Tool: OPENGDC			
Web-page: http://bioinformatics.iasi.cnr.it/opengdc/			
Subject: OPENGDC file format definition			
Document class: Final			
Release: 1.0	*Date:* 19/03/2018	Authors: Eleonora Cappelli; Fabio Cumbo, Marco Masseroli, Emanuel Weitschek	OPENGDC

OPENGDC file format definition

Contents

Tool: OPENGDC			
Web-page: http://bioinformatics.iasi.cnr.it/opengdc/			
Subject: OPENGDC file format definition			
Document class: Final			
Release: 1.0	*Date:* 19/03/2018	Authors: Eleonora Cappelli; Fabio Cumbo, Marco Masseroli, Emanuel Weitschek	**OPENGDC**

Introduction

TCGA2BED

TCGA2BED is a Java software application for the automatic extraction, extension and conversion of genomic and clinical data of cancer retrieved from The Cancer Genome Atlas (TCGA), one of the most relevant repositories containing data of about 33 different tumor types and more than 1000 involved healthy and diseased patients. More precisely, TCGA offers data regarding different types of experiments including Copy Number Variations, DNA-Sequencing, DNA-Methylation, miRNA-Sequencing, and RNA-Sequencing. The main goals of TCGA2BED are (i) to automate the extraction of these data from the TCGA repository and the proprietary tab-delimited format in which TCGA provide them, (ii) extend them by integrating information retrieved from different public sources such as NCBI, HGNC, UCSC, and MIRBase, and (iii) convert them into the BED format, which is more usable for biologists, bioinformaticians, and life scientists, and additionally it is fully supported by the GenoMetric Query Language (GMQL)[1]. GMQL is an innovative system, based on the Genomic Data Model (GDM)[2], able to process numerous and heterogeneous genomic data in the cloud in order to extract information about their metric (co)occurrence genome-wide (http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQL/).

TCGA2BED is also a FTP repository, available at ftp://bioinf.iasi.cnr.it/, containing the original public data from TCGA and the same data converted in BED format and extended with additional information, for a total of more than 650 GB. Additionally, the TCGA2BED software is accessible under GPL license and it is freely available from the project page at http://bioinf.iasi.cnr.it/tcga2bed/. This work has been described in the published article: *Cumbo F, Fiscon G, Ceri S, Masseroli M, Weitschek E.* ***TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas.*** *BMC Bioinformatics, 2017; 18(1), 6.*

Motivations and goal

In July 2016, TCGA closed its data portal, making its data unavailable. Recently the U.S. National Cancer Institute (NCI), the same authors of the TCGA project, have opened a new portal that currently hosts data from TCGA and from TARGET, a project to collect genomic experiments on children affected by different tumors. The new portal is called Genomic Data Commons (GDC), and aims to make available genomic data of cancer projects. TCGA data have been updated, both in the contents and in the structure, but TCGA2BED can no longer be updated, because of the changes in the TCGA data portal.

The goal of this work is the creation of a new software for the automatic extraction, extension, and conversion of the public experiments of the TCGA and TARGET projects available in GDC. The final formats of the conversion will be the BED, CSV, GTF, JSON, and XML to make data as much usable as possible for all domain experts. Additionally, a public FTP repository with original and

[1] *Masseroli M, Pinoli P, Venco F, Kaitoua A, Jalili V, Paluzzi F, Muller H, Ceri S*: **GenoMetric Query Language: A novel approach to large-scale genomic data management**. *Bioinformatics* 2015; 31(12):1881-1888.
[2] *Masseroli M, Kaitoua A, Pinoli P, Ceri S*. **Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying**. *Methods* 2016; 111: 3-11.

converted data sets will be created and will be accessible through the following address http://bioinformatics.iasi.cnr.it/opengdc/.

Input data sets

For the conversion of GDC TCGA and TARGET data files into the BED format, we actually take into account the following data sets, which include all the genomic data that the Genomic Data Commons (GDC) is currently providing publicly:

- Masked Somatic Mutation (msm)
- Gene Expression Quantification (geq)
- Methylation Beta Value (mbv)
- Copy Number Segment (cns)
- Masked Copy Number Segment (mcns)
- miRNA Expression Quantification (meq)
- Isoform Expression Quantification (ieq)
- Meta data: Biospecimen Supplement
- Meta data: Clinical Supplement

All data are retrieved from the "*GDC Application Programming Interface (API)*", available at https://gdc.cancer.gov/developers/gdc-application-programming-interface-api.

Following abbreviations are used for referring to the experimental data sets

Experiment	Abbreviation	Access
Copy Number Segment	cns	Open
Gene Expression Quantification	geq	Open
Isoform Expression Quantification	ieq	Open
Masked Copy Number Segment	mcns	Open
Masked Somatic Mutation	msm	Open
Methylation Beta Value	mbv	Open
miRNA Expression Quantification	meq	Open
Aligned Reads	ar	Controlled
Aggregated Somatic Mutation	agsm	Controlled
Annotated Somatic Mutation	ansm	Controlled
Raw Simple Somatic Mutation	rssm	Controlled

Data granularity

We consider the aliquot as the basic data granularity; it is the elementary unit of GDC (TARGET and TCGA), which identifies a single experiment on a tissue. The aliquot is the unit of analysis for GDC genomic data. Aliquots are the products shipped by the Biospecimen Core Resources to analysis centers. A Biospecimen Core Resource (BCR) is a TCGA or TARGET center where samples are carefully catalogued, processed, quality-checked and stored along with participant clinical information.
More details are available at https://docs.gdc.cancer.gov/Data/Data_Model/GDC_Data_Model/.
In GDC aliquots are encoded with the Universal Unique Identifier (uuid), a 128-bit number used to uniquely identify an object or entity in a system. More details about the uuid are available at https://docs.gdc.cancer.gov/Encyclopedia/pages/UUID/. Uuids are also used for identifying samples and patients in GDC. It is worth noting that the aliquot is encoded in the "biospecimen__bio__bcr_aliquot_uuid" meta data attribute. See meta data section for further details. For indexing our output data we use an internal ID called OpenGDC ID, which is composed of the "biospecimen__bio__bcr_aliquot_uuid" concatenated with the acronym of the considered experiment, i.e., biospecimen__bio__bcr_aliquot_uuid-experiment_acronym. See subsection "Input data sets" for the acronyms associated to the experiments.

Output data

We provide the user with all the data sets properly converted in BED format.
In particular, for each data set the data are provided as follows:

(i) a .bed file for each aliquot uuid, containing the experiment data converted in BED / CSV / GTF / JSON / XML formats;
(ii) a .meta file for each aliquot, with meta data including the patient clinical and biospecimen data;
(iii) a header.schema file in xml format that describes the structure of the BED files.

Several other files containing general and statistical information about the experiments and metadata are produced as output (e.g., MD5 checksum files, metadata dictionary file, experiment information files, experiments annotations). We point the reader to the section Additional output files of this document for further details.

We use the 1-based (1-start or base-counted or fully-closed) genomic coordinate representation, as adopted in the GDC data files.

Missing values in case present in original data, are homogeneously labeled in the output format with the string "null" for numerical attributes, or with an empty string "" for text attributes.

Reference assembly

The genomic coordinates in all GDC and converted data sets refer to the human reference assembly GRCh38[3]. In particular[4]:
Genome Reference Consortium Human Build 38

[3] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26
[4] https://gdc.cancer.gov/about-data/data-harmonization-and-generation/gdc-reference-files

Organism: Homo sapiens (human)
Submitter: Genome Reference Consortium
Date: 2013/12/17
Assembly type: haploid-with-alt-loci
Assembly level: Chromosome
Genome representation: full
Synonyms: hg38
GenBank assembly accession: GCA_000001405.15 (replaced)
RefSeq assembly accession: GCF_000001405.26 (replaced).

Tumor tags and tumor names

The following tumor tags of TCGA are available at GDC and correspond to the following tumor names:

TCGA-ACC	Adrenocortical carcinoma
TCGA-BLCA	Bladder Utothelial Carcinoma
TCGA-BRCA	Breast Invasive Carcinoma
TCGA-CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
TCGA-CHOL	Cholangiocarcinoma
TCGA-COAD	Colon adenocarcinoma
TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
TCGA-ESCA	Esophageal carcinoma
TCGA-GBM	Glioblastoma multiforme
TCGA-HNSC	Head and Neck squamous cell carcinoma
TCGA-KICH	Kidney Chromophobe
TCGA-KIRC	Kidney renal clear cell carcinoma
TCGA-KIRP	Kidney renal papillary cell carcinoma
TCGA-LAML	Acute Myeloid Leukemia
TCGA-LGG	Brain Lower Grade Glioma
TCGA-LIHC	Liver hepatocellular carcinoma
TCGA-LUAD	Lung adenocarcinoma
TCGA-LUSC	Lung squamous cell carcinoma
TCGA-MESO	Mesothelioma
TCGA-OV	Ovarian serous cystadenocarcinoma
TCGA-PAAD	Pancreatic adenocarcinoma
TCGA-PCPG	Pheochromocytoma and Paraganglioma
TCGA-PRAD	Prostate adenocarcinoma
TCGA-READ	Rectum adenocarcinoma
TCGA-SARC	Sarcoma
TCGA-SKCM	Skin Cutaneous Melanoma
TCGA-STAD	Stomach adenocarcinoma

TCGA-TGCT	Testicular Germ Cell Tumors
TCGA-THCA	Thyroid carcinoma
TCGA-THYM	Thymoma
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma
TCGA-UCS	Uterine Carcinosarcoma
TCGA-UVM	Uveal Melanoma

The following tumor tags of TARGET are available at GDC and correspond to the following tumor names:

TARGET-AML	Acute Myeloid Leukemia
TARGET-CCSK	Clear Cell Sarcoma of the Kidney
TARGET-NBL	Neuroblastoma
TARGET-OS	Osteosarcoma
TARGET-RT	Rhabdoid Tumor
TARGET-WT	High-Risk Wilms Tumor

Tool: OPENGDC			
Web-page: http://bioinformatics.iasi.cnr.it/opengdc/			
Subject: OPENGDC file format definition			
Document class: Final			
Release: 1.0	*Date:* 19/03/2018	Authors: Eleonora Cappelli; Fabio Cumbo, Marco Masseroli, Emanuel Weitschek	OPENGDC

Masked Somatic Mutation

This type of Next Generation Sequencing (NGS) experiment discovers mutations by aligning DNA sequences derived from tumor samples to sequences derived from normal samples and a reference sequence. A Mutation Annotation Format (MAF) file is used to specify, for each sample, the discovered putative or validated mutations and to categorize those mutations (SNP, deletion, or insertion) as somatic (originating in the tissue) or germline (originating from the germline), as well as to specify additional information for those mutations.
More details are available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization/high-level-data-generation/dna-seq-somatic-variation

Input: multiple MAF files for each tumor are provided by GDC, each with DNA-sequencing data; each of those files includes 125 attributes (columns), which are described at https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Example of the first 13 attributes (columns) of a GDC MAF file

#version 2.4

Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2
CTBS	1486	BCM	GRCh38	chr1	84570701	84570701	+	Missense_Mutation	SNP	C	C	T
ATF6	22926	BCM	GRCh38	chr1	161791444	161791444	+	Missense_Mutation	SNP	C	C	G
SLC35F3	148641	BCM	GRCh38	chr1	234309160	234309160	+	Missense_Mutation	SNP	C	C	A
TTN	7273	BCM	GRCh38	chr2	178704929	178704929	+	Missense_Mutation	SNP	T	T	A
SP140	11262	BCM	GRCh38	chr2	230238312	230238312	+	Missense_Mutation	SNP	G	G	A
ITPR1	3708	BCM	GRCh38	chr3	4673355	4673355	+	Missense_Mutation	SNP	G	G	C
BRPF1	7862	BCM	GRCh38	chr3	9739306	9739306	+	Missense_Mutation	SNP	G	G	C
BRPF1	7862	BCM	GRCh38	chr3	9745646	9745646	+	Missense_Mutation	SNP	G	G	C
OGG1	4968	BCM	GRCh38	chr3	9751153	9751153	+	Missense_Mutation	SNP	G	G	A
GOLGA4	2803	BCM	GRCh38	chr3	37327015	37327015	+	Missense_Mutation	SNP	G	G	T
XIRP1	165904	BCM	GRCh38	chr3	39186471	39186471	+	Missense_Mutation	SNP	G	G	A
HRG	3273	BCM	GRCh38	chr3	186669019	186669019	+	Missense_Mutation	SNP	G	G	C
PRMT9	90826	BCM	GRCh38	chr4	147683889	147683889	+	Silent	SNP	C	C	T
SLC6A19	340024	BCM	GRCh38	chr5	1213975	1213975	+	Missense_Mutation	SNP	A	A	G
DNAH5	1767	BCM	GRCh38	chr5	13753451	13753451	+	Missense_Mutation	SNP	C	C	A
HMMR	3161	BCM	GRCh38	chr5	163484133	163484133	+	Missense_Mutation	SNP	A	A	G
RP11-1277A3.2	0	BCM	GRCh38	chr5	177632498	177632498	+	RNA	SNP	G	G	A
RP3-420J14.1	0	BCM	GRCh38	chr6	11862180	11862180	+	RNA	SNP	C	C	A
ADGRB3	577	BCM	GRCh38	chr6	68956041	68956041	+	Missense_Mutation	SNP	G	G	C
AC013470.6	0	BCM	GRCh38	chr7	12471568	12471568	+	RNA	SNP	C	C	A
RP11-700P18.1	0	BCM	GRCh38	chr7	56291205	56291205	+	RNA	SNP	C	C	A
PKHD1L1	93035	BCM	GRCh38	chr8	109507790	109507790	+	Missense_Mutation	SNP	G	G	T
MURC	347273	BCM	GRCh38	chr9	100578481	100578481	+	Missense_Mutation	SNP	A	A	T
HNRNPF	3185	BCM	GRCh38	chr10	43387009	43387009	+	Missense_Mutation	SNP	G	G	C
CEP57L1P1	221017	BCM	GRCh38	chr10	70390293	70390293	+	RNA	SNP	A	A	C
SFXN4	119559	BCM	GRCh38	chr10	119164169	119164169	+	Missense_Mutation	SNP	T	T	A
PRDX3	10935	BCM	GRCh38	chr10	119172446	119172446	+	Missense_Mutation	SNP	C	C	G
CCDC73	493860	BCM	GRCh38	chr11	32614115	32614115	+	Missense_Mutation	SNP	T	T	G
NR1H3	10062	BCM	GRCh38	chr11	47261308	47261308	+	Silent	SNP	C	C	T
GAS2L3	283431	BCM	GRCh38	chr12	100623835	100623835	+	Missense_Mutation	SNP	G	G	C
WBP4	11193	BCM	GRCh38	chr13	41068702	41068702	+	Missense_Mutation	SNP	A	A	C

BED output format: a tab-separated BED file, in which each original DNA-seq .maf file is converted, with the following 18 fields, the main ones in the original MAF file:

1. **chrom** (i.e., the name of the chromosome, e.g., “chr3”, “chrY”, “chr2_random”, equal to the 5. field of the GDC MAF file)
2. **start** (i.e., the starting position of the feature in the chromosome or scaffold, e.g., 999, equal to the 6. field of the GDC MAF file)
3. **end** (i.e., the ending position of the feature in the chromosome or scaffold, e.g., 1000, equal to the 7. field of the GDC MAF file)
4. **strand** (i.e., the DNA strand where the feature is observed, either '+' or '-', equal to the 8. field of the GDC MAF file)

5. **gene_symbol** (i.e., the symbol of the gene related to the reported variant, if it exists, e.g., "HRG", equal to the 1. field of the GDC MAF file)
6. **entrez_gene_id** (i.e., the Entrez gene ID of the gene related to the reported variant, if it exists, e.g., "3273", equal to the 2. field of the GDC MAF file)
7. **variant_classification** (i.e., the classification of the reported variant, e.g., "Missense_Mutation", equal to the 9. field of the GDC MAF file)
8. **variant_type** (i.e., the type of mutation, e.g., "INS", equal to the 10. field of the GDC MAF file)
9. **reference_allele** (i.e., the plus strand reference allele at the variant position, e.g., "A", equal to the 11. field of the GDC MAF file)
10. **tumor_seq_allele1** (i.e., the tumor sequencing (discovery) allele 1, e.g., "C", equal to the 12. field of the GDC MAF file)
11. **tumor_seq_allele2** (i.e., the tumor sequencing (discovery) allele 2, e.g., "G", equal to the 13. field of the GDC MAF file)
12. **dbsnp_rs** (i.e., the latest dbSNP rs ID, e.g., "rs12345" or "novel" if not present in dbSNP, equal to the 14. field of the GDC MAF file)
13. **tumor_sample_barcode** (i.e., the BCR aliquot barcode for the tumor sample, e.g., "TCGA-02-0021-01A-01D-0002-04", equal to the 16. field of the GDC MAF file)
14. **matched_norm_sample_barcode** (i.e., the BCR aliquot barcode for the matched normal sample, e.g., "TCGA-02-0021-10A-01D-0002-04", equal to the 17. field of the GDC MAF file)
15. **match_norm_seq_allele1** (i.e., the matched normal sequencing allele 1, e.g., "T", equal to the 18. field of the GDC MAF file)
16. **match_norm_seq_allele2** (i.e., the matched normal sequencing allele 2, e.g., "ACGT", equal to the 19. field of the GDC MAF file)
17. **tumor_sample_uuid** (i.e., the BCR aliquot UUID for the tumor sample, e.g., "b2804bb2-70f4-471a-b6db-70c0ef457df3", equal to the 33. field of the GDC MAF file)
18. **matched_norm_sample_uuid** (i.e., the BCR aliquot UUID for the matched normal sample, e.g., "567e8487-e29b-32d4-a716-446655443246", equal to the 34. field of the GDC MAF file)

Notes about GDC MAF format

- This format is not to be confused with the UCSC Multiple Alignment Format
- The GDC MAF format regards a tab-delimited file containing only somatic mutations (open access portion of the GDC Data Portal for the TCGA and TARGET projects)
- Mutations are discovered by aligning DNA sequences derived from tumor samples to sequences derived from normal samples and a reference sequence. A MAF file specifies, for each sample, the discovered putative or validated mutations and categorizes those mutations (SNP, deletion, or insertion) as somatic (originating in the tissue), as well as specifies additional information for those mutations.
- Types of specified somatic mutations:
    - Missense and nonsense mutation
    - Splice site mutation, defined as SNP within 2 bp of the splice junction
    - Silent mutation
    - Indel mutation, that overlaps the coding region or splice site of a gene or the targeted region of a genetic element of interest
    - Frameshift mutation
    - Mutation in regulatory regions
- Included SNPs:
    - Any germline SNP with validation status "unknown" is included
    - SNPs already validated in dbSNP are not included, since they are unlikely to be

involved in cancer

- The 125 MAF format attributes (columns) are described at https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
- Column headers and values are case sensitive where specified
- Columns may allow null values (i.e., blank cells) and/or have enumerated values; when converted to BED format, null values for numeric columns (attributes) are marked with the "null" label, whereas those for not numeric (textual) columns (attributes) are left as blank cells

Gene Expression Quantification

GDC provides gene expression quantification data in three files for each aliquot:

- FPKM (i.e., Fragments Per Kilobase of transcript per Million mapped reads)
- FPKM-UQ (i.e., Upper Quartile normalized FPKM values)
- counts (i.e., raw mapping counts of reads mapped to each gene)

More details are described in the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization/high-level-data-generation/rna-seq-quantification.

Input: **FPKM file**

One tab delimited file is provided by GDC for each aliquot, with the following fields:

1. Gene_Ensembl (i.e., the Ensembl ID of the gene, including its version with "." notation);
2. FPKM (i.e., number of Fragments Per Kilobase of transcript per Million mapped reads).

FPKM file example

```
ENSG00000242268.2       0.0
ENSG00000270112.3       0.456673501724
ENSG00000167578.15      10.555943415
ENSG00000273842.1       0.0
ENSG00000078237.5       5.70425402923
ENSG00000146083.10      8.95127291553
ENSG00000225275.4       0.0
ENSG00000158486.12      0.0754083909194
ENSG00000198242.12      131.076819733
ENSG00000259883.1       0.0261281621307
ENSG00000231981.3       0.0
ENSG00000269475.2       0.0
ENSG00000201788.1       0.0
ENSG00000134108.11      33.7943884797
ENSG00000263089.1       0.00470563256313
ENSG00000172137.17      0.721569931396
ENSG00000167700.7       19.2386831804
ENSG00000234943.2       0.106869034497
ENSG00000240423.1       0.0478120561774
ENSG00000060642.9       2.96632669289
```

Input: **FPKM-UQ file**

Another tab-delimited file is provided by GDC for each aliquot, with the following fields:

1. Gene_Ensembl (i.e., the Ensembl ID of the gene, including its version with "." notation);
2. UQ-FPKM (i.e., Upper Quartile normalized FPKM value).

FPKM-UQ file example

```
ENSG00000242268.2       0.0
ENSG00000270112.3       7687.60487006
ENSG00000167578.15      631320.10322
ENSG00000273842.1       0.0
ENSG00000078237.5       294156.121221
ENSG00000146083.10      239960.786896
ENSG00000225275.4       3670.9102389
ENSG00000158486.12      207.59291859
ENSG00000198242.12      3081029.80076
ENSG00000259883.1       1342.6675582
ENSG00000231981.3       0.0
ENSG00000269475.2       0.0
ENSG00000201788.1       0.0
ENSG00000134108.11      723421.846124
ENSG00000263089.1       0.0
ENSG00000172137.17      909368.317267
ENSG00000167700.7       376486.782077
ENSG00000234943.2       0.0
ENSG00000240423.1       818.984721021
ENSG00000060642.9       142637.982352
```

Input: **Counts file**

Another tab-delimited file is provided by GDC for each aliquot, with the following fields:

1. Gene_Ensembl (i.e., the Ensembl ID of the gene, including its version with "." notation);
2. counts (i.e., the number of reads aligned to each gene, calculated by HT-Seq).

Counts file example

```
ENSG00000000003.13      3543
ENSG00000000005.5       1
ENSG00000000419.11      1050
ENSG00000000457.12      395
ENSG00000000460.15      98
ENSG00000000938.11      123
ENSG00000000971.14      757
ENSG00000001036.12      3713
ENSG00000001084.9       1649
ENSG00000001167.13      600
ENSG00000001460.16      187
ENSG00000001461.15      1259
ENSG00000001497.15      3482
ENSG00000001561.6       1672
ENSG00000001617.10      2739
ENSG00000001626.13      3
ENSG00000001629.8       3466
ENSG00000001630.14      2905
ENSG00000001631.13      1301
ENSG00000002016.15      398
```

BED output format: We merge the three original GDC files in one single BED file with the following fields:

1. **chrom** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5] according to the Ensembl ID of the gene, completed with "chr", e.g., "chr2")
2. **start** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5] according to the Ensembl ID of the gene, e.g., 32277910)

[5] *GDC.h38 GENCODE v22 GTF annotation file: https://api.gdc.cancer.gov/data/fe1750e4-fc2d-4a2c-ba21-5fc969a24f27*

3. **end** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5] according to the Ensembl ID of the gene, e.g., 32316594)
4. **strand** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5] according to the Ensembl ID of the gene, e.g., '+')
5. **ensembl_gene_id (**equal to the 1. field of any of the GDC gene expression quantification files, e.g., "ENSG00000119820.9")
6. **entrez_gene_id (**retrieved from the Genome annotation file of NCBI[6] according to the human gene symbol. If it is not found, than it is retrieved from the gene history file of NCBI[7] according to the human gene symbol. Otherwise, if it is not found from the NCBI sources, it is retrieved from HUGO Gene Nomenclature Committee (HGNC)[8] according to the human gene symbol, e.g., "YIPF4")
7. **gene_symbol** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5] according to the Ensembl ID of the gene, e.g., "YIPF4")
8. **type** (retrieved from GDC.h38 GENCODE v22 GTF annotation files[5] according to the Ensembl ID of the gene, e.g., "gene")
9. **htseq_count (**equal to the 2. field of the GDC counts file, e.g., 1320)
10. **fpkm_uq (**equal to the 2. field of the GDC FPKM-UQ file, e.g., 88737.5390983
11. **fpkm (**equal to the 2. field of the GDC FPKM file, e.g., 2.44783943057)

BED file example

chr2	32277910	32316594	+	ENSG00000119820.9	84272	YIPF4	gene	1320	88737.5390983	2.44783943057
chr15	20835372	20866314	-	ENSG00000230031.9	100287399	POTEB2	gene	0	0.0	0.0
chr6	166240290	166240493	-	ENSG00000213536.2	2789	GNG5P1	gene	1	4031.21775026	0.111201796473
chrX	50202713	50351910	+	ENSG00000147082.16	85417	CCNB3	gene	44	6690.86733105	0.184568662193
chr3	3799437	3847703	+	ENSG00000175928.5	57633	LRRN1	gene	78	15043.3247754	0.414972557569
chr22	29438583	29442455	+	ENSG00000128250.5	5988	RFPL1	gene	3	1649.1345342	0.0454916440115
chrX	152698752	152702347	+	ENSG00000221867.7	4102	MAGEA3	gene	0	0.0	0.0
chr5	70925030	70953942	+	ENSG00000172062.15	6606	SMN1	gene	136	36113.0465816	0.996184256163
chr14	105672308	105673314	-	ENSG00000213140.3	2003	ELK2AP	gene	0	0.0	0.0

[6] ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/ANNOTATION_RELEASE.107/GFF/ref_GRCh38.p2_top_level.gff3.gz
[7] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_history.gz
[8] Queries to HUGO Gene Nomenclature Committee (HGNC) are performed according to the following REST query http://rest.genenames.org/fetch/symbol/ followed by gene symbol, e.g., http://rest.genenames.org/fetch/symbol/BRCA1

Methylation Beta Value

A wide-spread NGS experiment is the large-scale analysis of DNA methylation, which consists in deep sequencing of bisulfite-treated DNA. DNA methylation can be defined as the covalent modification of cytosine bases at the C-5 position, generally within a CpG sequence context. If DNA methylation occurs in promoter regions, it is an epigenetic mark that represents the inactivity of the transcripts of the promoter gene.
More details are described in the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_LO_Pipeline/.
We consider both Illumina Infinium HumanMethylation27 (HM27) and HumanMethylation450 (HM450) DNA methylation platforms. They are used for measuring the level of methylation at 27578 / 485577 known CpG sites as beta values. Using probe sequence information provided in the manufacturer's manifest, HM27 and HM450 probes were remapped to the GRCh38 reference genome. The HM27 and HM450 manifest files are available at https://www.ncbi.nlm.nih.gov/geo/download/?acc=GPL8490&format=file&file=GPL8490%5FHumanMethylation27%5F270596%5Fv%2E1%2E2%2Ecsv%2Egz and ftp://webdata2:webdata2@ussd-ftp.illumina.com/downloads/ProductFiles/HumanMethylation450/HumanMethylation450_15017482_v1-2.csv, respectively.
These probe coordinates were then used to identify the associated transcripts from GENCODE v22, the associated CpG island (CGI), and the CpG sites' distance from each of these features. Multiple transcripts overlapping the target CpG were separated with semicolons. Beta values were inherited from existing TCGA Level 3 DNA methylation data (hg19-based) based on Probe IDs.
GDC reports for each methylated site a list of gene symbols that are associated with it. Genes that fall within 1,500 bp from the methylated site are used, considering the gene as starting from the transcription start site (TSS) to the end of the gene body.

Input:
One tab-delimited file is provided by GDC for each aliquot, with the following fields:

1. composite_element_ref (i.e., the composite element reference, used to record the location of what is aligned to the considered assembly; it is an unique ID for the array probe associated with a CpG site; the IDs that begin with the prefix "cg" are Illumina probe IDs of CpG-targeting probes; the IDs that begin with the prefix "ch" are illumina probe IDs of non-CpG-targeting probes; the IDs that start with the prefix "rs" refer to methylated sites, which overlap well known SNPs, therefore NCBI SNP IDs are used);
2. beta_value (i.e., the ratio between the methylated array intensity and total array intensity, falling between 0 (lower levels of methylation) and 1 (higher levels of methylation); missing values (i.e., not measured or with unreliable measurement) are encoded with "NA");
3. chr (i.e., the chromosome in which the probe binding site is located);

4. start (i.e., the starting position of the probed CpG dinucleotide (a CpG island is where a cytosine nucleotide occurs next to a guanine nucleotide));
5. end (i.e., the ending position of the probed CpG dinucleotide (a CpG island is where a cytosine nucleotide occurs next to a guanine nucleotide));
6. gene_symbol (i.e., the symbol of each of the genes (can be more than one, separated by the ; char) associated with the CpG site. Genes that fall within 1,500 bp from the methylated site are used, considering the gene as starting from the transcription start site (TSS) to the end of the gene body. The same gene symbol is repeated if more than one transcript_id (reported in field 8) is associated with it.)
7. gene_type (i.e., a general classification for each associated gene (e.g., protein coding, miRNA, pseudogene), separated by the ; char);
8. transcript_id (i.e., Ensembl transcript ID of each transcript associated with the genes detailed above, separated by the ; char);
9. position_to_tss (i.e., distance in base pairs of the CpG site from each associated transcript's start site, separated by the ; char; negative values indicate that the CpG site is located downstream with respect to the TSS);
10. cgi_coordinate (CpG island coordinate, i.e., the start and end coordinates of the CpG island associated with the CpG site);
11. feature_type (i.e., the position of the CpG site in reference to the island: Island, or N_Shore, or S_Shore (0-2 kb upstream, or downstream from CGI), or N_Shelf, or S_Shelf (2-4 kbp upstream or downstream from CGI)) CpG island shores are 0–2 kb from CGI, CpG island shelves are 2–4 kb from CGI, N stands for upstream, S for downstream. For more details the reader may refer to http://www.sciencedirect.com/science/article/pii/S0888754311001807.
    "Methylated cytosines can be in CpG islands, shores, shelves, open sea, and sites surrounding transcription sites [−200 to −1500 bp, 5′ untranslated region (UTR), and exons 1] for coding genes as well as gene bodies and 3′UTR and other/open sea regions derived from genome-wide association studies. Shores are considered regions 0–2 kb from CpG islands, shelves are regions 2–4 kb from CpG islands, and other/open sea regions are isolated CpG sites in the genome that do not have a specific designation." In this last case the feature_type is not defined and encoded with ".". [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3387424/]

Each row in the input file refers to a single CpG island.

Input example

```
cg00000029      0.464333545084658       chr16   53434200        53434201        RBL2;RBL2;RBL2  protein_coding;protein_coding;protein_coding
        ENST00000262133.9;ENST00000544405.5;ENST00000567964.5   -221;-1420;222  CGI:chr16:53434489-53435297     N_Shore

cg00024396      0.0393555284862584      chr6    53349210        53349211        ELOVL5;ELOVL5;ELOVL5;ELOVL5;ELOVL5;ELOVL5;ELOVL5;RP3-483K16.4
        protein_coding;protein_coding;protein_coding;protein_coding;protein_coding;protein_coding;protein_coding;lincRNA
ENST00000304434.9;ENST00000370913.5;ENST00000370918.7;ENST00000465983.4;ENST00000485336.4;ENST00000486973.1;ENST00000542638.4;ENST00000605281.1
        -202;-283;-30;-259;-236;-259;-30;-949   CGI:chr6:53347819-53349245      Island

cg00000289      0.775168406929978       chr14   68874422        68874423        ACTN1;ACTN1;ACTN1;ACTN1
protein_coding;protein_coding;protein_coding;protein_coding     ENST00000193403.9;ENST00000394419.7;ENST00000553882.1;ENST00000556083.1
105019;104818;4601;13842        CGI:chr14:68874710-68875103     N_Shore
```

BED output format: Tab-separated BED file, in which the DNA methylation file is converted, with the following fields:

1. **chrom** (equal to the 3. field of the GDC DNA methylation file, i.e., the chromosome in which the probe binding site is located, e.g., "chr16")

2. **start** (equal to the 4. field of the GDC DNA methylation file, i.e., the starting position of the probed CpG dinucleotide, since methylation involves a single base and the used genomic coordinate system is 1-based, e.g., 53434200)
3. **end** (equal to the 5. field of the GDC DNA methylation file, i.e., the ending position of the probed CpG dinucleotide since methylation involves a single base and the used genomic coordinate system is 1-based, e.g., 53434201)
4. **strand** (retrieved from GDC.h38 GENCODE v22 GTF annotation file[5], based on the human gene symbol provided in 6. field of this output file, e.g., '+')
5. **composite_element_ref** (equal to the 1. field of the GDC DNA methylation file, e.g., "cg00000092")
   beta_value (equal to the 2. field of the GDC DNA methylation file, e.g., 0.157004810973011; it is worth to note that we filter out the methylation sites with missing beta values (i.e., not measured or with unreliable measurement), which were originally encoded with "NA".)
6. **gene_symbol** (the symbol of the gene region where the CpG dinucleotide is located, e.g., "RBL2"; retrieved from field 6 of the input file and GDC.h38 GENCODE v22 GTF annotation file[5]; if the CpG dinucleotide is outside a gene region, we report the gene symbol that is at minimum bp distance from the CpG dinucleotide, as retrieved from field 6 of the input file and GDC.h38 GENCODE v22 GTF annotation file[5])
7. **entrez_gene_id** (retrieved from the Genome annotation file of NCBI[6] according to the human gene symbol. If it is not found, than it is retrieved from the gene history file of NCBI[7] according to the human gene symbol. Otherwise, if it is not found from the NCBI sources, it is retrieved from HUGO Gene Nomenclature Committee (HGNC)[8] according to the human gene symbol provided in the 6. field of this output file, e.g., 5934)
8. **gene_type** (type of gene provided in the 6. field of this output file, e.g., "protein_coding"; retrieved from the 7. field of the GDC DNA methylation file)
9. **ensembl_transcript_id** (Ensembl IDs of the transcripts related to the gene provided in the 6. field of this output file, e.g., "ENST00000544405.5|ENST00000262133.9", retrieved from the 8. field of the GDC DNA methylation file)
10. **position_to_tss** (distances in base pairs of the CpG site from each associated transcript's start site, related to the transcripts provided in the 9. field of this output file; negative values indicate that the CpG site is located downstream with respect to the TSS, e.g., "-221|-1420|222"; retrieved from the 9. field of the GDC DNA methylation file)
11. **all_gene_symbols** (equal to the 6. field of the GDC DNA methylation file, i.e., the symbol of each of the genes (can be more than one, separated by the ; char) associated with the CpG site, e.g., "RBL2, COX")
12. **all_entrez_gene_ids** (retrieved from HUGO Gene Nomenclature Committee (HGNC)[8] according to the gene symbols provided in the 11. field of this output file, e.g., 5934;1253;4861)
13. **all_gene_types** (equal to the 7. field of the GDC DNA methylation file, by taking into account the corresponding gene symbol (can be more than one, separated by the ; char) in field 11 of this output file, e.g., "protein_coding")
14. **all_ensembl_transcript_ids** (equal to the 8. field of the GDC DNA methylation file, i.e., Ensembl transcript ID of each transcript associated with the corresponding gene symbol (can be more than one, separated by the ; char) in field 11 of this output file, e.g., "ENST00000155840.8|ENST00000335475.5;ENST00000597346.1"), pipe delimits transcript IDs related to the same gene, semicolon the ones related to different genes
15. **all_positions_to_tss** (equal to the 9. field of the GDC DNA methylation file, i.e., distance in base pairs of the CpG site from each associated transcript's start site, by taking into account the corresponding gene symbol (can be more than one, separated by the ; char), negative values indicate that the CpG site is located downstream with respect to the TSS, e.g., "254241|237796;762"), pipe delimits positions_to_tss related to the same gene, semicolon the ones related to different genes
16. **cgi_coordinate** (equal to the 10. field of the GDC DNA methylation file, i.e., the start and end coordinates of

the CpG island associated with the CpG site, e.g., "CGI:chr16:53434489-53435297")

17. **feature_type** (equal to the 11. field of the GDC DNA methylation file, i.e., the position of the CpG site in reference to the island, e.g., "N_Shore")

BED file example

```
chr16   53434200        53434201        +       cg00000029      0.464333545084658       RBL2    5934    protein_coding
ENST00000262133.9|ENST00000544405.5|ENST00000567964.5   -221|-1420|222  RBL2    5934    protein_coding  ENST00000262133.9|
ENST00000544405.5|ENST00000567964.5     -221|-1420|222  CGI:chr16:53434489-53435297     N_Shore

chr1    43365370        43365371        -       cg00001446      0.918395118276287       ELOVL1  64834   protein_coding
ENST00000372458.6|ENST00000413844.3|ENST00000464204.4|ENST00000465321.4|ENST00000468865.5|ENST00000470769.4|ENST00000470968.5|
ENST00000478481.4|ENST00000479439.4|ENST00000479686.4|ENST00000482302.4|ENST00000487209.4|ENST00000496932.1|ENST00000497050.4|
ENST00000497569.4|ENST00000621943.3     2649|2705|2638|2634|-101|1218|2656|-155|960|2247|2047|2630|2596|2369|1735|2369
ELOVL1;MIR6734  64834;102466723 protein_coding;miRNA    ENST00000372458.6|ENST00000413844.3|ENST00000464204.4|ENST00000465321.4|
ENST00000468865.5|ENST00000470769.4|ENST00000470968.5|ENST00000478481.4|ENST00000479439.4|ENST00000479686.4|ENST00000482302.4|
ENST00000487209.4|ENST00000496932.1|ENST00000497050.4|ENST00000497569.4|ENST00000621943.3;ENST00000621166.1     2649|2705|2638|
2634|-101|1218|2656|-155|960|2247|2047|2630|2596|2369|1735|2369;-654    CGI:chr1:43367143-43367402      N_Shore


chr14   68874422        68874423        -       cg00000289      0.775168406929978       ACTN1   87      protein_coding
ENST00000193403.9|ENST00000394419.7|ENST00000553882.1|ENST00000556083.1 105019|104818|4601|13842        ACTN1   87
protein_coding  ENST00000193403.9|ENST00000394419.7|ENST00000553882.1|ENST00000556083.1 105019|104818|4601|13842
CGI:chr14:68874710-68875103     N_Shore
```

Copy Number Segment and Masked Copy Number Segment

A copy number variation (CNV) is a variation in the number of copies of a given genomic segment per cell.
More details are described in the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/.

Two different data types (both related to CNVs) are provided by GDC:

a) Copy Number Segment (includes both germline and somatic CNVs)
b) Masked Copy Number Segment (includes only somatic CNVs)

For the Copy Number Segment data type, the experiments have the suffix "grch38.seg" and they include both germline and somatic CNVs. Instead, for the Masked Copy Number Segment data type, the suffix for each experiment is "nocnv_grch38.seg" and it includes only somatic CNVs.

The internal representation of the files for both Copy Number Segment and Masked Copy Number Segment is the same. This is the reason why the following Input and Output paragraph is reported only once.

Input:
A single experiment is represented by a tab delimited file with the following fields:

1. Sample (i.e., the GDC internal sample ID)
2. Chromosome (i.e., the name or number of the chromosome where the CNV is located)
3. Start (i.e., the starting position of the CNV feature in the chromosome)
4. End (i.e., the ending position of the CNV feature in the chromosome)
5. Num_Probes (i.e., the number of consecutive probes that comprise the genome segment with the CNV)
6. Segment_Mean (i.e., the estimated Copy Number (CN) ratio for the segment, that is the $\log_2$ ratio of the tumor intensity of CN to the normal intensity of CN; use $(2^{Segment\_Mean}) * 2$ to convert to absolute CN)[9]

Each row in the input file refers to a single CNV.

[9] https://www.biostars.org/p/112310/

Sample	Chromosome	Start	End	Num_Probes	Segment_Mean
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	61735	1628826	229	0.1756
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	1642103	1688058	20	0.8677
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	1688192	16149915	8139	0.0169
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	16153497	16154239	8	1.105
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	16154966	25570830	5697	0.0116
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	25571269	25696602	56	-0.4542
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	25698469	35091674	4921	0.0113
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	35102654	35104491	20	-0.608
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	35114268	72768916	23688	0.0027
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	72768936	72811133	44	-1.8052
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	72811148	76050844	1908	-0.0045
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	76054763	76054854	2	-2.6875
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	76059509	86573546	7067	-0.0077
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	86573802	86577211	2	-2.1489
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	86577870	99732202	8251	0.0046
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	99732737	99737222	2	-1.956
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	99737524	104163499	2699	0.003
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	104163787	104303403	27	-0.7798
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	104303501	110224427	3562	-0.0077
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	110225642	110232974	14	-0.5318
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	110233053	110240178	14	-1.2134
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	110242953	152759678	10146	0.009
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	152761923	152768700	37	-1.5703
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	152773905	161479438	5226	0.0031
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	161496900	161648237	56	0.847
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	161648621	210071062	32856	0.0011
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	210081613	210083984	3	-2.6172
AQUAE_p_TCGA_112_304_b2_N_GenomeWideSNP_6_A01_1348356	1	210086552	222366668	8539	-1e-04

BED output format: Tab separated BED file, in which the CNV_file is converted, with the following fields:

1. **chrom** (equal to the 2. field of the GDC CNV file, e.g., "1")
2. **start** (equal to the 3. field of the GDC CNV file, e.g., 61735)
3. **end** (equal to the 4. field of the GDC CNV file, e.g., 1628826)
4. **strand** (unknown, set to '*')
5. **num_probes** (equal to the 5. field of the GDC CNV file, e.g., 229)
6. **segment_mean** (equal to the 6. field of the GDC CNV file, e.g., 0.1756)

BED file example

chr1	61735	6016361	*	2835	-0.3124
chr1	6019570	6019642	*	2	-2.2437
chr1	6020227	13326062	*	3737	-0.2954
chr1	13338980	13362453	*	8	-1.3935
chr1	13366082	15823420	*	1815	-0.3037
chr1	15827002	15827706	*	7	0.4437
chr1	15827744	16684955	*	350	-0.3384
chr1	16685015	16721910	*	33	0.1203
chr1	16721984	16864367	*	26	-0.5167
chr1	16868660	16935752	*	61	-0.0202
chr1	16949746	21992508	*	3344	-0.3036
chr1	21994022	22019085	*	12	-0.8921
chr1	22019154	25256800	*	1908	-0.2857
chr1	25256850	25278567	*	13	0.5545
chr1	25284629	25335514	*	18	0.1529
chr1	25335721	45151640	*	10907	-0.2779
chr1	45153815	64961923	*	13039	-0.3037
chr1	64963532	64964114	*	6	-1.2978
chr1	64969380	72167620	*	4695	-0.3166
chr1	72171216	72303233	*	88	-0.2252
chr1	72303253	72345450	*	44	-0.87
chr1	72345465	99681022	*	17648	-0.308
chr1	99682647	99683312	*	2	-2.4127

miRNA Expression Quantification

miRNA-seq data are derived from the sequencing of micro RNAs (miRNA). They contain information about both nucleotide sequence and expression. More details are described in the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/.

One file for each aliquot is provided by GDC, containing the expression calculated based on all reads aligning to a particular miRNA.

Input:
One tab delimited file is provided by GDC for each aliquot, with the following fields:

1. miRNA_ID (i.e., a valid miRBase ID (http://www.mirbase.org/))
2. read_count (i.e., the sum of fractions of reads that mapped to a miRNA)
3. reads_per_million_miRNA_mapped (i.e., millions of reads that mapped to a miRNA)
4. cross-mapped (i.e., cross-mapped to other miRNA forms (Y or N))

Each row in the input file refers to a single miRNA.

miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-mapped
hsa-let-7a-1	76213	13484.031491	N
hsa-let-7a-2	151321	26772.560183	Y
hsa-let-7a-3	77498	13711.380899	N
hsa-let-7b	85979	15211.886995	N
hsa-let-7c	11107	1965.112747	Y
hsa-let-7d	9740	1723.255438	N
hsa-let-7e	15161	2682.369168	N
hsa-let-7f-1	261	46.177584	N
hsa-let-7f-2	94960	16800.855895	N
hsa-let-7g	6601	1167.885950	N
hsa-let-7i	1550	274.234695	N
hsa-mir-1-1	0	0.000000	N
hsa-mir-1-2	30	5.307768	N
hsa-mir-100	1677	296.704247	N
hsa-mir-101-1	45395	8031.538051	N
hsa-mir-101-2	377	66.700955	N
hsa-mir-103-1	126526	22385.689691	Y
hsa-mir-103-2	57	10.084760	N
hsa-mir-105-1	1	0.176926	N
hsa-mir-105-2	2	0.353851	N
hsa-mir-106a	11	1.946182	Y
hsa-mir-106b	1060	187.541146	N
hsa-mir-107	143	25.300362	Y
hsa-mir-10a	195986	34674.942539	N
hsa-mir-10b	1655780	292949.885998	N
hsa-mir-1178	0	0.000000	N
hsa-mir-1179	2	0.353851	N
hsa-mir-1180	258	45.646807	N

BED output format: Tab separated BED file, in which the miRNA-seq Mirna quantification file is converted, with the following fields:

1. **chrom** (retrieved from miRBase database[10], according to the miRNA id provided in field 5, e.g., "chr9")
2. **start** (retrieved from miRBase database[10], according to the miRNA id provided in field 5, e.g., 94175957)
3. **end** (retrieved from miRBase database[10], according to the miRNA id provided in field 5, e.g., 94176036)
4. **strand** (retrieved from miRBase database[10], according to the miRNA id provided in field 5, e.g., '+')
5. **mirna_id** (equal to the 1. field of the GDC miRNA-seq file, e.g., "hsa-let-7a-1")
6. **read_count** (equal to the 2. field of the GDC miRNA-seq file, e.g., 29726)
7. **reads_per_million_mirna_mapped** (equal to the 3. field of the GDC miRNA-seq file, e.g., 12429.699816)
8. **cross-mapped** (equal to the 4. field of the GDC miRNA-seq file, e.g., 'N')
9. **entrez_gene_id** (retrieved from HUGO Gene Nomenclature Committee (HGNC)[11] starting from the **mirna_id** provided in field 5)
10. **gene_symbol** (retrieved from HUGO Gene Nomenclature Committee (HGNC)[12] starting from the entrez_gene_id retrieved in field 9)

BED file example

chr9	94175957	94176036	+	hsa-let-7a-1	141272	21050.8717	N	406881	MIRNLET7A1
chr11	122146522	122146593	-	hsa-let-7a-2	141458	21078.58747	N	406882	MIRNLET7A2
chr22	46112749	46112822	+	hsa-let-7a-3	141840	21135.5091	N	406883	MIRNLET7A3
chr22	46113686	46113768	+	hsa-let-7b	78222	11655.822	N	406884	MIRLET7B
chr21	16539828	16539911	+	hsa-let-7c	12732	1897.189099	N	406885	MIRLET7C
chr9	94178834	94178920	+	hsa-let-7d	1876	279.541843	N	406886	MIRLET7D
chr19	51692786	51692864	+	hsa-let-7e	38600	5751.767141	N	406887	MIRLET7E
chr9	94176347	94176433	+	hsa-let-7f-1	123324	18376.44899	N	406888	MIRNLET7F1
chrX	53557192	53557274	-	hsa-let-7f-2	126337	18825.41465	N	406889	MIRNLET7F2
chr3	52268278	52268361	-	hsa-let-7g	5619	837.284445	N	406890	MIRLET7G
chr12	62603686	62603769	+	hsa-let-7i	1190	177.321319	N	406891	MIRLET7I
chr11	122152229	122152308	-	hsa-mir-100	8211	1223.517098	N	406892	MIRN100
chr1	65058434	65058508	-	hsa-mir-101-1	46212	6886.027542	N	406893	MIR101-1
chr9	4850297	4850375	+	hsa-mir-101-2	47482	7075.269622	N	406894	MIR101-2

[10] Used GRCh38 data are retrieved from the version 21 of the miRBase database at ftp://mirbase.org/pub/mirbase/21/
[11] Queries to HUGO Gene Nomenclature Committee (HGNC) are performed according to the following rest query http://rest.genenames.org/fetch/hgnc_id/ followed by the **hgnc_id**; the **hgnc_id** is also retrieved from HUGO starting from the **mirna_id** provided in field 1 of the input file
[12] Queries to HUGO Gene Nomenclature Committee (HGNC) are performed according to the following REST query http://rest.genenames.org/fetch/symbol/ followed by the entrez id

Isoform Expression Quantification

The miRNA Isoform Expression Quantification data contain expression profiles calculated for each individual miRNA sequence isoform observed.
More details are described in the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf and at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/.
GDC provides one file for each aliquot.

Input:
One tab delimited file is provided by GDC for each aliquot, with the following fields:

1. miRNA_ID (i.e., a valid miRBase ID (http://www.mirbase.org/))
2. isoform_coords (i.e., Alignment coordinates as <version>:<Chromosome>:<Start position>-<End position>:<Strand>)
3. read_count (i.e., count of raw reads that mapped to a miRNA isoform)
4. reads_per_million_miRNA_mapped (i.e., millions of reads that mapped to a miRNA isoform)
5. cross-mapped (i.e., cross-mapped to other miRNA forms (Y or N))
6. miRNA_region (i.e., miRBase accession number of a class of miRNA sequence, e.g., mature, stemloop, ...)

Each row in the input file refers to a single isoform.

miRNA_ID	isoform_coords	read_count	reads_per_million_miRNA_mapped	cross-mapped	miRNA_region
hsa-let-7a-1	hg38:chr9:94175961-94175979:+	1	0.213099	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175980:+	2	0.426199	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175981:+	1	0.213099	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175982:+	13	2.770.290	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175983:+	17	3.622.687	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175984:+	45	9.589.466	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175985:+	2	0.426199	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175981:+	373	79.486.022	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175982:+	15219	3.243.157.543	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175983:+	13148	2.801.828.988	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175984:+	43064	9.176.906.263	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175985:+	817	174.102.090	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175986:+	25	5.327.481	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175963-94175982:+	1	0.213099	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175963-94175984:+	10	2.130.993	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175965-94175982:+	2	0.426199	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175965-94175983:+	2	0.426199	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175965-94175984:+	4	0.852397	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175966-94175984:+	2	0.426199	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175967-94175988:+	1	0.213099	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175984-94176007:+	2	0.426199	N	stemloop

BED output format: Tab separated BED file, in which the miRNA-seq Isoform quantification file is converted, with the following fields:

1. **chrom** (retrieved from the 2. field of the GDC miRNA-seq file, part just after the first ":", e.g., "chr9")
2. **start** (retrieved from the 2. field of the GDC miRNA-seq file, part just after the second ":", e.g., 96938243)
3. **end** (retrieved from the 2. field of the GDC miRNA-seq file, part just before the third ":", e.g., 96938264)
4. **strand** (retrieved from the 2. field of the GDC miRNA-seq file, part just after the third ":", e.g., '+')
5. **genome_version** (retrieved from the 2. field of the GDC miRNA-seq file, part just before the first ":", e.g., "hg38")
6. **mirna_id** (equal to the 1. field of the GDC miRNA-seq file, e.g., "has-let-7a-1")
7. **read_count** (equal to the 3. field of the GDC miRNA-seq file, e.g., 4)
8. **reads_per_million_mirna_mapped** (equal to the 4. field of the GDC miRNA-seq file, e.g., 0.707702)
9. **cross-mapped** (equal to the 5. field of the GDC miRNA-seq file, e.g., 'N')
10. **mirna_region** (equal to the 6. field of the GDC miRNA-seq file, e.g., "mature, MIMAT0000062")
11. **entrez_gene_id** (retrieved from HUGO Gene Nomenclature Committee (HGNC)[11] starting from the **mirna_id** provided in field 6)
12. **gene_symbol** (retrieved from HUGO Gene Nomenclature Committee (HGNC)[12] starting from the **entrez_gene_id** provided in field 11)

BED file example

chr9	94175943	94175962	+	hg38	hsa-let-7a-1	1	0.097527	N	precursor	406881	MIRNLET7A1
chr9	94175961	94175982	+	hg38	hsa-let-7a-1	18	1.755491	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175961	94175983	+	hg38	hsa-let-7a-1	17	1.657963	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175961	94175984	+	hg38	hsa-let-7a-1	47	4.583781	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175981	+	hg38	hsa-let-7a-1	426	41.546615	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175982	+	hg38	hsa-let-7a-1	14255	1390.251155	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175983	+	hg38	hsa-let-7a-1	13823	1348.119377	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175984	+	hg38	hsa-let-7a-1	48839	4763.13407	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175985	+	hg38	hsa-let-7a-1	790	77.046539	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175962	94175986	+	hg38	hsa-let-7a-1	14	1.365382	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175963	94175981	+	hg38	hsa-let-7a-1	1	0.097527	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175963	94175982	+	hg38	hsa-let-7a-1	5	0.487636	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175963	94175983	+	hg38	hsa-let-7a-1	5	0.487636	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175963	94175984	+	hg38	hsa-let-7a-1	18	1.755491	N	mature,MIMAT0000062	406881	MIRNLET7A1
chr9	94175963	94175985	+	hg38	hsa-let-7a-1	1	0.097527	N	mature,MIMAT0000062	406881	MIRNLET7A1

Meta data: Clinical and Biospecimen Supplements

Clinical and Biospecimen Supplements contain information about the patients (e.g., gender, race, weight, vital status, treatment, etc.) and the experiments conducted on normal and/or tumoral tissues of such patients (e.g., experiment name, disease type, tissue type, etc.), respectively.

More details about the attributes contained in Clinical Supplement data are available at https://gdc.cancer.gov/about-data/data-harmonization-and-generation/clinical-data-harmonization. The attributes contained in Biospecimen Supplement data are listed and explained at https://gdc.cancer.gov/about-data/data-harmonization-and-generation/biospecimen-data-harmonization. The reader may also refer to the GDC Data User's Guide available at https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf.

Input: **Clinical and Biospecimen Supplements**
For the TCGA project GDC provides two XML files for each patient, the first one (Clinical) containing patient clinical data, the second one (Biospecimen) containing specimen data. An example of these files is available at:

1. Clinical – https://gdc-api.nci.nih.gov/data/0bf20449-4129-4183-80ad-5e1eec2f84ea
2. Biospecimen – https://gdc-api.nci.nih.gov/data/1be29e3c-c23d-4870-9329-972a28ccf160

For the TARGET project GDC provides several Spreadsheet XLSX files for each patient, containing patient clinical data (Clinical) and specimen data (Biospecimen). An example of these files is available at:

1. Clinical –https://gdc-api.nci.nih.gov/data/06f24280-5215-484b-bb12-0722cee4b7d4
2. Biospecimen – https://gdc-api.nci.nih.gov/data/f42b0592-caa4-4801-b54b-956b26b7094b

Meta data output format:
One meta data tab delimited file for each aliquot (.meta), whose rows contain all the meta data attribute-value pairs for the specific aliquot, with each attribute fully specified through the double underscore ("__") delimited composition of the name of the group it belongs to and the name of the attribute. It is worth noting that every attribute contained in the file is codified to be a valid Java variable. This characteristic is required for each attribute to be correctly interpreted as valid search key. The name of these files corresponds to the aliquot UUID of a single experiment concatenated with the acronym of the considered experiment, e.g., `007a5a35-5614-52d3-8393-7642ecf84933-geq.bed.meta`, where geq stands for gene expression quantification. See subsection "Input data sets" for the acronyms associated to the experiments. When no experiment is associated to the meta data file, than we use the acronym "xxx", e.g., 0003c0e6-4e9e-544e-8ee7-55749e121895-xxx.bed.meta

Meta data in the TCGA project

For the TCGA project the meta data attribute biospecimen__bio__bcr_aliquot_uuid identifies a single experiment on a tissue of a patient and is used as primary identifier for the sequencing/array experiment. The tissue is identified with the biospecimen__bio__bcr_sample_uuid and the patient by the biospecimen__shared__bcr_patient_uuid.

Moreover, we add additional meta data attributes, within a specific group named `manually_curated`. These attributes are not present in the input files, but are retrieved separately with ad-hoc queries to the GDC API. We consider the following ones (each one is reported with an output example value):

1) **manually_curated__analysis__analysis_id** (*required*)
   5a0d2d27-d752-4109-b9fd-92575b363ecf | …
2) **manually_curated__analysis__workflow_link** (*required*)
   https://github.com/NCI-GDC/met-liftover-tool | ...
3) **manually_curated__analysis__workflow_type** (*required*)
   Liftover
4) **manually_curated__audit_warning**
   missed required metadata | …
5) **manually_curated__cases__case_id** (*required*)
   c8898b42-b704-45a0-9829-144b98f416e0 | …
6) **manually_curated__cases__demographic__year_of_birth** (*required*)
   1977 | …
7) **manually_curated__cases__disease_type** (*required*)
   Adrenocortical Carcinoma | …
8) **manually_curated__cases__primary_site** (*required*)
   Adrenal Gland | …
9) **manually_curated__cases__project__program__name** (*required*)
   TCGA | TARGET
10) **manually_curated__cases__project__program__program_id** (*required*)
    b80aa962-9650-5110-b3eb-bd087da808db | …
11) **manually_curated__cases__submitter_id** (*required*)
    TCGA-OR-A5J6 | ..
12) **manually_curated__data_category** (*required*)
    DNA Methylation | …
13) **manually_curated__data_format** (*required*)
    BED
14) **manually_curated__data_type** (*required*)
    Methylation Beta Value
15) **manually_curated__exp_data_bed_url** (*required*)
    ftp://bioinformatics.iasi.cnr.it/opengdc/bed/tcga/tcga-acc/copy_number_segment/c00b53a9-bb48-4841-974d-7087eacd5420-cns.bed

16) **manually_curated__exp_metadata_url** (required)
ftp://bioinformatics.iasi.cnr.it/opengdc/bed/tcga/tcga-acc/clinical_and_biospecimen_supplements/c00b53a9-bb48-4841-974d-7087eacd5420-cns.meta
17) **manually_curated__experimental_strategy** (*required*)
Methylation Array
18) **manually_curated__file_id** (*required*)
471788f6-ce7f-4267-a30a-ff179bceff3b | …
19) **manually_curated__file_name** (*required*)
jhu-usc.edu_ACC.HumanMethylation450.1.lvl-3.TCGA-OR-A5J6-01A-31D-A29J-05.gdc_hg38.txt | …
20) **manually_curated__file_size** (*required*)
141305308 | …
21) **manually_curated__opengdc_id** (*required*)
00b8b899-6191-4169-91bd-a507c326e44d-msm
22) **manually_curated__platform** (*required*)
Illumina Human Methylation 450 | …
23) **manually_curated__source_data_format** (*required*)
txt
24) **manually_curated__tissue_status** (*required*)
control | normal | tumoral | undefined

The keyword *required* indicates that the metadata attributes are mandatory and always present in GDC for public data. For private data not all required fields are released.

Values of attribute *manually_curated__tissue_status* are defined based on value of attribute *biospecimen__bio__sample_type_id* (whose value in range 01 – 09 and 40 indicates a tumor type, in range 10 – 14 indicates normal type, and 20 indicates control type; the comprehensive list of sample type codes is available at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes).

Additionally, the meanings of the alphanumeric values of the attributes *biospecimen__shared__tissue_source_site* and *clinical__shared__tissue_source_site* are available at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes.

In the following we report the explanation of the identifiers present in the manually curated meta data fields:

- *manually_curated__analysis__analysis_id* is the GDC analysis UUID associated with the experiment analysis workflow

- *manually_curated__cases__case_id* is the GDC case UUID associated with the study participant (see *manually_curated__cases__submitter_id*)
- *manually_curated__cases__submitter_id* is the barcode associated with the participant or patient (i.e.; someone who contributes one or more samples to aid the research for a particular disease study)
- *manually_curated__cases__project__program__program_id* is the id of program, the highest level of organization of GDC datasets
- *manually_curated__file_id* is the GDC file UUID associated with the experimental file
- *manually_curated__opengdc_id* is the OpenGDC id associated with the experimental output file; it is composed by the aliquot *biospecimen__bio__bcr_aliquot_uuid* and the acronym of the experiment type (data type), e.g., "00b8b899-6191-4169-91bd-a507c326e44d-msm" is related to the Masked Somatic Mutations data type

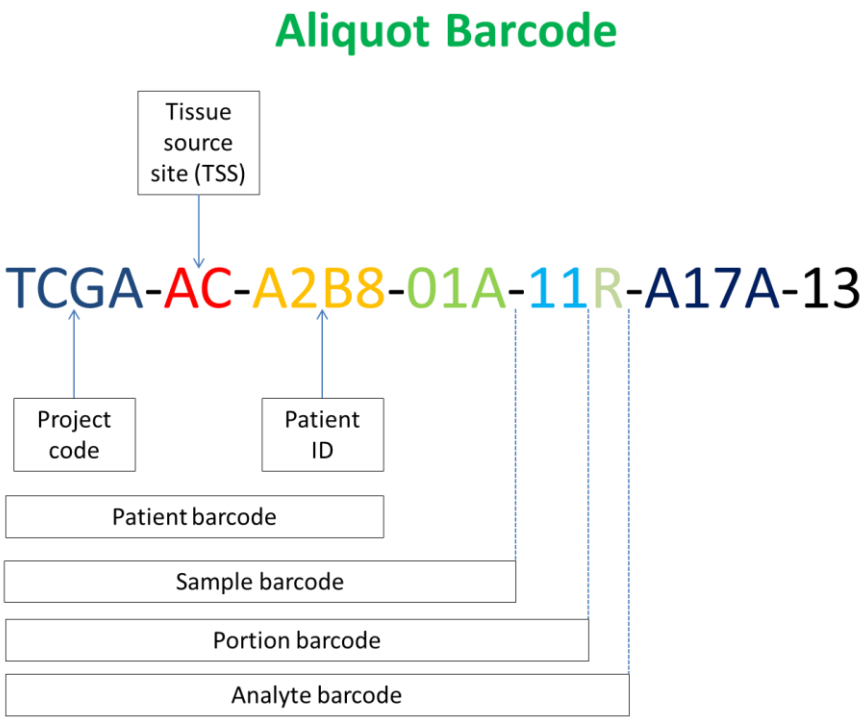
It is worth to note that the meta data *manually_curated__file_size*, *manually_curated__file_id*, *manually_curated__file_name*, *manually_curated__analysis__analysis_id*, and *manually_curated__analysis__workflow_type* of Gene Expression Quantification and Masked Somatic Mutation data provided in BED format have multiple values since such data combine data originally from three GDC files (FPKM, FPKM-UQ and counts) for Gene Expression Quantification and from four GDC files for Masked Somatic Mutation each one obtained with a different Variant caller (MuSE, MuTect2, VarScan2, and SomaticSniper)[13]; values reported in each meta data are ordered accordingly to the here above reported order of the original files they refer to. Other identifiers present in GDC meta data are the following:

Attribute	**Description**	**Example**
biospecimen__admin__file_uuid	UUID of the biospecimen file	4D24E7D2-B9CB-480B-8EFF-E8458E2C6432
biospecimen__admin__project_code	Code of the project in biospecimen file	TCGA
biospecimen__bio__bcr_aliquot_barcode	Aliquot barcode in biospecimen file	TCGA-AC-A2B8-01A-11R-A17A-13
biospecimen__bio__bcr_aliquot_uuid	Aliquot UUID in biospecimen file	40EF6B7B-8A4B-4BDB-A5F3-3742522F60E6
biospecimen__bio__bcr_analyte_barcode	Analyte barcode in biospecimen file	TCGA-AC-A2B8-01A-11R
biospecimen__bio__bcr_analyte_uuid	Analyte UUID in biospecimen file	4B2C5035-2DD2-476D-BC10-80175DF876F7
biospecimen__bio__bcr_portion_barcode	Portion barcode in biospecimen file	TCGA-AC-A2B8-01A-11
biospecimen__bio__bcr_portion_uuid	Portion UUID in biospecimen file	874BD013-CBA9-46AD-AFA5-

[13] https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/#masked-somatic-aggregation-workflow

		F2C30BB83475
biospecimen__bio__bcr_sample_barcode	Sample barcode in biospecimen file	TCGA-AC-A2B8-01A
biospecimen__bio__bcr_sample_uuid	Sample UUID in biospecimen file	FABFEC99-220B-4A94-A8FE-01DFD71487F6
biospecimen__shared__bcr_patient_barcode	Patient barcode in the biospecimen file	TCGA-AC-A2B8
biospecimen__shared__bcr_patient_uuid	Patient UUID in the biospecimen file	6E126B73-D3E8-4641-A128-306F3B313E40
biospecimen__shared__patient_id	Code of the patient in biospecimen file	A2B8
biospecimen__shared__tissue_source_site	Code of the Tissue Source Site (TSS) in biospecimen file	AC
clinical__admin__file_uuid	UUID of the clinical file	3B1D4CD1-F81B-439A-8F45-C78844AE5E08
clinical__admin__project_code	Code of the project in clinical file	TCGA
clinical__clin_shared__bcr_followup_barcode	Followup barcode in clinical file	TCGA-AC-A2B8-F43210
clinical__clin_shared__bcr_followup_uuid	Followup UUID in clinical file	E01D57EB-5162-4BCE-8AC8-D37DDBE74B3C
clinical__rad__bcr_radiation_barcode	Radiation barcode in clinical file	TCGA-AC-A2B8-R43215
clinical__rad__bcr_radiation_uuid	Radiation UUID in clinical file	DAB5FC3E-2668-4D3F-B2AD-9411CCACBF9C
clinical__rx__bcr_drug_barcode	Drug barcode in clinical file	TCGA-AC-A2B8-D43212
clinical__rx__bcr_drug_uuid	Drug UUID in clinical file	FE83C33F-CAF2-4C0F-8447-2905F8864C89
clinical__shared__bcr_patient_barcode	Patient barcode in the clinical file	TCGA-AC-A2B8
clinical__shared__bcr_patient_uuid	Patient UUID in the clinical file	6E126B73-D3E8-4641-A128-306F3B313E40
clinical__shared__patient_id	Code of the patient in clinical file	A2B8
clinical__shared__tissue_source_site	Code of the Tissue Source Site (TSS) in clinical file	AC
manually_curated__analysis__analysis_id	Id of the analysis	b758c52d-edbd-423a-9fca-52a480174137
manually_curated__cases__case_id	Id of the case (patient)	6e126b73-d3e8-4641-a128-306f3b313e40
manually_curated__cases__project__program__name	Name of the program	TCGA
manually_curated__cases__project__program__program_id	Id of the program	b80aa962-9650-5110-b3eb-bd087da808db
manually_curated__file_id	File id of the experiment file related to this metadata file	a2a05d33-e85d-4c03-84f5-69626a8236ce
manually_curated__file_name	File name of the experiment file related to this metadata file	0bc0e7ff-d391-4334-82a5-b95461fbf5ab.mirbase21.isoforms.quantification.txt

The hierarchy of the ids is depicted in the Aliquot Barcode figure:



It is worth to note that the *clinical__clin_shared__days_to_birth* meta data represents the time interval from a person's date of birth to the date of initial pathologic diagnosis, represented as a calculated negative number of days[14].

Meta data in the TARGET project

In the TARGET project the meta data attribute *manually_curated__cases__samples__portions__analytes__aliquots__aliquot_id* identifies a single experiment on a tissue of a patient and is used as primary identifier for the sequencing/array experiment. The tissue (sample) is identified with the *manually_curated__cases__samples__sample_id* and the patient by the *manually_curated__cases__case_id*. Also in this project, we add additional meta data attributes, named `manually_curated` reported above. .

[14] https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition-view&id=demographic&anchor=days_to_birth

An example of produced meta data file is shown below.

biospecimen__bio__analyte_type	DNA
biospecimen__bio__analyte_type_id	D
biospecimen__bio__bcr_aliquot_barcode	TCGA-PA-A5YG-01A-11D-A29H-01
biospecimen__bio__bcr_aliquot_uuid	0AAC83D9-DFB0-4AA8-9FC4-B12ACEF7FDEE
biospecimen__bio__bcr_analyte_barcode	TCGA-PA-A5YG-01A-11D
biospecimen__bio__bcr_analyte_uuid	8D4BD29F-A1F4-4138-B9A5-5E381137EA5F
biospecimen__bio__bcr_portion_barcode	TCGA-PA-A5YG-01A-11
...	...
clinical__acc_shared__mitotane_therapy	NO
clinical__admin__batch_number	304.63.0
clinical__admin__bcr	Nationwide Children's Hospital
clinical__admin__day_of_dcc_upload	31
clinical__admin__disease_code	ACC
clinical__admin__file_uuid	3FCC3913-872E-40A8-87FD-3B341C195A9D
...	...
manually_curated__analysis__analysis_id	af49e1d7-a57a-457f-978f-818b010ff3c6
manually_curated__analysis__workflow_link	https://github.com/NCI-GDC/dnacopy-tool
manually_curated__analysis__workflow_type	DNAcopy
manually_curated__cases__case_id	8b3649e3-16e2-4044-8ab8-f4d28a87e513
manually_curated__cases__demographic__year_of_birth	1961
manually_curated__cases__disease_type	Adrenocortical Carcinoma
...	...

Additional output files

We also provide the following output files:

MD5 checksum files

One tab-separated .txt ("*md5checksum.txt*") file for each experiment of each tumor with all the meta data and genomic data files, containing the name of the file and its md5 checksum.

Meta data dictionary file

One meta data dictionary tab-delimited file ("*meta_dictionary.txt*"), which contains all the possible values of any meta data attribute, for example:

biospecimen__bio__menopause_status
- Pre (<6 months since LMP AND no prior bilateral ovariectomy AND not on estrogen replacement)
- Peri (6-12 months since last menstrual period)
- [Unknown]
- Post (prior bilateral ovariectomy OR >12 mo since LMP with no prior hysterectomy)
- CDE_ID:2957270

clinical__clin_shared__histologic_diagnosis_other
- Mixed infiltrating lobular and grade 1 ductal carcinoma
- MUCINOUS & PAPILLARY
- CDE_ID:3124492
- Lobular carcinoma with ductal features
- ductal/lobular
- IDC+ mucinous carcinoma
- Ductal/Lobular
- Infiltrating ductal & lobular
- Infiltrating ductal and lobular carcinoma
- ductal and lobular
- Invasive ductal and lobular carcinoma
- lobular/ductal
- Mixed invasive ductal and invasive lobular
- Lobular/Ductal
- [Not Applicable]
- Mixed diagnosis
- with ductal and lobular phenotypes

When performing batch data format conversions, a meta data dictionary file is generated with all the converted data for each genomic experiment (data type) (e.g., DNA-Seq, DNA methylation, RNA-seq, miRNA-seq, and CNV) of each tumor.

Meta data information files

We output a comma separated values (CSV) file containing the occurrences of all the meta data attributes related to each experiment (data type) of each tumor ("*meta2disease_table.csv*").
Furthermore, we generate the following additional output files for each tumor:

- a CSV file containing the number of occurrences of each meta data attribute related to the tumor ("*meta2dataType_table.csv*")
- a CSV file containing a table with a list of all meta data attributes with all their possible values on the rows and the list of all available data types for the considered tumor on the columns; a generic cell of this table contains the number of occurrences of a specific attribute-value pair in a specific data type ("*meta_values2dataTypes_table.csv*")
- a tab-separated values (TSV) file containing a list of all meta data attributes with all their possible values followed by the number of occurrences of each of these pairs (attribute-value) in all data types for the considered tumor ("*meta_values2sample_list.tsv*")

Experiment information files

We generate an additional output file for each subtype of all the genomic experiments (data types), regardless the related tumor and called "*exp_info.tsv*". It is a tab-delimited file that includes:

- number of aliquots;
- number of samples (tissues);
- number of patients.

Annotations files

Gene Expression Quantification
We provide following additional annotation output files for the Gene Expression Quantification datasets:

(i) "*gene_expression_annotations.bed*", a bed file that contains the following fields for each gene in the considered genomic experiment:
   1) chrom
   2) start
   3) end
   4) strand
   5) ensembl_gene_id
   6) entrez_gene_id
   7) gene_symbol
   8) type

(ii) "*gene_expression_annotations.schema*", an xml file containing the structure and the fields of "*gene_expression_annotations.bed*"

(iii) "*gene_expression_annotations.bed.meta*", a metadata file containing following metadata related to the "*gene_expression_annotations.bed*" file:

1) **annotation_type**
   gene
2) **assembly**
   GRCh38
3) **platform**
   Illumina
4) **external_annotations_source**
   HUGO Gene Nomenclature Committee (HGNC)
5) **external_annotations_source_url**
   http://rest.genenames.org
6) **gdc_annotations_source**
   GDC.h38 GENCODE v22 GTF annotation file
7) **gdc_annotations_source_url**
   https://api.gdc.cancer.gov/data/fe1750e4-fc2d-4a2c-ba21-5fc969a24f27
8) **name**
   gene regions for GDC Gene Expression Quantification
9) **original_provider**
   GENCODE
10) **provider**
    GDC

DNA methylation

We provide following additional annotation output files for the DNA methylation datasets:

(i) "*humanMethylation27_annotations.bed*", a bed file that contains the following fields for each methylated site in the considered genomic experiment:

1) chrom
2) start
3) end
4) strand
5) composite_element_ref
6) gene_symbol
7) entrez_gene_id
8) gene_type
9) ensembl_transcript_id
10) position_to_tss
11) all_gene_symbols
12) all_entrez_gene_ids
13) all_gene_types
14) all_ensembl_transcript_ids

15) all_positions_to_tss
16) cgi_coordinate
17) feature_type

(ii) "*humanMethylation27_annotations.schema*", an xml file containing the structure and the fields of "*humanMethylation27_annotations.bed*"

(iii) "*humanMethylation27_annotations.bed.meta*", a metadata file containing following metadata related to the "*humanMethylation27_annotations.bed*" file:

1) **annotation_type**
   CpG site
2) **assembly**
   GRCh38
3) **platform**
   Illumina Human Methylation 27
4) **external_annotations_source**
   HUGO Gene Nomenclature Committee (HGNC)
5) **external_annotations_source_url**
   http://rest.genenames.org
6) **gdc_annotations_source**
   GDC.h38 GENCODE v22 GTF annotation file
7) **gdc_annotations_source_url**
   https://api.gdc.cancer.gov/data/fe1750e4-fc2d-4a2c-ba21-5fc969a24f27
8) **name**
   genomic coordinates related to the CpG site and gene regions associated to it
9) **original_provider**
   GENCODE
10) **provider**
    GDC

(iv) "*humanMethylation450_annotations.bed*", a bed file that contains the following fields for each methylated site in the considered genomic experiment:

1) chrom
2) start
3) end
4) strand
5) composite_element_ref
6) gene_symbol
7) entrez_gene_id
8) gene_type
9) ensembl_transcript_id
10) position_to_tss

11) all_gene_symbols
12) all_entrez_gene_ids
13) all_gene_types
14) all_ensembl_transcript_ids
15) all_positions_to_tss
16) cgi_coordinate
17) feature_type

(v) "*humanMethylation450_annotations.schema*", an xml file containing the structure and the fields of "*humanMethylation450_annotations.bed*"

(vi) "*humanMethylation450_annotations.bed.meta*", a metadata file containing following metadata related to the "*humanMethylation27_annotations.bed*" file:

1) **annotation_type**
   CpG site
2) **assembly**
   GRCh38
3) **platform**
   Illumina Human Methylation 450
4) **external_annotations_source**
   HUGO Gene Nomenclature Committee (HGNC)
5) **external_annotations_source_url**
   http://rest.genenames.org
6) **gdc_annotations_source**
   GDC.h38 GENCODE v22 GTF annotation file
7) **gdc_annotations_source_url**
   https://api.gdc.cancer.gov/data/fe1750e4-fc2d-4a2c-ba21-5fc969a24f27
8) **name**
   genomic coordinates related to the CpG site and gene regions associated to it
9) **original_provider**
   GENCODE
10) **provider**
    GDC

Summary table of the additional output files

Meta data	*meta_dictionary.txt*
Gene Expression Quantification	*gene_expression_annotations.bed* *gene_expression_annotations.bed.meta* *gene_expression_annotations.schema*
DNA methylation	*humanMethylation27_annotations.bed* *humanMethylation27_annotations.bed.meta* *humanMethylation27_annotations.schema* *humanMethylation450_annotations.bed* *humanMethylation450_annotations.bed.meta* *humanMethylation450_annotations.schema*
For each data type	*exp_info.tsv* *md5checksum.txt*
General	*meta2dataType_table.csv* *meta2disease_table.csv* *meta_values2dataTypes_table.csv* *meta_values2sample_list.tsv*

Additional data file formats

Besides the BED format, to ensure maximum usage, we also support the set of additional data file formats following specified.

CSV format

The standard Comma Separated Values (CSV) file format defines the structure and content of the genomic data files as equal to the ones of the BED format, but a comma (instead of a tabulator) is used to separate the different fields.
The structure of the meta data files is the same as for the BED format.

XML format

The standard eXtended Markup Language (XML) file format specifies the content of the genomic data files as equal to the one of the BED format, but the file structure is designed according to the XML style. In particular, we define one genomic data XML file for each aliquot and experiment type; the content of this file starts with the XML heading line

```
<?xml version="1.0" encoding="UTF-8"?>
```

and with the root tag called `<aliquot>`.
Then, for each genomic measure (row of the input data file) we define a `<data>` tag containing the measured attributes and their values as sub-tags.
In the following, we provide an example of XML file of DNA methylation:

```
<?xml version="1.0" encoding="UTF-8"?>
<aliquot>
      <data>
            <chr>chr17</chr>
            <start>62503072</start>
            <stop>62503072</stop>
            <strand>+</strand>
            <composite_element_ref>cg00003784</composite_element_ref>
            <beta_value>0.0286291327274318</beta_value>
            <gene_symbol>CEP95</gene_symbol>
      </data>
      <data>
            <chr>chr19</chr>
            <start>17336525</start>
            <stop>17336525</stop>
            <strand>+</strand>
            <composite_element_ref>cg00003818</composite_element_ref>
            <beta_value>null</beta_value>
            <gene_symbol>OCEL1</gene_symbol>
```

```
        </data>
        …
</aliquot>
```

The structure of the meta data files is the same as for the BED format.

JSON format

The standard JavaScript Object Notation (JSON) format specifies the content of the genomic data files as equal to the one of the BED format, but the file structure is designed according to the JSON style. In particular, we define one genomic data JSON file for each aliquot and experiment type; the content of this file starts with the root tag called "aliquot".

Then, for each genomic measure (row of the input data file) we define a "data" tag containing the measured attributes and their values as sub-tags.

In the following, we provide an example of JSON file of DNA methylation:

```
{
        "aliquot": {
                "data": [
                        {
                                "chr": "chr17",
                                "start": "62503072",
                                "stop": "62503072",
                                "strand": "+",
                                "composite_element_ref": "cg00003784",
                                "beta_value": "0.0286291327274318",
                                "gene_symbol": "CEP95"
                        },
                        {
                                "chr": "chr19",
                                "start": "17336525",
                                "stop": "17336525",
                                "strand": "+",
                                "composite_element_ref": "cg00003818",
                                "beta_value": "null",
                                "gene_symbol": "OCEL1"
                        },
….
}
```

The structure of the meta data files is the same as for the BED format.

GTF format

The bioinformatics standard Gene Transfer Format (GTF) specifies the content of the genomic data files as equal to the one of the BED format, but the file structure is designed according to the GTF style. In particular, we define one genomic data GTF file for each aliquot and experiment type.

The nine tab-separated GTF fields are[15]:

1. **seqname** - the name of the sequence; it must be a chromosome or scaffold (in our case, the chromosome).
2. **source** - the program that generated this feature (in our case, OPENGDC)
3. **feature** - the name of this type of feature; some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon" (in our case, “GDC_Region”).
4. **start** - the starting position of the feature in the sequence; the first base is numbered 1.
5. **end** - the ending position of the feature in the sequence (inclusive).
6. **score** - a score between 0 and 1000. In UCSC Genome Browser, if the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value determines the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, "." is entered.
7. **strand** - valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - if the feature is a coding exon, *frame* should be a number between 0 and 2 that represents the reading frame of the first base; if the feature is not a coding exon, the value should be '.'.
9. **group** - a list of attributes; each attribute consists of a name-value pair (in our case, we include the fields of the genomic data file and their values, e.g., composite_element_ref "cg00003784"; beta_value "0.0286291327274318"; gene_symbol "CEP95"). Attributes must end with a semi-colon and be separated from any following attribute by exactly one space.

In the following, we provide an example of GTF file of DNA methylation:

```
chr17 OPENGDC GDC_Region  62503072  62503072  . + . composite_element_ref "cg00003784"; beta_value "0.0286291327274318"; gene_symbol "CEP95";
chr19 OPENGDC GDC_Region  17336525  17336525  . + . composite_element_ref "cg00003818"; beta_value "null"; gene_symbol "OCEL1";
chr1  OPENGDC GDC_Region  45080600  45080600  . + . composite_element_ref "cg00003858"; beta_value "null"; gene_symbol "RNF220";
chr3  OPENGDC GDC_Region  108476878 108476878 . - . composite_element_ref "cg00003965"; beta_value "null"; gene_symbol "RETNLB";
chr7  OPENGDC GDC_Region  15725862  15725862  . - . composite_element_ref "cg00003994"; beta_value "0.0493941711402823"; gene_symbol "MEOX2";
chr16 OPENGDC GDC_Region  66586745  66586745  . + . composite_element_ref "cg00004055"; beta_value "0.073911219948775"; gene_symbol "CKLF";
chr3  OPENGDC GDC_Region  36981714  36981714  . - . composite_element_ref "cg00004067"; beta_value "0.965022265629378"; gene_symbol "TRANK1";
chr19 OPENGDC GDC_Region  39898015  39898015  . + . composite_element_ref "cg00004072"; beta_value "0.0999956612897953"; gene_symbol "ZFP36";
chr15 OPENGDC GDC_Region  23034447  23034447  . - . composite_element_ref "cg00000622"; beta_value "0.0143491154061897"; gene_symbol "NIPA2";
chr2  OPENGDC GDC_Region  237027592 237027592 . + . composite_element_ref "cg00004073"; beta_value "null"; gene_symbol "AGAP1";
chr9  OPENGDC GDC_Region  139997924 139997924 . + . composite_element_ref "cg00000658"; beta_value "0.837545212449724"; gene_symbol "MAN1B1";
chr19 OPENGDC GDC_Region  54695678  54695678  . + . composite_element_ref "cg00000714"; beta_value "0.164030705433507"; gene_symbol "TSEN34";
chr6  OPENGDC GDC_Region  25282779  25282779  . + . composite_element_ref "cg00000721"; beta_value "0.956370606771304"; gene_symbol "LRRC16A";
chr3  OPENGDC GDC_Region  128902377 128902377 . - . composite_element_ref "cg00000734"; beta_value "0.0626386186322679"; gene_symbol "CNBP";
chr12 OPENGDC GDC_Region  124086477 124086477 . + . composite_element_ref "cg00000769"; beta_value "0.0233990802366794"; gene_symbol "DDX55";
```

The structure of the meta data files is the same as for the BED format.

[15] https://genome.ucsc.edu/FAQ/FAQformat#format4