariants_v99_GRCh37.vcf.gz
- README_Cosmic_NonCodingVariants_v99_GRCh37.txt

Cosmic_NonCodingVariants_VcfNormal_v99_GRCh37.tar contains:

- Cosmic_NonCodingVariants_Normal_v99_GRCh37.vcf.gz
- README_Cosmic_NonCodingVariants_Normal_v99_GRCh37.txt

Main changes

- Including Complex - compound substitution (id_mut_type=29)
- File name changed for consistency
- 270 rows with '.' (id_mut_type=13), These are now defined the same way as deletion (e.g: GATATG G instead of GATATG .)
- Header is now specific to each VCFs to avoid having CDS and AA information in non-coding header.

```
#CHROM  POS    ID    REF    ALT    QUAL    FILTER  INFO
1      10108   COSV70831266    C      T      .      .
GENE=WASH7P;TRANSCRIPT=ENST00000538476.1;STRAND=-;LEGACY_ID=COSN28762392;HGVSG
=1:g.10108C>T;SAMPLE_COUNT=1;IS_CANONICAL=n;SO_TERM=SNV;
```

README file

```
-------------------
COSMIC Non Coding Variants VCF
-------------------
 VCF file of all non coding variants in the current release. [ Cosmic_NonCodingVariants_v99_GRCh37.vcf.gz ]

  File Description

##fileformat=VCFv4.1
##source=COSMICv99
##reference=GRCh37
##fileDate=20210917
##comment="Missing nucleotide details indicate ambiguity during curation process"
##comment="URL stub for ID field (use the whole COSV
identifier)='https://cancer.sanger.ac.uk/cosmic/search?genome=37&q='"
```

```
##comment="REF and ALT sequences are both forward strand
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene name">
##INFO=<ID=TRANSCRIPT,Number=1,Type=String,Description="Transcript accession">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
##INFO=<ID=LEGACY_ID,Number=1,Type=String,Description="Legacy Mutation ID">
##INFO=<ID=HGVSG,Number=1,Type=String,Description="HGVS genomic syntax">
##INFO=<ID=SAMPLE_COUNT,Number=1,Type=Integer,Description="How many samples have this mutation">
##INFO=<ID=IS_CANONICAL,Number=1,Type=String,Description="The Ensembl Canonical transcript is a single,
representative transcript identified at every locus. For details see:
https://www.ensembl.org/info/genome/genebuild/canonical.html">
##INFO=<ID=TIER,Number=1,Type=String,Description="Indicates to which tier of the Cancer Gene Census the gene
belongs (1/2)">
##INFO=<ID=SO_TERM,Number=1,Type=String,Description="SO term for this mutation">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
```

CELL LINES PROJECT DOWNLOAD FILES

No changes have been made to the file contents or structure but they have been renamed to follow the new convention and are also available as a tar file that contains a descriptive README.txt file

23) CellLinesProject_CompleteCNA_v99_GRCh37.tsv

File package

CellLinesProject_CompleteCNA_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_CompleteCNA_v99_GRCh37.tsv.gz
- README_CellLinesProject_CompleteCNA_v99_GRCh37.txt

24) CellLinesProject_CompleteGeneExpression_v99_GRCh37.tsv

File package

CellLinesProject_CompleteCNA_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_CompleteGeneExpression_v99_GRCh37.tsv.gz
- README_CellLinesProject_CompleteGeneExpression_v99_GRCh37.txt

25) CellLinesProject_GenomeScreensMutant_v99_GRCh37.tsv

File package

CellLinesProject_GenomeScreensMutant_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_GenomeScreensMutant_v99_GRCh37.tsv.gz
- README_CellLinesProject_GenomeScreensMutant_v99_GRCh37.txt

26) CellLinesProject_MutationTracking_v99_GRCh37.tsv

File package

CellLinesProject_MutationTracking_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_MutationTracking_v99_GRCh37.tsv.gz
- README_CellLinesProject_MutationTracking_v99_GRCh37.txt

27) CellLinesProject_NonCodingVariants_v99_GRCh37.tsv

File package

CellLinesProject_NonCodingVariants_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_NonCodingVariants_v99_GRCh37.tsv.gz
- README_CellLinesProject_NonCodingVariants_v99_GRCh37.txt

28) CellLinesProject_RawGeneExpression_v99_GRCh37.tsv

File package

CellLinesProject_RawGeneExpression_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_RawGeneExpression_v99_GRCh37.tsv.gz
- README_CellLinesProject_RawGeneExpression_v99_GRCh37.txt

29) CellLinesProject_Sample_v99_GRCh37.tsv

File package

CellLinesProject_Sample_Tsv_v99_GRCh37.tar contains:

- CellLinesProject_Sample_v99_GRCh37.tsv.gz
- README_CellLinesProject_Sample_v99_GRCh37.txt

30) CellLinesProject_GenomeScreensMutant_v99_GRCh37.vcf

File package

CellLinesProject_GenomeScreensMutant_Vcf_v99_GRCh37.tar contains:

- CellLinesProject_GenomeScreensMutant_v99_GRCh37.vcf.gz
- README_CellLinesProject_GenomeScreensMutant_v99_GRCh37.txt

CellLinesProject_GenomeScreensMutant_VcfNormal_v99_GRCh37.tar contains:

- CellLinesProject_GenomeScreensMutant_Normal_v99_GRCh37.vcf.gz
- README_CellLinesProject_GenomeScreensMutant_Normal_v99_GRCh37.txt

31) CellLinesProject_NonCodingVariants_v99_GRCh37.vcf

File package

CellLinesProject_NonCodingVariants_Vcf_v99_GRCh37.tar contains:

- CellLinesProject_NonCodingVariants_v99_GRCh37.vcf.gz
- README_CellLinesProject_NonCodingVariants_v99_GRCh37.txt

CellLinesProject_NonCodingVariants_VcfNormal_v99_GRCh37.tar contains:

- CellLinesProject_NonCodingVariants_Normal_v99_GRCh37.vcf.gz
- README_CellLinesProject_NonCodingVariants_Normal_v99_GRCh37.txt

ACTIONABILITY AND CANCER MUTATION CENSUS (CMC) DOWNLOAD FILE CHANGES

No changes have been made to the file contents or structure but they have been renamed to follow the new convention and are also available as a tar file that contains a descriptive README file

32) Actionability_AllData_v10_GRCh38.tsv

File package
Actionability_AllData_Tsv_v10_GRCh37.tar contains:

- Actionability_AllData_v10_GRCh37.tsv
- README_Actionability_AllData_v10_GRCh37.pdf

33) CancerMutationCensus_AllData_v99_GRCh38.tsv

File package
CancerMutationCensus_AllData_Tsv_v99_GRCh37.tar contains:

- CancerMutationCensus_AllData_v99_GRCh37.tsv.gz
- README_CancerMutationCensus_AllData_v99_GRCh37.txt

End.