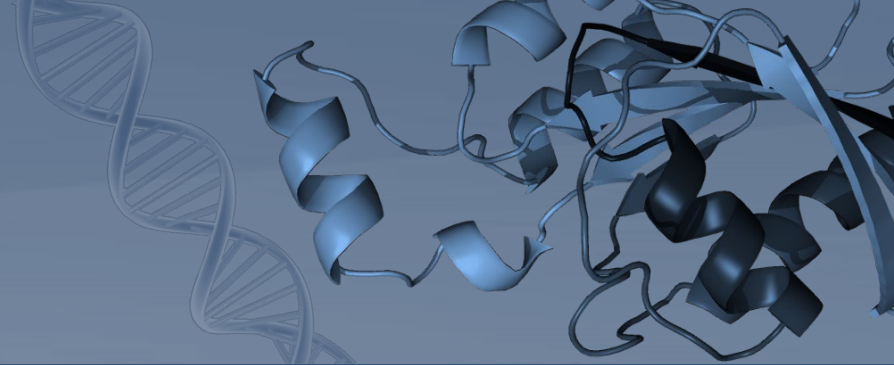




OFFICE OF CANCER CLINICAL
PROTEOMICS RESEARCH



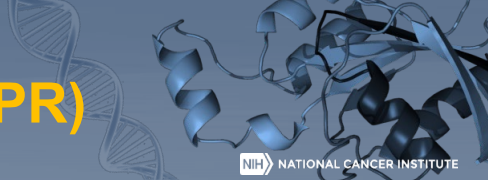
CPTAC Data at the GDC

Ana I. Robles, Ph.D.

Program Director

Office of Cancer Clinical Proteomics Research (OCCPR)
Center for Strategic Scientific Initiatives
Office of the Director, NCI

ana.robles@nih.gov



Tumor Characterization Programs

- **Clinical Proteomic Tumor Analysis Consortium (CPTAC)**
nationwide effort to apply proteogenomics to the characterization of tumors and generate public resources of proteogenomic data and targeted proteomic assays
- **Applied Proteogenomics Organizational Learning and Outcomes (APOLLO)**
multi-federal agency partnership: NCI, DoD, VA
- **International Cancer Proteogenome Consortium (ICPC)**
forum for collaboration among the world's leading cancer and proteogenomic researcher centers

Builds on TCGA

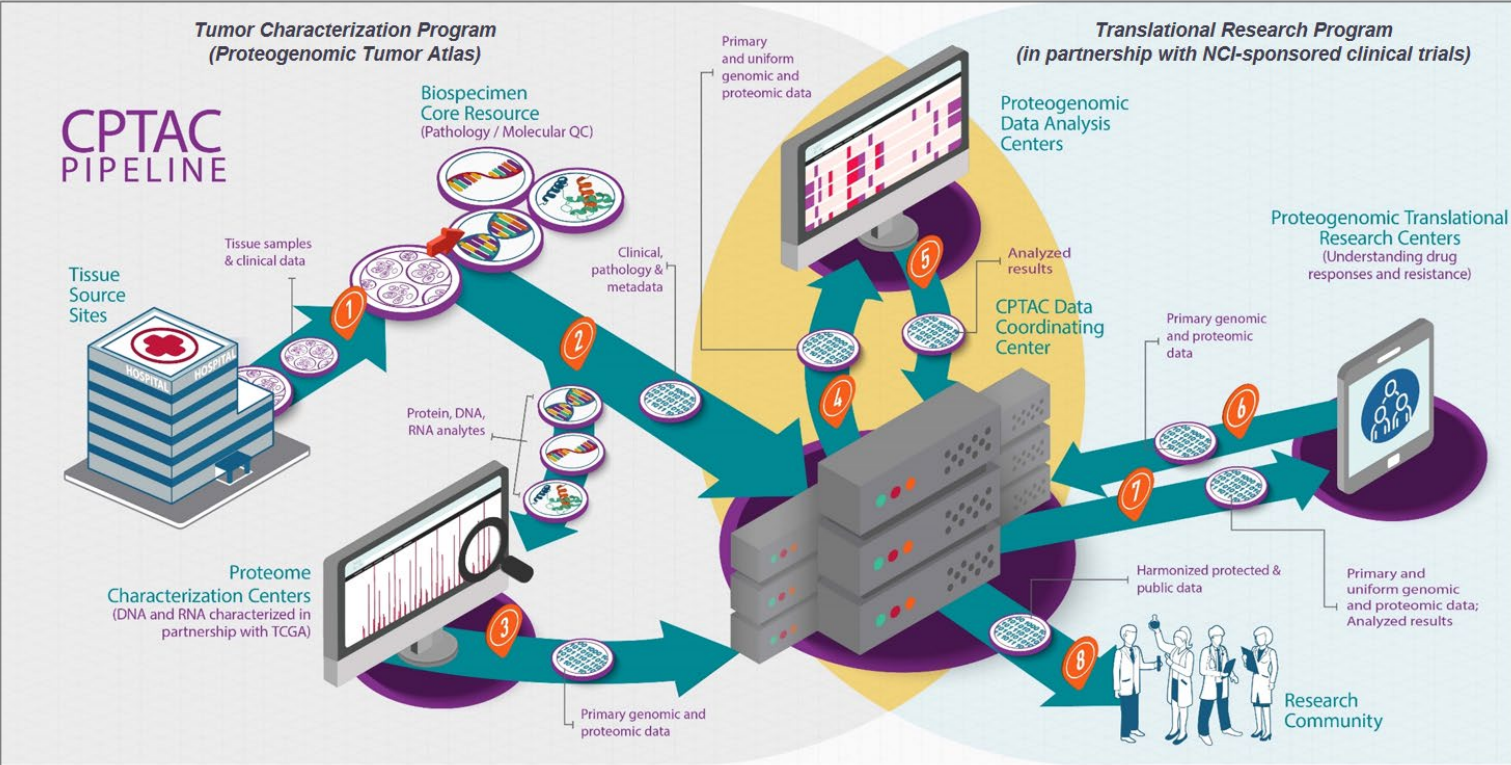
- Tumor Characterization Program (treatment naïve – tumor and NAT)
 - Characterize proteins and genes to better understand the molecular basis of cancer
- Translational Research Program (pre-clinical and clinical trials)
 - Understand [predict] drug response and resistance to therapies **in context of a clinical trial**

- Public Resources

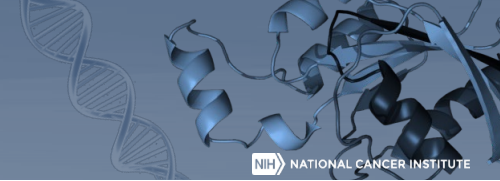


CPTAC Pipeline: Proteogenomics builds on genomics

TSSs
 ↓
 BCR
 ↓
 PCCs
 GCCs (TCGA)
 ↓
 DCC
 ↓
 PGDACs



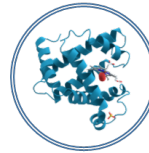
Where is the Data?



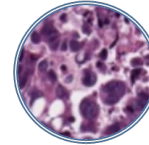
Genomic Data



Clinical Data



Proteomic Data



Imaging Data



SRA



Genomic Data Commons



Data Portal



Proteomic Data Commons

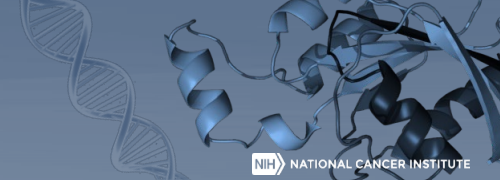


TCIA

The Cancer Imaging Archive

Genomic Data Commons

<https://portal.gdc.cancer.gov/>



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Search: e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 17.0 - June 05, 2019

PROJECTS

47

PRIMARY SITES

68

CASES

33,605

FILES

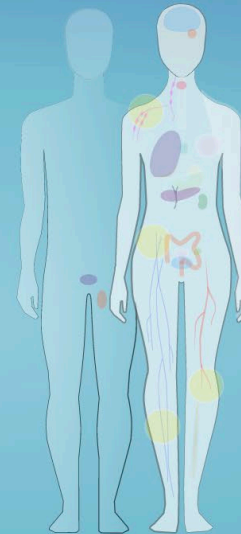
374,699

GENES

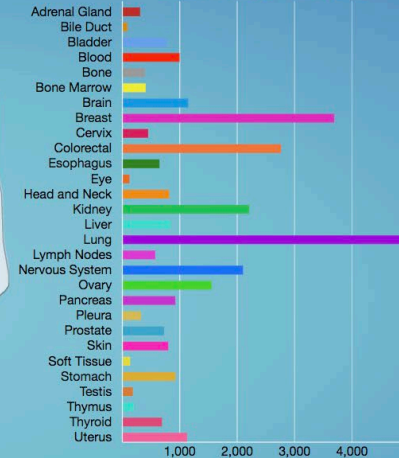
22,872

MUTATIONS

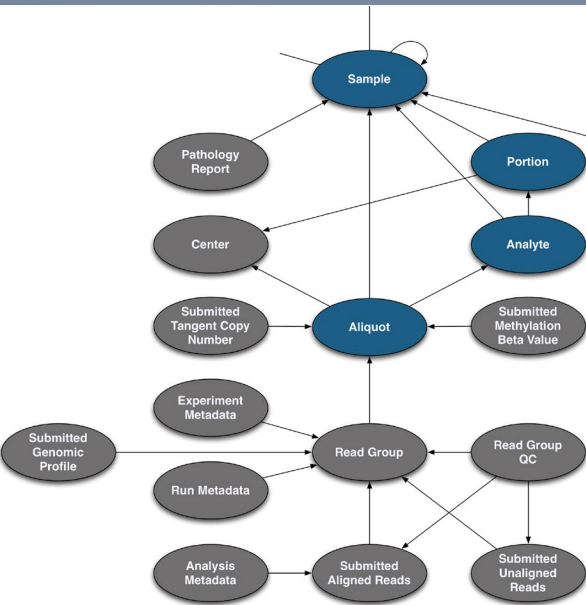
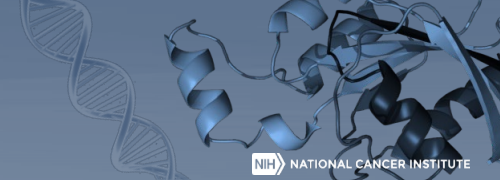
3,142,246



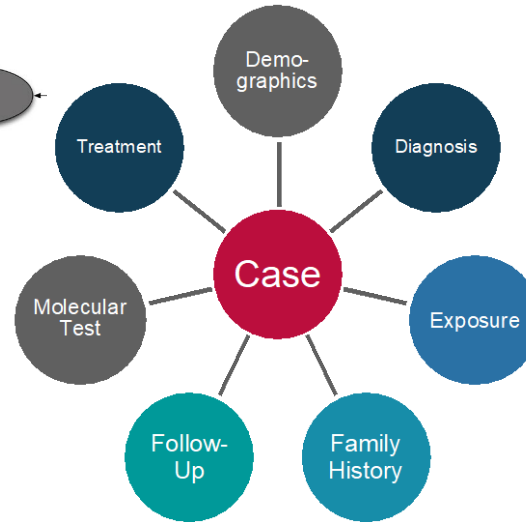
Cases by Major Primary Site



GDC Data submission metadata are as important as molecular data



Biospecimen Data



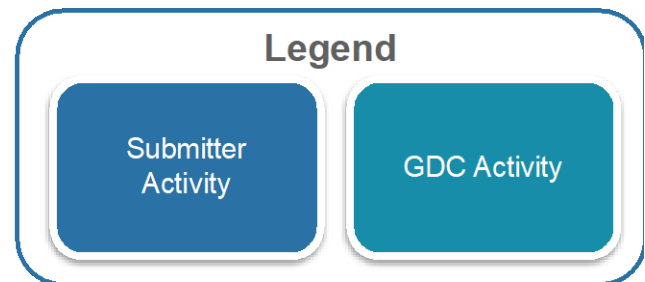
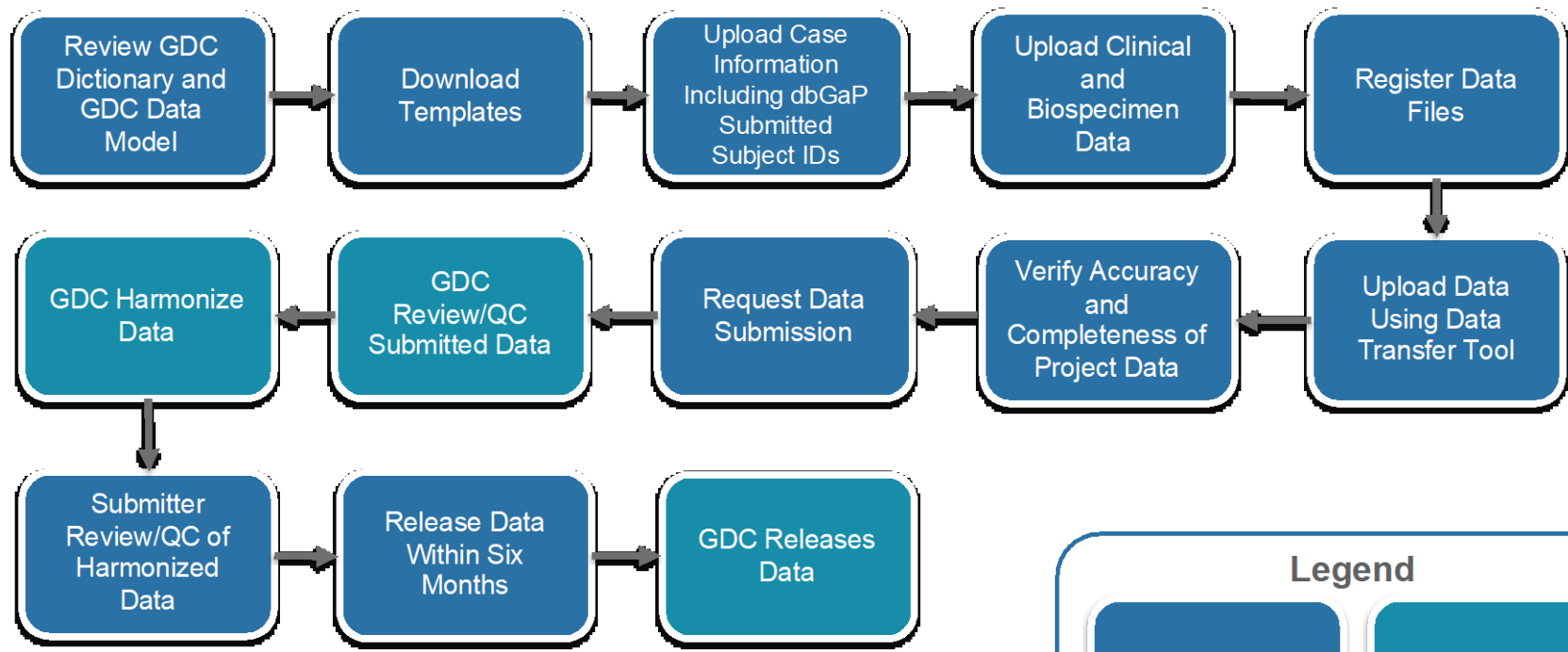
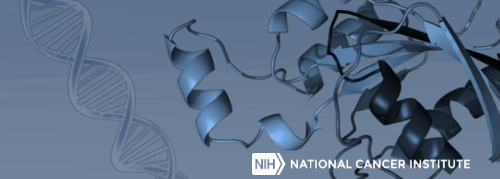
Clinical Data

Data File	Data Category	File Format
Aligned/ Unaligned Reads	Sequencing Reads (WGS, WXS, RNA-Seq)	BAM, FASTQ
Raw Methylation Array	DNA Methylation	IDAT
Slide Image	Biospecimen	SVS, JPEG, TIFF
Supplement Files	Clinical, Biopecimen	TXT, XML

Molecular Data

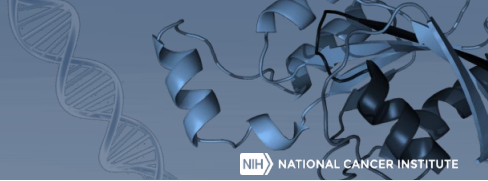
- The GDC supports the submission of biospecimen and clinical data, and experimental data files
- The [GDC Data Dictionary](#) identifies the GDC supported data elements. Relationships between elements are described in the [GDC Data Model](#).

GDC Data submission workflow



Courtesy of Zhining Wang, Ph.D., CCG

GDC Data access, exploration and analysis



Choice of three supporting infrastructures for data access:

1. Repository: Download and analyze locally
 - Fastest access if you know what you want to download
2. Exploration: Preview the data before downloading
 - If you want to take a look at the data before download
 - Explore data in the GDC using advanced filters/facets, including on specific Cases, Genes, and/or Mutations
 - The Gene/Mutation data for these visualizations come from the Open-Access MAF files on the GDC Data Portal.
3. Analysis: build your own cohorts and analyze online

Find and download data

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

- Projects
- Exploration
- Analysis
- Repository

Search: e.g. BRAF, Breast, TCGA-BLCA, TCGA-AS-ADG2

Data Portal Summary Data Release 17.0 - June 08, 2019

PROJECTS	PRIMARY SITES	CASES
47	68	33,605
FILES	GENES	MUTATIONS
374,699	22,872	3,142,246

Cases by Major Primary Site

Primary Site	Cases
Adrenal Gland	~10
Bile Duct	~10
Bladder	~10
Blood	~10
Bone	~10
Bone Marrow	~10
Brain	~10
Breast	~3,500
Cervix	~10
Colorectal	~10
Esophagus	~10
Eye	~10
Head and Neck	~10
Kidney	~10
Liver	~10
Lung	~10
Lymph Nodes	~10
Nervous System	~10
Ovary	~10
Pancreas	~10
Pituitary	~10
Prostate	~10
Skin	~10
Soft Tissue	~10
Stomach	~10
Testis	~10
Thymus	~10
Thyroid	~10
Uterus	~10

3 ways to get data

GUI

DTT

(Data Transfer Tool)

API

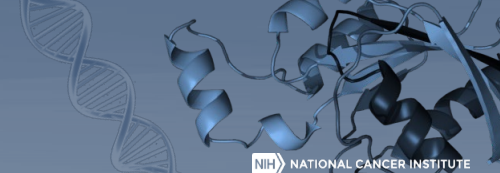
(Application Programming Interface)

API URL Endpoint URL Parameters Query Parameters

https://api.gdc.cancer.gov/projects?fields=project_id,primary_site&pretty=true

```
C:\ZhiningWang>
C:\ZhiningWang>gdc-client.exe download -t token.txt 7e0de2e1-db9d-4962-862c-33873129cb8
2 9887a43b-a514-4112-8aec-473ee9dda4d9
WARNING: Your token file 'C:\ZhiningWang\token.txt' is not properly secured. Please sec
ure your token file by ensuring that it is not readable or writable by anyone other th
an the owner of the file. Contact your system administrator for assistance.
100% [#####] Time: 0:00:00
100% [#####] Time: 0:00:00 2.65 GB/s
100% [#####] Time: 0:00:00
100% [#####] Time: 0:00:00 2.81 GB/s
Successfully downloaded: 2
```

Find CPTAC data at the GDC



Cases Genes Mutations <

[Add a Case Filter](#)

Case ?

Q e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Primary Site

- Bronchus and lung 111
- Kidney 110
- Uterus, NOS 101

Program ↻

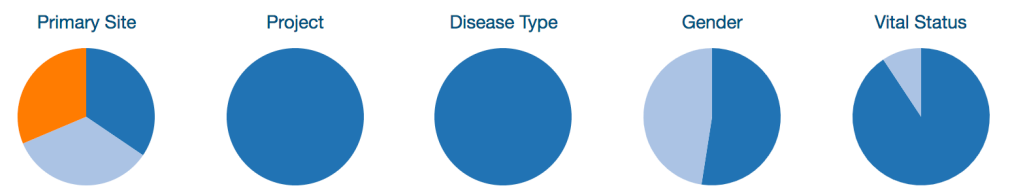
- FM 18,004
- TCGA 11,315
- TARGET 3,360
- NCICCR 489
- CPTAC 322

4 More...

Clear Program Name IS CPTAC

View Files in Repository

Cases (322) Genes (0) Mutations (0) OncoGrid

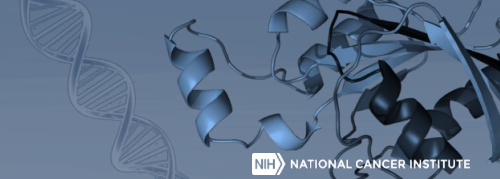


Showing 1 - 20 of 322 cases

Biospecimen Clinical JSON TSV Save/Edit Case Set

Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category						# Mutations	# Genes	
					Seq	Exp	SNV	CNV	Meth	Clinical			Bio
C3N-01071	CPTAC-3	Bronchus and lung	Male	30	12	10	8	0	0	0	0	0	0
C3N-00383	CPTAC-3	Uterus, NOS	Female	30	12	10	8	0	0	0	0	0	0
C3L-00913	CPTAC-3	Bronchus and lung	Male	30	12	10	8	0	0	0	0	0	0
C3L-00001	CPTAC-3	Bronchus and lung	Female	29	11	10	8	0	0	0	0	0	0

CPTAC metadata at the GDC



▼ Project

CPTAC-3 **322**

▼ Disease Type

Adenomas and Adenocarcinomas **322**

▼ Gender

female **169**

male **153**

▼ Age at Diagnosis

Years Days

From: eg. 30 To: eg. 90 Go!

▼ Vital Status

Alive **292**

Dead **30**

▼ Days to Death

From: eg. 23 To: eg. 732 Go!

▼ Race

not reported **156**

white **154**

black or african american **5**

asian **3**

unknown **3**

1 More...

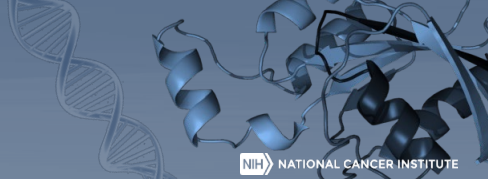
▼ Ethnicity

not reported **187**

not hispanic or latino **123**

hispanic or latino **12**

CPTAC molecular data at the GDC



▼ Data Category

- Sequencing Reads **12**
- Transcriptome Profiling **10**
- Simple Nucleotide Variation **8**

▼ Data Type

- Aligned Reads **12**
- Gene Expression Quantification **8**
- Annotated Somatic Mutation **4**
- Raw Simple Somatic Mutation **4**
- Splice Junction Quantification **2**

▼ Experimental Strategy

- RNA-Seq **16**
- WXS **11**
- WGS **3**

▼ Workflow Type

- BWA with Mark Duplicates and BQSR **6**
- STAR - Counts **4**
- HTSeq - Counts **2**
- HTSeq - FPKM **2**
- HTSeq - FPKM-UQ **2**
- STAR 2-Pass Chimeric **2**
- STAR 2-Pass Genome **2**
- STAR 2-Pass Transcriptome **2**
- MuSE **1**
- MuSE Annotation **1**
- MuTect2 **1**
- MuTect2 Annotation **1**
- SomaticSniper **1**
- SomaticSniper Annotation **1**
- VarScan2 **1**
- VarScan2 Annotation **1**

Less...

▼ Data Format

- BAM **12**
- VCF **8**
- TXT **6**
- TSV **4**

▼ Platform

- Illumina **12**

▼ Access

- controlled **22**
- open **8**

Download GDC data directly in R package



Home

Install

Help

Home » Bioconductor 3.8 » Software Packages » GenomicDataCommons

GenomicDataCommons

platforms all rank 188 / 1649 posts 1 / 1 / 1 / 0 in Bioc 2 years

build warnings updated before release

DOI: [10.18129/B9.bioc.GenomicDataCommons](https://doi.org/10.18129/B9.bioc.GenomicDataCommons)  

NIH / NCI Genomic Data Commons Access

Bioconductor version: Release (3.8)

Programmatically access the NIH / NCI Genomic Data Commons RESTful service.

Author: Martin Morgan [aut], Sean Davis [aut, cre]

TCGAbiolinks

platforms all rank 97 / 1649 posts 10 / 0.6 / 2 / 0 in Bioc 3.5 years

build ok updated < 1 week

DOI: [10.18129/B9.bioc.TCGAbiolinks](https://doi.org/10.18129/B9.bioc.TCGAbiolinks)  

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

Bioconductor version: Release (3.8)

The aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses and iv) to easily reproduce earlier research results. In more detail, the package provides multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., survival plots, volcano plots, starburst plots) in order to easily develop complete analysis pipelines.

Author: Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Davide Garolini, Claudia Cava, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr

Maintainer: Antonio Colaprico <antonio.colaprico@ulb.ac.be>, Tiago Chedraoui Silva <tiagochst@usp.br>

Citation (from within R, enter `citation("TCGAbiolinks")`):

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." *Nucleic Acids Research*. doi: [10.1093/nar/gkv1507](https://doi.org/10.1093/nar/gkv1507), <http://doi.org/10.1093/nar/gkv1507>.



Link to slide images (TCGA only)

Reset | [Add a File Filter](#)

[Clear](#) [Data Format](#) [IS](#) [SVS](#) [Advanced Search](#)

[Add All Files to Cart](#) [Manifest](#) [View 11,007 Cases in Exploration](#) [View Images](#) [Browse Annotations](#)

Files (30,072) Cases (11,007) 16.98 TB

Primary Site Project Data Category Data Type Data Format

[Show More](#)

Showing 1 - 20 of 30,072 files

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	TCGA-30-1853-01A-01-BS1.97d7d1f1-3b08-47ed-84d2-9c7439d1720c.svs	1	TCGA-OV	Biospecimen	SVS	174.72 MB	0
open	TCGA-29-1692-01B-01-TS1.2f5a7806-6d45-4669-81be-06a732632f54.svs	1	TCGA-OV	Biospecimen	SVS	28.25 MB	0
open	TCGA-13-0766-01A-02-BS2.3710a5cd-90f0-4d8a-9b23-a85b0640f834.svs	1	TCGA-OV	Biospecimen	SVS	144.16 MB	0
open	TCGA-25-1634-01A-01-BS1.d2bf3a3f-c51b-40a9-956d-2f77195d7a69.svs	1	TCGA-OV	Biospecimen	SVS	127.53 MB	0
open	TCGA-25-2042-01A-01-TS1.a7d0928a-7742-4678-8183-b9f192787900.svs	1	TCGA-OV	Biospecimen	SVS	78.84 MB	0
open	TCGA-13-0884-01B-01-TS1.4bcf3062-6cc6-4759-9515-4d1674763ca5.svs	1	TCGA-OV	Biospecimen	SVS	249.82 MB	0
open	TCGA-04-1655-01A-01-BS1.55f36120-46fb-4c88-b137-81798b011a9f.svs	1	TCGA-OV	Biospecimen	SVS	72.82 MB	0
open	TCGA-09-0365-01A-02-BS2.1d81c850-3161-4aab-b772-cfba72a6d66e.svs	1	TCGA-OV	Biospecimen	SVS	62.59 MB	0

[Menu](#) [Filter](#) [JSON](#) [TSV](#)

Mean Coverage ✕
From: eg. 0 To: eg. 0 Go!

File ?
e.g. 142682.bam, 4f6e2e7a-b...

Data Category
 Biospecimen 30,072

Data Type
 Slide Image 30,072

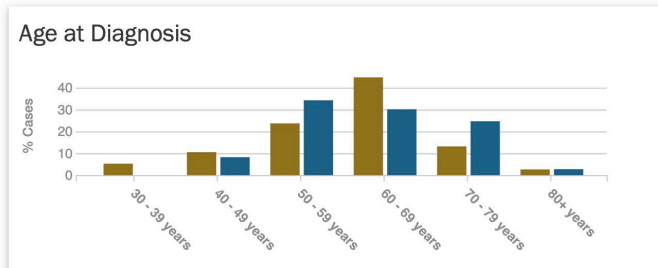
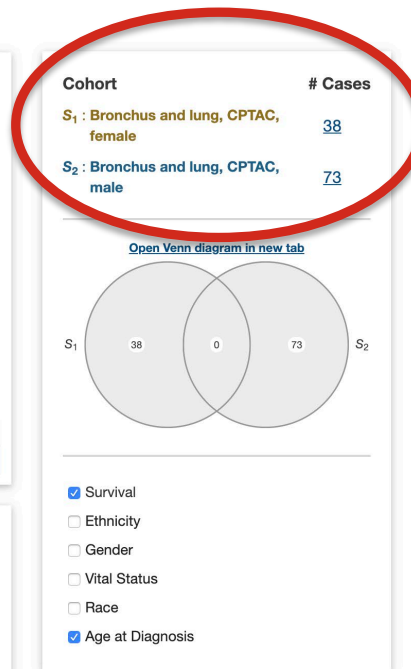
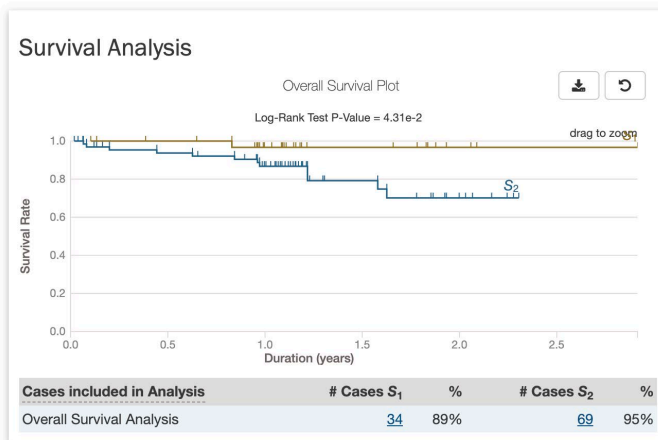
Experimental Strategy
 Tissue Slide 18,306
 Diagnostic Slide 11,766

Workflow Type
No data for this field

Data Format ↻
 VCF 131,416
 TXT 119,261
 BAM 53,580
 SVS 30,072
 BCR XML 22,496
[8 More...](#)

GDC Analysis (gender comparison CPTAC LUAD)

Cohort Comparison



THANK YOU

<https://proteomics.cancer.gov>



CPTAC Investigator Retreat
April 2019 | Bethesda, MD