**TCGA-Expedition User Guide**

**1. Overview**

TCGA-Expedition provides the research community access to the TCGA public and controlled data. The software generates versioned, participant and sample-centered locally stored data and metadata. It supports flexible searches of the data via a web portal.

**2. Access Requirements**

For accessing controlled data users will need to have [dbGAP](http://www.ncbi.nlm.nih.gov/gap) Data Use Certificate.

**3. System and Software Requirements**

Unix 64-bit

[Java](https://java.com/en/download/) 1.7+

[PostgreSQL](http://www.postgresql.org/download/) **OR** [Virtuoso](http://virtuoso.openlinksw.com/download/)

[Drupal](https://www.drupal.org/download)

**4. Storage Structure**

Basic Structure

Almost all TCGA data are stored in the following structure.

- <top directory> [*repository]*

- <disease abbreviation> [*brca*]

- <patient barcode> [*TCGA-BH-A0AY*]

- clinical

- bio

- <sample barcode> [*TCGA-BH-A0AY-01*]

- <analysis type> [*protected-mutations*]

- <**analysis platform>** [*IlluminaGA-DNASeq-Cont*]

NOTES:

* label in <> would be replaced by real name
* label in [] - example name
* **bio** folder contains all bio .txt files
* **analysis platform** folder contains all the files for a particular analysis without splitting them to the data levels

**Exceptions**:

- Expression\_Protein level 0 data which belongs to the whole analysis sets are stored under <disease>

- Data from the sources other than TCGA and cgHub are stored in <top directory>/other folder.

**5. File Naming Convention**

Data file name consists of 9 parts plus original file extension:

<TCGA barcode>\_<PUBL/CTRL><disease abbreviation>\_<analysis type>\_<platform>\_<TCGA level>\_<reference genome>\_<note><note number>\_<revision>.<original extension>

For example:

TCGA-A1-A0SN\_PUBL\_brca\_clin\_patient\_Level-2\_\_1\_V1

TCGA-A1-A0SN-01A-11D-A145-05\_PUBL\_brca\_jhu-usc.edu\_humanmethylation450\_Level-1\_hg19\_grn1\_V1

NOTES:

* <TCGA barcode> in the name includes as many fields as could be found for this analysis, up to full aliquot (*https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode*)
* <revision> = "v"+number - is purely TCGA-Expedition revision version
* <revision> would be updated if and only if the original TCGA file is modified
* <note> - used only if need to differentiate multiple files from the same analysis (*'red' or 'grn' note for DNA\_Methylation*)
* <note number> - default is "1"

**6. How to Run**

Usage: java -jar tcgaExpedition-<vx.x.x>.jar --diseaseList <list> --analysistype <string> --accesstype <string>

================================================================================

--diseaseList Comma separated list of disease abbreviations. Use ALL for the whole data set.

--analysistype See table below for available analysis types based on the data source.

--accesstype Use public or controlled. Access type depends on analysis type ans data level. See table below.

**Example**

Download Clinical data for acc:

java -jar tcgaExpedition-<vx.x.x>.jar acc clinical public

**7. Available Analysis / Access Types**

|  |  |  |  |
| --- | --- | --- | --- |
| **DataSource** | **AnalysisType** | **AccessType** | **Level** |
| TCGA | clinical | public | 2 |
| TCGA | cnv\_(cn\_array) | public | 1,2,3 |
| TCGA | cnv\_(snp\_array) | controlled | 1,2 |
| TCGA | cnv\_(snp\_array) | public | 3 |
| TCGA | cnv\_(low\_pass\_dnaseq) | controlled | 2 |
| TCGA | cnv\_(low\_pass\_dnaseq) | public | 3 |
| TCGA | dna\_methylation | public | 1,2,3 |
| TCGA | expression\_gene | public | 1,2,3 |
| TCGA | expression\_protein | public | 0,1,2,3 |
| TCGA | fragment\_analysis | controlled | 1 |
| TCGA | images | public | 1 |
| TCGA | mirnaseq | public | 3 |
| TCGA | protected\_mutations | controlled | 2 |
| TCGA | protected\_mutations\_maf | public | 2 |
| TCGA | rnaseq | controlled | 2 |
| TCGA | rnaseq | public | 3 |
| TCGA | rnaseqv2 | public | 3 |
| TCGA | somatic\_mutations | public | 2 |
| Firehose | CN\_Level4 | controlled | 4 |
| Georgetown | mass\_spectrometry | public | 4 |
| cgHub\* | bisulfite-seq\_(cghub) | controlled | 1 |
| cgHub\* | mirna-seq\_(cghub) | controlled | 1 |
| cgHub\* | rna-seq\_(cghub) | controlled | 1 |
| cgHub\* | validation\_(cghub) | controlled | 1 |
| cgHub\* | wgs\_(cghub) | controlled | 1 |
| cgHub\* | wxs\_(cghub) | controlled | 1 |

* - coming soon