|  |
| --- |
| **OpenGDC** |
| **User Guide** |
| Eleonora Cappelli, Fabio Cumbo, Anna Bernasconi, Arif Canakoglu, Stefano Ceri,  Marco Masseroli, Emanuel Weitschek |

****

**Contents**

Introduction 2

OpenGDC procedure steps 2

Executing OpenGDC 2

Download of genomic and clinical data 3

Conversion into the the BED, GTF, CSV, and JSON formats 4

Data repository 5

Citation 5

Contacts 5

Appendix: tumor tags and tumor names 6

# Introduction

This user guide is intended for all the users that want to learn how to use the OpenGDC tool for downloading and converting all the public available genomic and clinical data from the Genomic Data Commons (GDC) portal into the BED, CSV, GTF, JSON, and XML formats. Please refer also to the OpenGDC\_readme.txt file (that is included in the software package) for additional details.

# OpenGDC procedure steps

The following steps are necessary to perform the download of public GDC data and their conversion into the BED, CSV, GTF, JSON, and XML formats. These steps are thoroughly explained in the following sections of this tutorial:

* Download of genomic and clinical data;
* Conversion into the BED, CSV, GTF, JSON, and XML formats.

Installation

JAVA

The OpenGDC tool requires a working JAVA Virtual Machine (VM) installed. Thus, if not done yet, first download and install the free Oracle JAVA Runtime Environment from <http://www.java.com/getjava/>.

Several versions for the most common operating systems are available (e.g., Windows x86 for Windows 32 bit, Windows x64 for Windows 64 bit, MacOsX, or Linux). Please choose the right version according to your operating system.

OpenGDC

You can download and unzip the multi-platform (Windows, Linux and MacOS) Java software OpenGDC from [http://bioinf.iasi.cnr.it/opengdc/](http://bioinf.iasi.cnr.it/tcga2bed/) (“OpenGDC-v1.0.zip”), that allows to retrieve the GDC data and convert them into the BED, CSV, GTF, JSON, and XML formats.

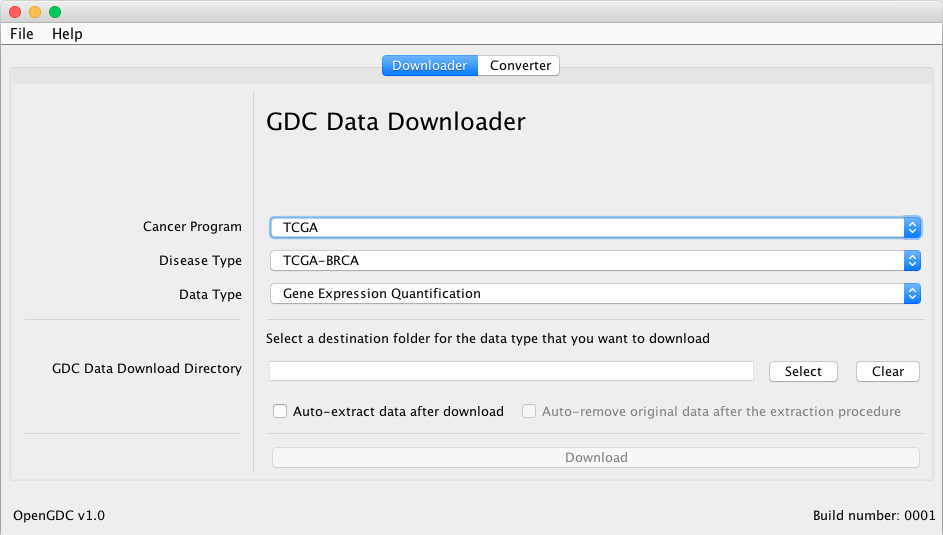
# Executing OpenGDC

Go to the directory where you extracted the OpenGDC archive and execute OpenGDC.jar by double clicking it (for supported operating systems) or by executing the following command line from a terminal window: java -jar OpenGDC.jar

The software is composed of two main panels plus one to take trace of the software log. The first panel is the Downloader, which permits the retrieval of GDC data. The other one is the Converter, which allows the conversion of the downloaded data into the BED, CSV, GTF, JSON, and XML formats.

## Download of genomic and clinical data

The first step to perform is downloading genomic and/or clinical meta data (clinical and biospecimen biotab files) for the cancer type you want to analyze. Please select the *Cancer Program* (e.g., TCGA) and the tumor tag from the drop down menu (*Disease Type*); a list with the available tumor tags and names is provided at the end of this tutorial. Additionally, select the experiment type from the *Data Type* dropdown menu. The available experiment types are: Copy Number Segment, Masekd Copy Number Segment, Gene-, Isoform-, miRNA- Expression Quantification, Methylation Beta Value, and Masked Somatic Mutation, in addition to the Clinical and Biospecimen Supplements (called meta data in this document). Then choose a folder in which the selected data set will be downloaded and press the *Download* button to start the process.



The download will start and you can take trace of the progress from the OpenGDC Log window that will appear automatically.

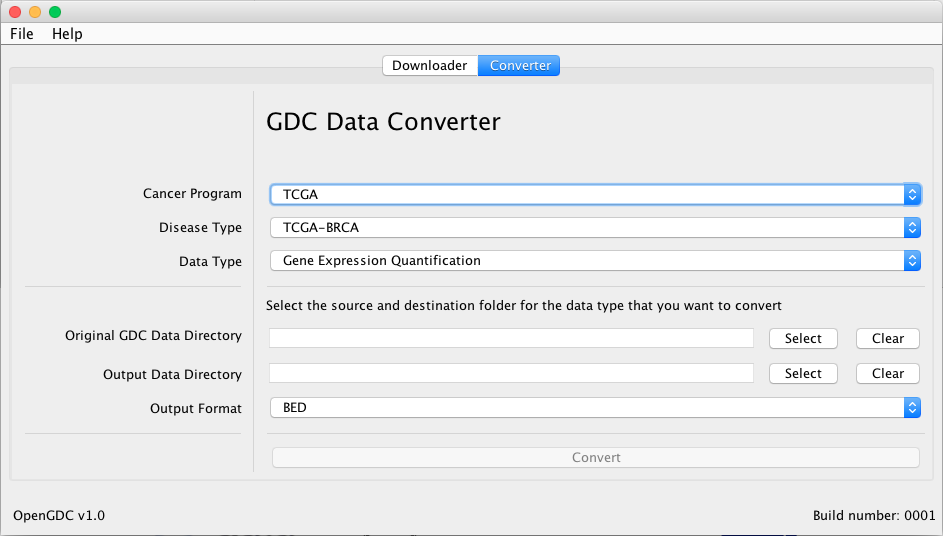
# Conversion into the the BED, GTF, CSV, and JSON formats

After the download of meta data and experimental data, you can start the conversion into the BED, CSV, GTF, and JSON formats (see the “OpenGDC\_format\_definition.pdf” format definition file that is available as Supplemental material and at <http://bioinf.iasi.cnr.it/opengdc> for further details).

Please select from the drop down menus the *Cancer Program*, the *Disease Type* (through the tumor tag) and the *Data Type* (Copy Number Segment, Masekd Copy Number Segment, Gene-, Isoform-, miRNA- Expression Quantification, Methylation Beta Value, Masked Somatic Mutation, and Clinical and Biospecimen Supplements) you want to convert.

Additionally, specify the folder where you downloaded the meta data or the experimental data (*Original GDC Data Directory*).

Finally, choose the output folder in which the converted data will be stored in (*Output Data Directory*) and the format (*Output Format*) with which you want to standardize your data among BED, GTF, CSV, JSON, and XML, before clicking the *Convert* button to start the process.



The conversion will start and you can take trace of the progress from the OpenGDC Log window that will appear automatically.

You can start the whole process again with new tumor or experiment types.

# Data repository

The ftp site <ftp://bioinformatics.iasi.cnr.it/opengdc/> contains an up-to-date archive with the experimental and meta data from GDC converted into the BED format.

# Citation

If you use OpenGDC for you research, please cite: “*OpenGDC: promoting data standardization in the Life Sciences. A concrete effort toward the harmonization of clinical and genomic cancer data*” *by Eleonora Cappelli, Fabio Cumbo, Anna Bernasconi, Arif Canakoglu, Stefano Ceri, Marco Masseroli, and Emanuel Weitschek*.

# Contacts

For comments and questions please contact Fabio Cumbo ([fabio.cumbo@iasi.cnr.it](mailto:fabio.cumbo@iasi.cnr.it)) or   
Eleonora Cappelli ([eleonora.cappelli@uniroma3.it](mailto:eleonora.cappelli@uniroma3.it)).

# Appendix: tumor tags and tumor names

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