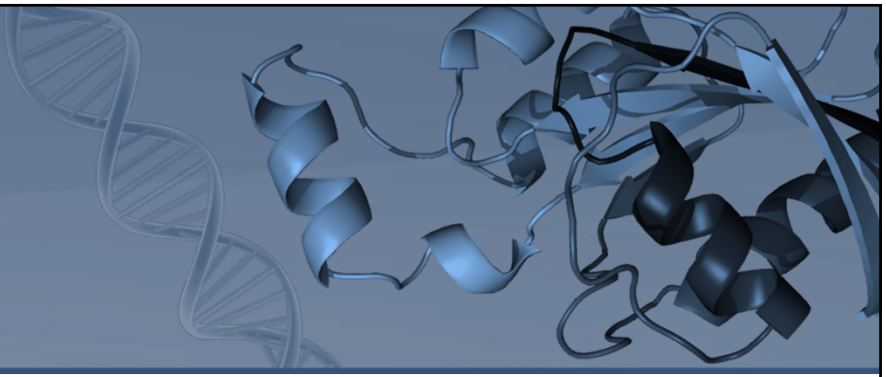


OFFICE OF CANCER CLINICAL PROTEOMICS RESEARCH

CPTAC Data at the GDC

Ana I. Robles, Ph.D.

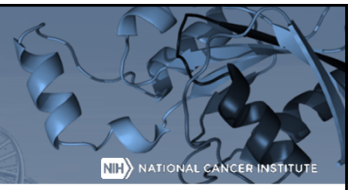
Program Director

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

ana.robles@nih.gov

Imaging IG
July 1, 2019

NIH NATIONAL CANCER INSTITUTE



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

Clinical Proteomic Tumor Analysis Consortium (CPTAC)

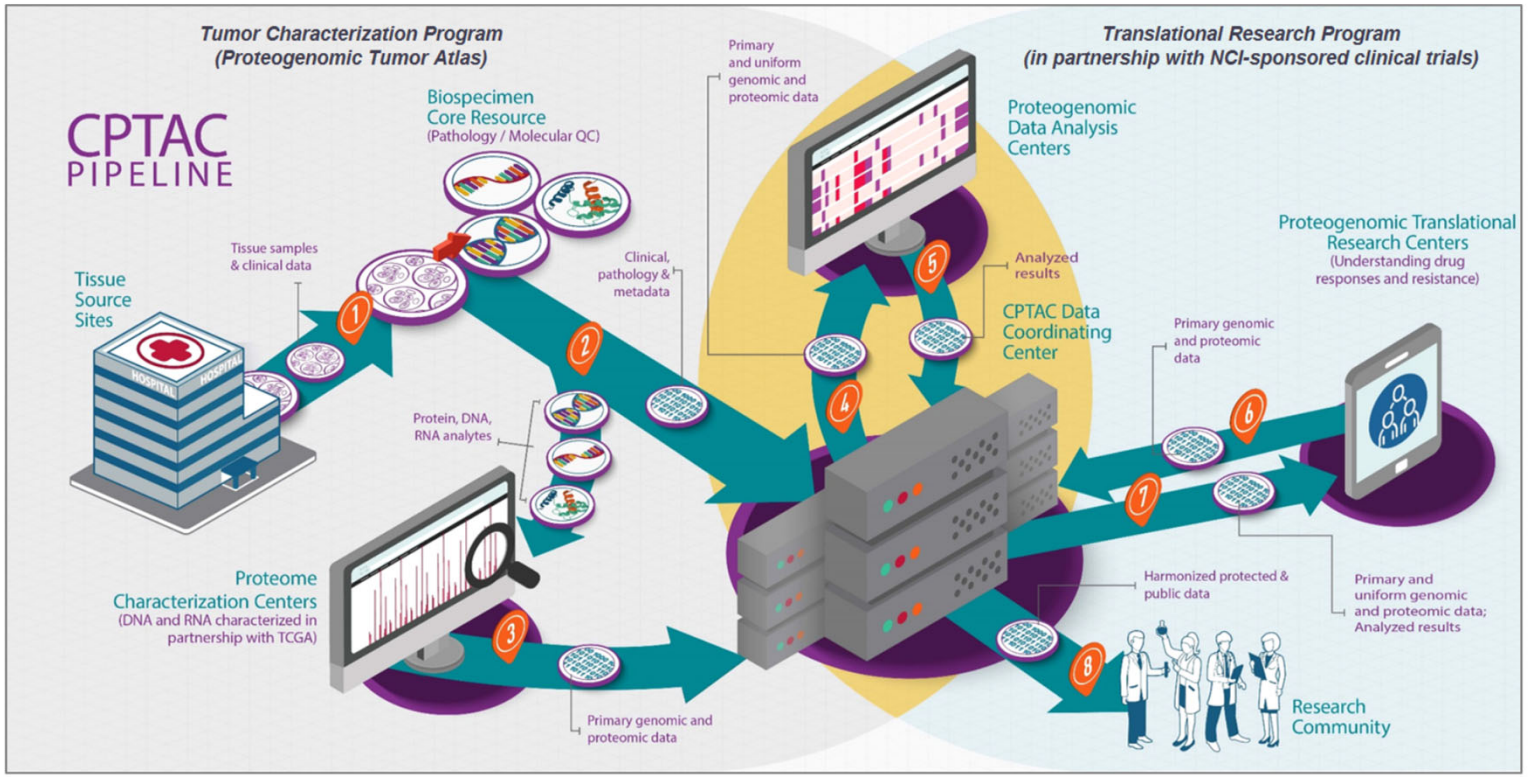
NIH NATIONAL CANCER INSTITUTE

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

NCBI

SRA

Genomic Data Commons

CPTAC

Data Portal

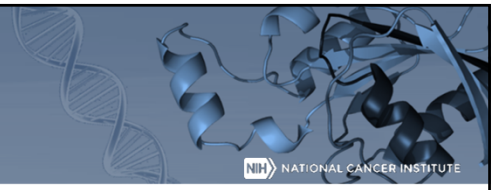
PDC

Proteomic Data Commons

TCIA

The Cancer Imaging Archive

Genomic Data Commons
https://portal.gdc.cancer.gov/



NIH NATIONAL CANCER INSTITUTE GDC Data Portal | Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

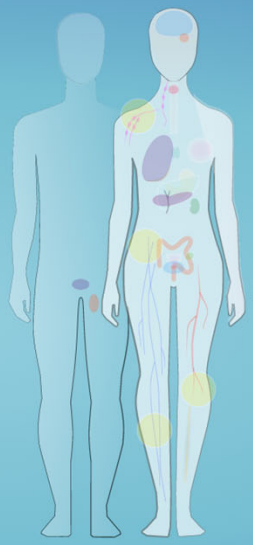
Projects | Exploration | Analysis | Repository

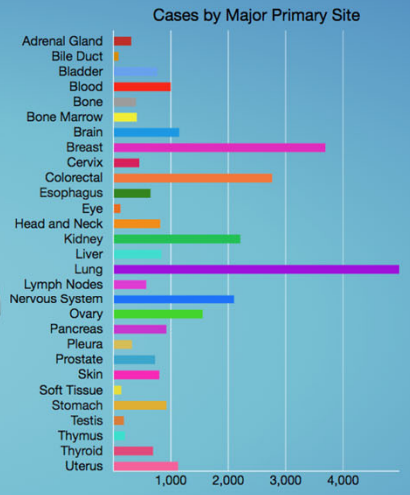
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

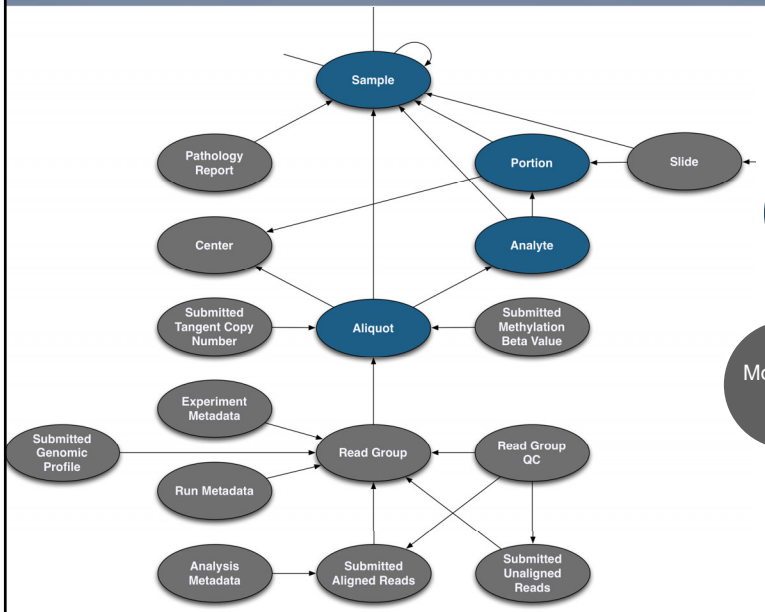
Data Release 17.0 - June 05, 2019

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

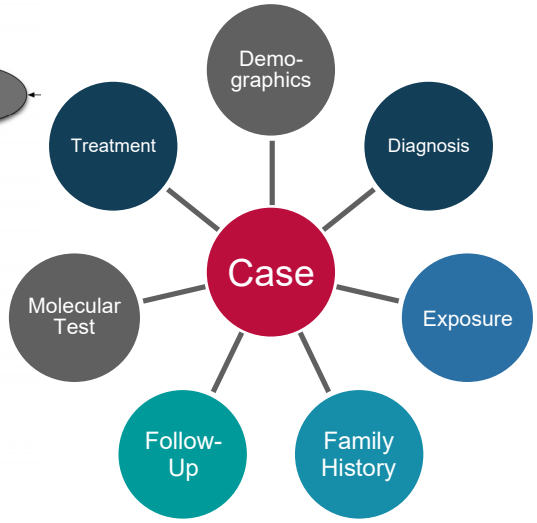




GDC Data submission metadata are as important as molecular data



Biospecimen Data



Clinical Data

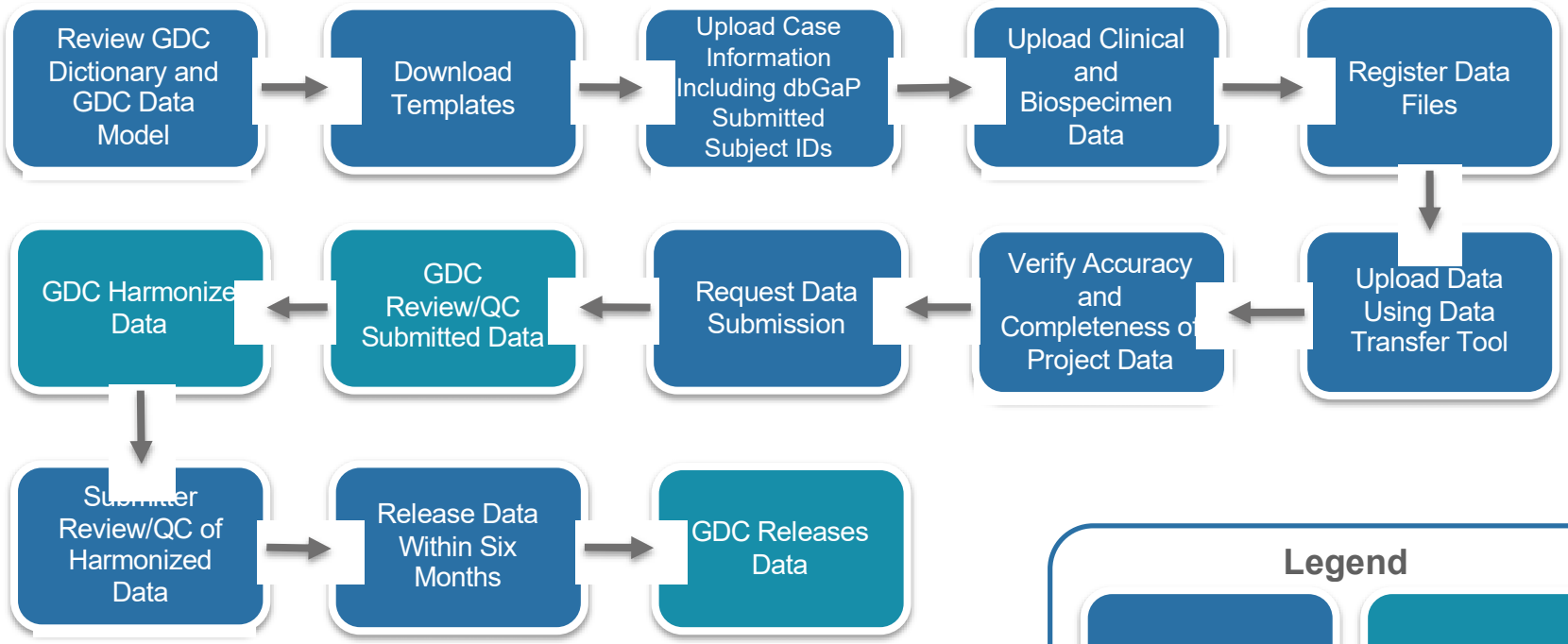
Data File	Data Category	File Format
Aligned/ Unaligned Reads	Sequencing Reads (WGS, WXS, RNA-Seq)	BAM, FASTQ
Raw Methylation Array	DNA Methylation	IDAT
Slide Image	Biospecimen	SVS, JPEG, TIFF
Supplement Files	Clinical, Biopecimen	TXT, XML

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.

Courtesy of Zhining Wang, Ph.D., CCG

GDC Data submission workflow



Courtesy of Zhining Wang, Ph.D., CCG

GDC Data access, exploration and analysis

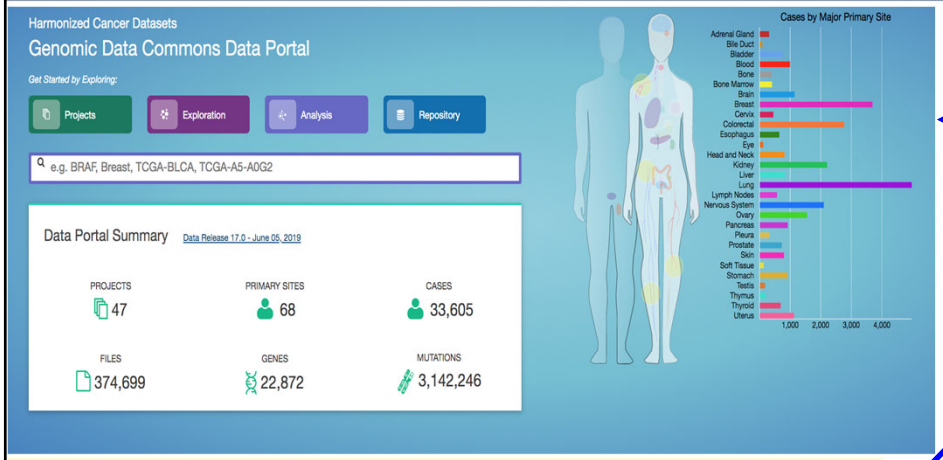
Choice of three supporting infrastructures for data access:

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



3 ways to get data

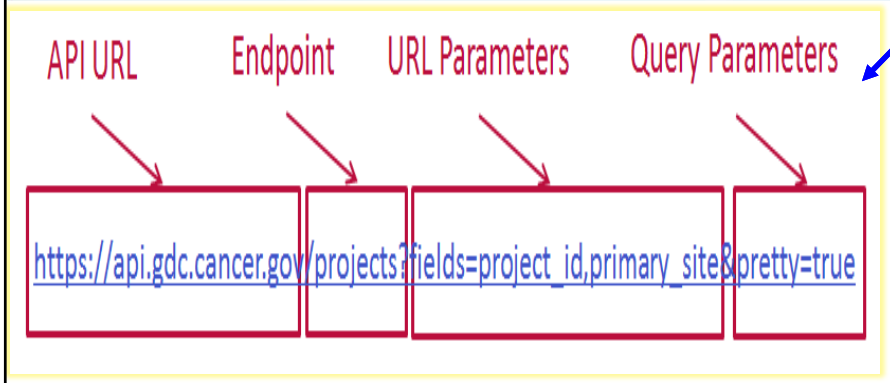
GUI

DTT

(Data Transfer Tool)

API

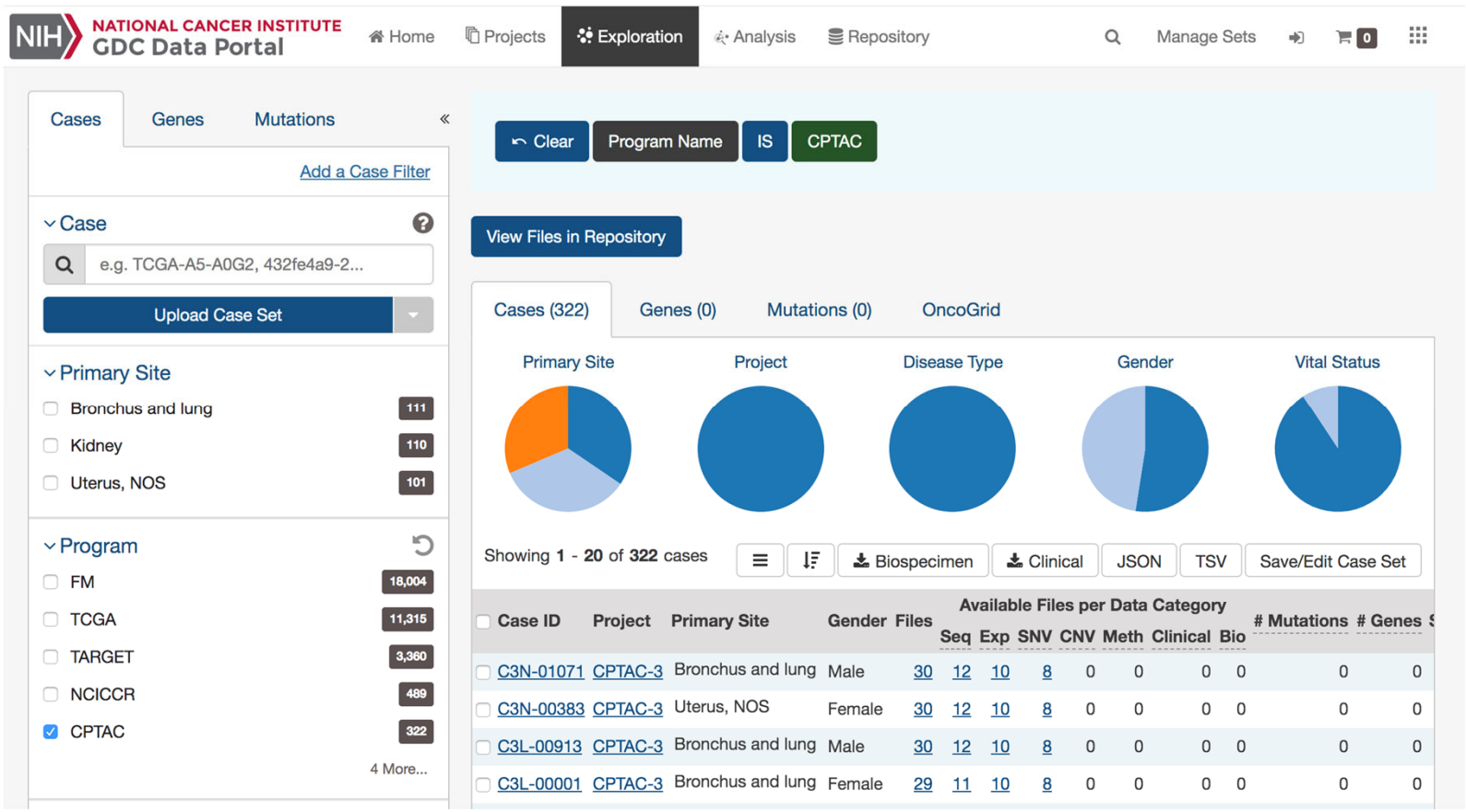
(Application Programming Interface)



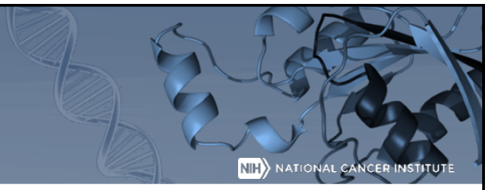

```
C:\ZhiningWang>
C:\ZhiningWang>gdc-client.exe download -t token.txt 7e0de2e1-db9d-4962-862c-33873129cb8
2 9887a43b-a514-4112-8aec-473ee9dda4d9
WARNING: Your token file 'C:\ZhiningWang\token.txt' is not properly secured. Please sec
ure your token file by ensuring that it is not readable or writeable 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
```

Courtesy of Zhining Wang, Ph.D., CCG

Find CPTAC data at the GDC



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

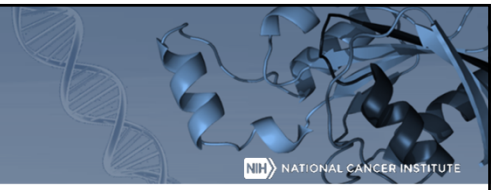
Days to Death
- From: eg. 23 To: eg. 732 Go!

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

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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

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

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

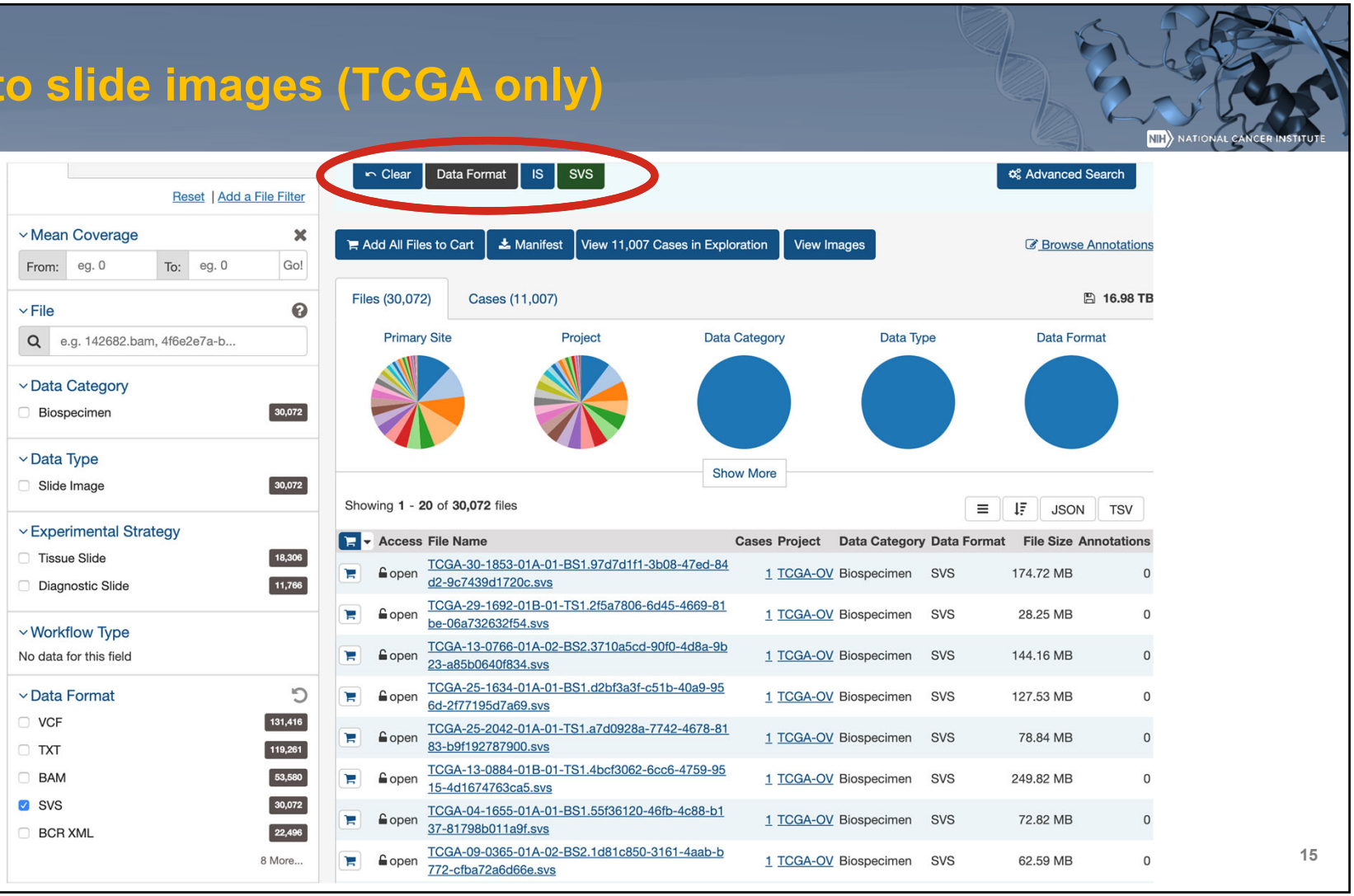
Maintainer: Antonio Colaprico <antonio.colaprico at ulb.ac.be>, Tiago Chedraoui Silva <tiagochst at usp.br>

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

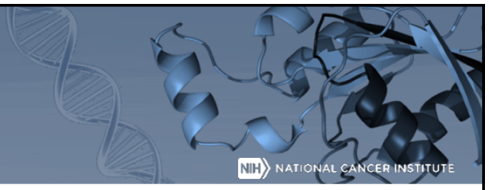
Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). " TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." *Nucleic Acids Research*. doi: 10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

Courtesy of Zhining Wang, Ph.D., CCG

Link to slide images (TCGA only)

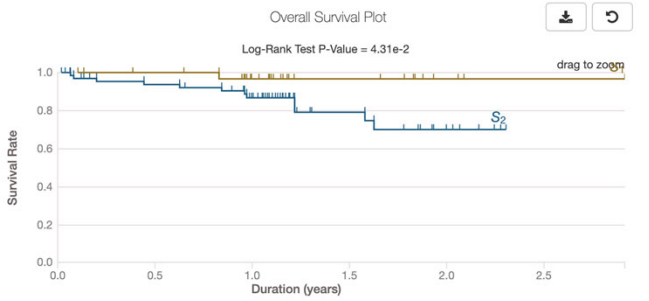


GDC Analysis
(gender comparison CPTAC LUAD)



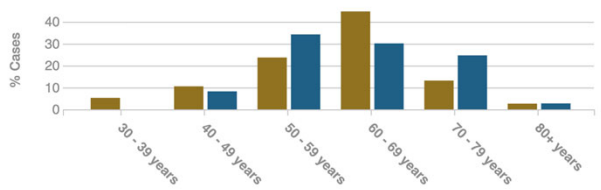
Cohort Comparison

Survival Analysis



Cases included in Analysis	# Cases S_1	%	# Cases S_2	%
Overall Survival Analysis	34	89%	69	95%

Age at Diagnosis



Cohort	# Cases
S_1 : Bronchus and lung, CPTAC, female	38
S_2 : Bronchus and lung, CPTAC, male	73

Open Venn diagram in new tab

- [x] Survival
- [ ] Ethnicity
- [ ] Gender
- [ ] Vital Status
- [ ] Race
- [x] Age at Diagnosis

THANK YOU

https://proteomics.cancer.gov

CPTAC Investigator Retreat
April 2019 I Bethesda, MD