**TCGAExpedition User Guide**

**1. Overview**

TCGAExpedition 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.

TCGAExpedition Modules:

* Data uploader: [tcga-expedition](https://github.com/TCGAExpedition/tcga-expedition) – performs TCGA data download, creates and stores metadata about the downloaded files.
* Web portal (optional): [ipm](https://github.com/TCGAExpedition/ipm) plus [ipm-ws](https://github.com/TCGAExpedition/ipm-ws) – serve for identifying files of interest based on metadata query and filtering, and to download a manifest of files and directory locations. It also provides notification mechanisms based on [tcga-expedition](https://github.com/TCGAExpedition/tcga-expedition) RDF graph or relational metadata store.

See TCGAExpedition installation and demo [video](https://youtu.be/bpcQiBNf8Fc).

**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. Installation**

**4.1 Module tcga-expedition**

#### Select and install one of the supported storages:

* PostgreSLQ*:*[*http://www.postgresql.org/download/*](http://www.postgresql.org/download/)
* Virtuoso*:*[*http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSDownload*](http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSDownload)

**NOTE:** We recommend to use PostgreSQL - it's much faster than RDF store.

#### Configure

Set parameters in resources/tcgaexpedition.conf file.

Required:

* TCGA credentials. Leave blank if download public data only.

**tcga.user**

**tcga.pwd**

* Storage selection. Uncomment 'postgres' or 'virtuoso'.

**storage.name**=postgres

#**storage.name**=virtuoso

* Set access parameters for PostgreSQL or Virtuoso.
* Set local repository location.

**repository.home**

Optional:

* Email used to send updates to the web portal subscribers.
* Sender and receiver emails for notification about ambiguous tissue source site names.

**4.2 Web Portal Module**

* **Download and install** [**ipm**](https://github.com/TCGAExpedition/ipm)  - a Drupal prerequisite. Follow the installation instructions on <https://www.drupal.org/start>.
* **Install** [**ipm-ws**](https://github.com/TCGAExpedition/ipm-ws)
* **Configure** [**ipm-ws**](https://github.com/TCGAExpedition/ipm-ws)
* Put war/ipm-ws.war on your Tomcat web server.
* Set ***useVitOrPostgre*** parameter in resources/jQueryPostgres.conf.
* Set connection parameters in *jQueryPostgres.conf* or in *jQueryVrt.conf* based on your  data [tcga-expedition](https://github.com/TCGAExpedition/tcga-expedition) store selection.
* Copy *jQueryPostgres.conf* to $CATALINA\_HOME/conf directory.
* If you use Virtuoso, copy *jQueryVrt.conf* file must be in $CATALINA\_HOME/conf directory as well.
* Start the Tomcat server.

**5. TCGAExpedition repository**

**5.1 Repository Structure**

Repository contains patient-centric TCGA data.

Basic Structure (Fig. 1)

Almost all TCGA data are stored in the following structure.

- <top directory> [*repository]*

- <disease abbreviation> [*brca*]

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

- clinical

- bio

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

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

- <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
* <analysis type> directory under the <patient barcode> contains multi-sample analysis files. In this case <analysis type> under <sample barcode> has been created but is empty.

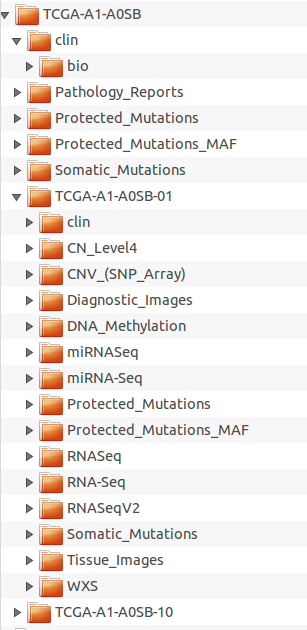


Figure 1. Repository structure

**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.

**5.2 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 TCGAExpedition 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"

**5.3. 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 brca clinical public

**5.4. 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

**6. Web Portal**

**6.1 Search Repository**

The web interface for identifying files of interest contains multiple filters (Fig. 2). It also provides a way to download the metadata information about the selected files as a table in csv or tsv format.

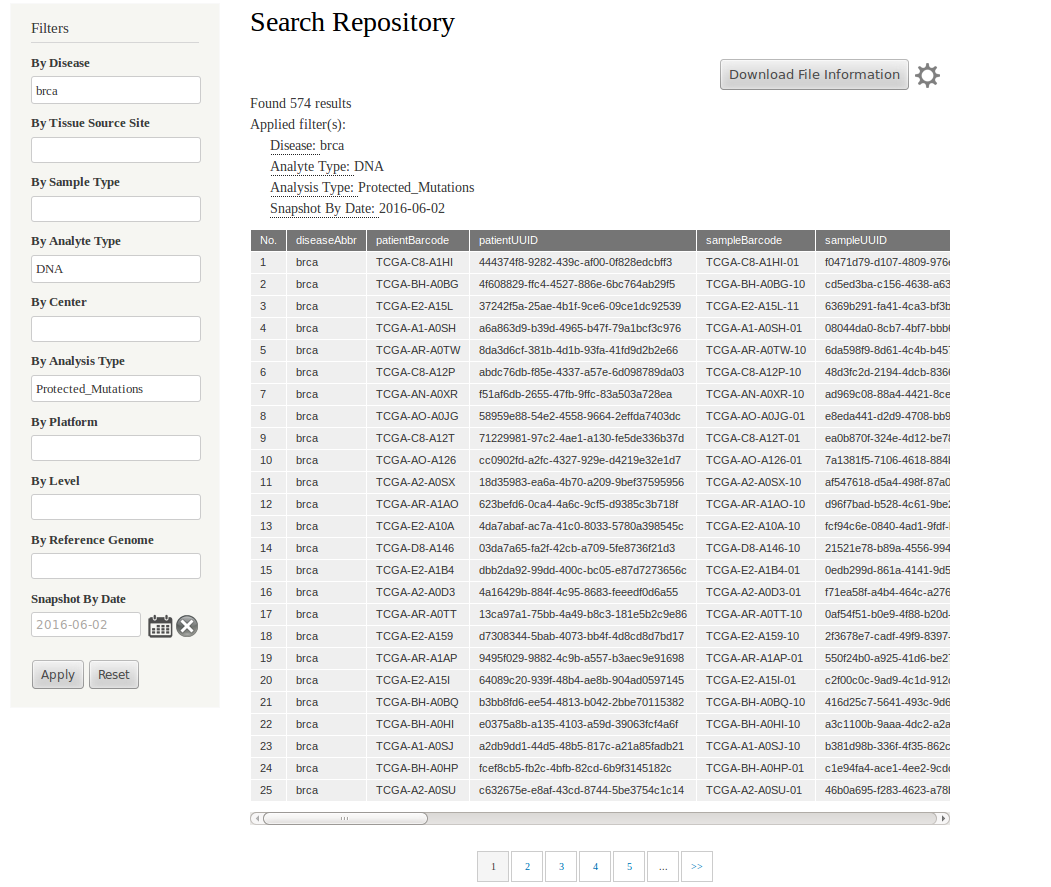


Figure 2. Search repository interface.

**6.2 Data notification request**

The investigators can request to be notified by email of selected data changes (added or updated) (Fig. 3).

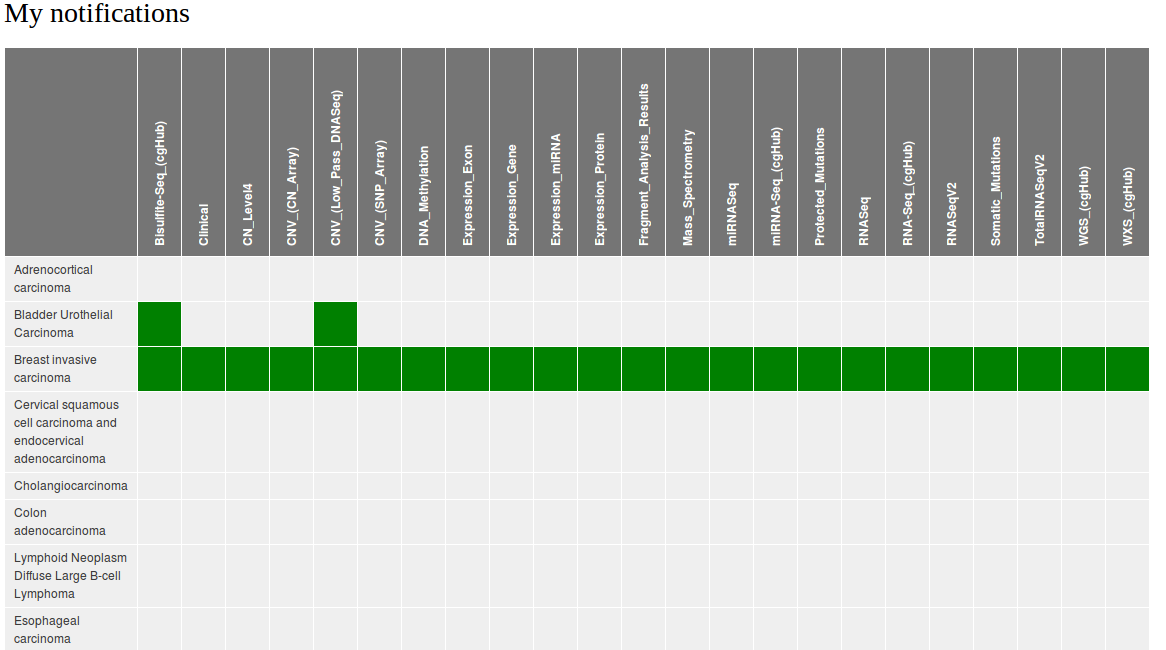


Figure 3. Notification request page.