An Introduction to the ISB-CGC Cancer Resources in the Google Cloud with BigQuery as a Statistical Tool

David Pot and Fabian Seidl for ISB-CGC

2024-03-27

ISB-CGC

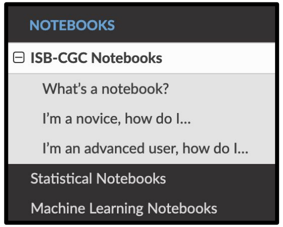
Outline of today’s agenda

- Introduction to ISB-CGC and the data commons
- ISB-CGC’s approach to derived data
- Data exploration in BigQuery (‘Excel-like data tables in the cloud’)
- Hands on demonstration of the Google Cloud with focus on BigQuery

The ISB-CGC homepage

isb-cgc.org

Focus today is on analysis of derived data in BigQuery



NOTEBOOKS

⊟ ISB-CGC Notebooks

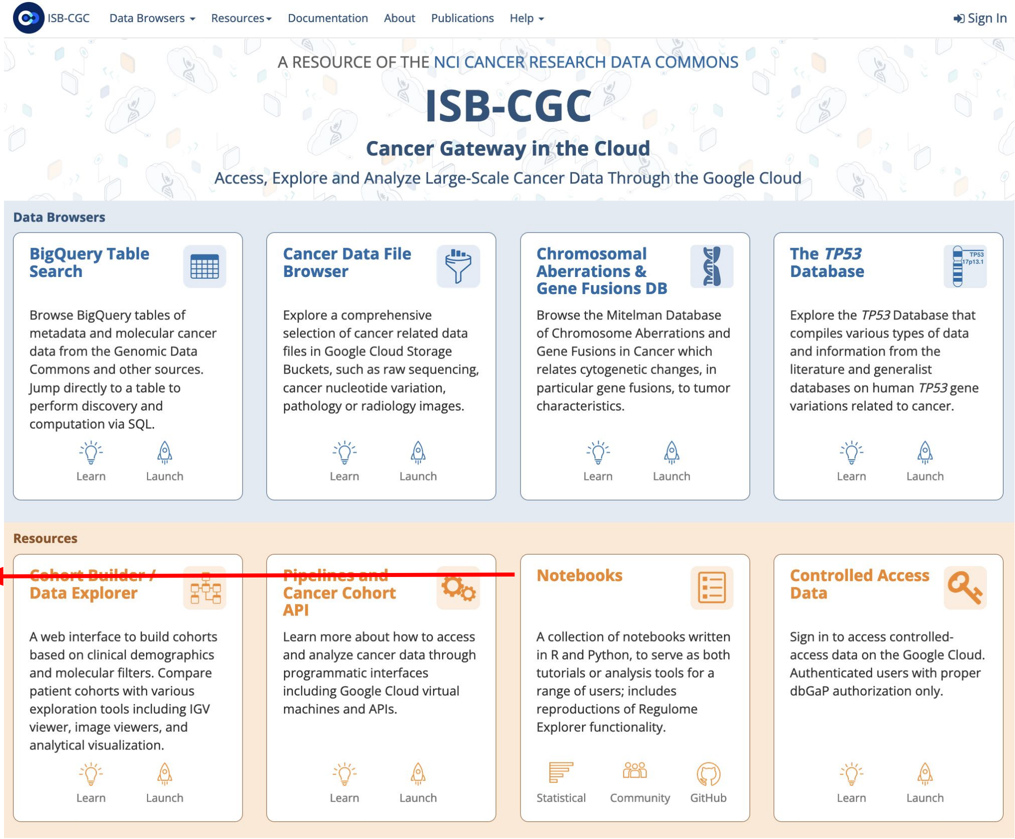
What's a notebook?

I'm a novice, how do I...

I'm an advanced user, how do I...

Statistical Notebooks

Machine Learning Notebooks



ISB-CGC | Data Browsers ▾ | Resources ▾ | Documentation | About | Publications | Help ▾ | Sign In

A RESOURCE OF THE NCI CANCER RESEARCH DATA COMMONS

ISB-CGC

Cancer Gateway in the Cloud

Access, Explore and Analyze Large-Scale Cancer Data Through the Google Cloud

Data Browsers

BigQuery Table Search

Browse BigQuery tables of metadata and molecular cancer data from the Genomic Data Commons and other sources. Jump directly to a table to perform discovery and computation via SQL.

Learn | Launch

Cancer Data File Browser

Explore a comprehensive selection of cancer related data files in Google Cloud Storage Buckets, such as raw sequencing, cancer nucleotide variation, pathology or radiology images.

Learn | Launch

Chromosomal Aberrations & Gene Fusions DB

Browse the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer which relates cytogenetic changes, in particular gene fusions, to tumor characteristics.

Learn | Launch

The *TP53* Database

Explore the *TP53* Database that compiles various types of data and information from the literature and generalist databases on human *TP53* gene variations related to cancer.

Learn | Launch

Resources

Cohort Builder / Data Explorer

A web interface to build cohorts based on clinical demographics and molecular filters. Compare patient cohorts with various exploration tools including IGV viewer, image viewers, and analytical visualization.

Learn | Launch

Pipelines and Cancer Cohort API

Learn more about how to access and analyze cancer data through programmatic interfaces including Google Cloud virtual machines and APIs.

Learn | Launch

Notebooks

A collection of notebooks written in R and Python, to serve as both tutorials or analysis tools for a range of users; includes reproductions of Regulome Explorer functionality.

Statistical | Community | GitHub

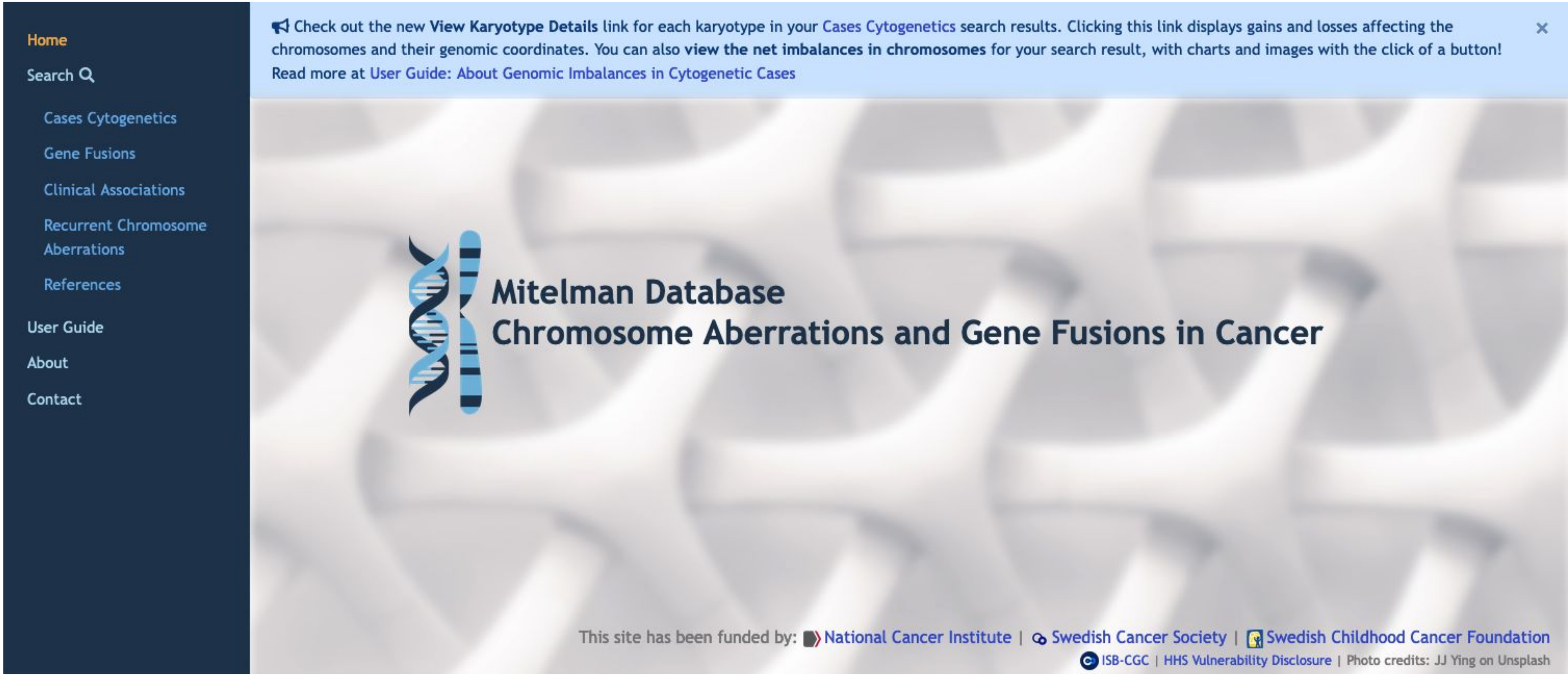
Controlled Access Data

Sign in to access controlled-access data on the Google Cloud. Authenticated users with proper dbGaP authorization only.

Learn | Launch

Ways to use data and tools:

- Explore datasets
- Create cohorts
- Run pipelines
- Specialized DBs
    - Mitelman DB of Chromosomal Aberrations & Gene Fusions
    - The *TP53* DB
    - caNanoLab

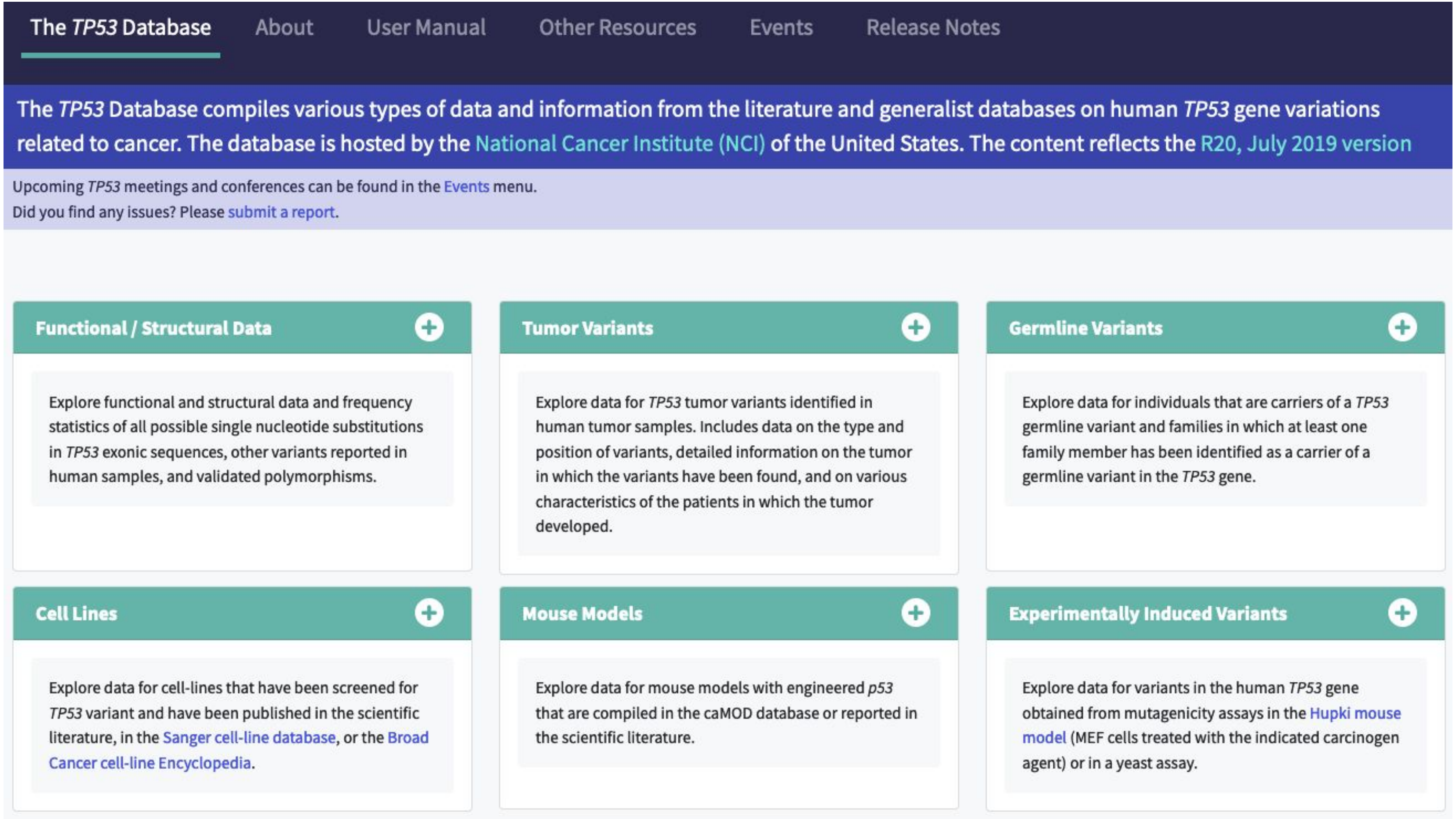


Check out the new **View Karyotype Details** link for each karyotype in your Cases Cytogenetics search results. Clicking this link displays gains and losses affecting the chromosomes and their genomic coordinates. You can also **view the net imbalances in chromosomes** for your search result, with charts and images with the click of a button! Read more at User Guide: About Genomic Imbalances in Cytogenetic Cases

Mitelman Database
Chromosome Aberrations and Gene Fusions in Cancer

This site has been funded by: National Cancer Institute | Swedish Cancer Society | Swedish Childhood Cancer Foundation
ISB-CGC | HHS Vulnerability Disclosure | Photo credits: JJ Ying on Unsplash

https://mitelmandatabase.isb-cgc.org/



The *TP53* Database compiles various types of data and information from the literature and generalist databases on human *TP53* gene variations related to cancer. The database is hosted by the National Cancer Institute (NCI) of the United States. The content reflects the R20, July 2019 version

Upcoming *TP53* meetings and conferences can be found in the Events menu.
Did you find any issues? Please submit a report.

Functional / Structural Data ⊕

Explore functional and structural data and frequency statistics of all possible single nucleotide substitutions in *TP53* exonic sequences, other variants reported in human samples, and validated polymorphisms.

Tumor Variants ⊕

Explore data for *TP53* tumor variants identified in human tumor samples. Includes data on the type and position of variants, detailed information on the tumor in which the variants have been found, and on various characteristics of the patients in which the tumor developed.

Germline Variants ⊕

Explore data for individuals that are carriers of a *TP53* germline variant and families in which at least one family member has been identified as a carrier of a germline variant in the *TP53* gene.

Cell Lines ⊕

Explore data for cell-lines that have been screened for *TP53* variant and have been published in the scientific literature, in the Sanger cell-line database, or the Broad Cancer cell-line Encyclopedia.

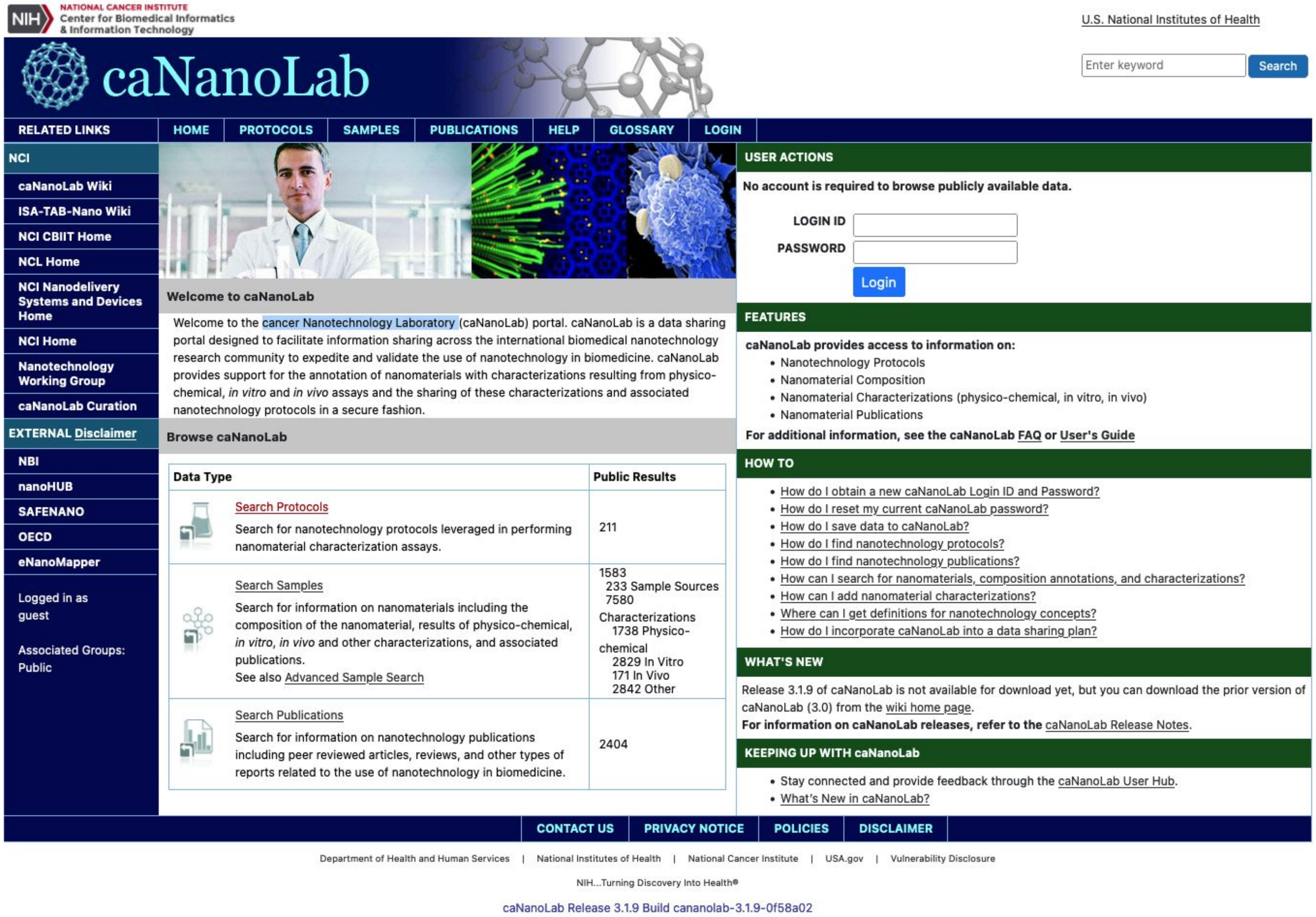
Mouse Models ⊕

Explore data for mouse models with engineered *p53* that are compiled in the caMOD database or reported in the scientific literature.

Experimentally Induced Variants ⊕

Explore data for variants in the human *TP53* gene obtained from mutagenicity assays in the Hupki mouse model (MEF cells treated with the indicated carcinogen agent) or in a yeast assay.

https://tp53.isb-cgc.org/



NIH NATIONAL CANCER INSTITUTE Center for Biomedical Informatics & Information Technology

U.S. National Institutes of Health

caNanoLab

Enter keyword Search

RELATED LINKS | HOME | PROTOCOLS | SAMPLES | PUBLICATIONS | HELP | GLOSSARY | LOGIN

NCI

- caNanoLab Wiki
- ISA-TAB-Nano Wiki
- NCI CBIIT Home
- NCL Home
- NCI Nanodelivery Systems and Devices Home
- NCI Home
- Nanotechnology Working Group
- caNanoLab Curation

EXTERNAL Disclaimer

- NBI
- nanoHUB
- SAFENANO
- OECD
- eNanoMapper

Logged in as guest

Associated Groups: Public

Welcome to caNanoLab

Welcome to the cancer Nanotechnology Laboratory (caNanoLab) portal. caNanoLab is a data sharing portal designed to facilitate information sharing across the international biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical, *in vitro* and *in vivo* assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.

Browse caNanoLab

Data Type	Public Results
Search Protocols Search for nanotechnology protocols leveraged in performing nanomaterial characterization assays.	211
Search Samples Search for information on nanomaterials including the composition of the nanomaterial, results of physico-chemical, *in vitro*, *in vivo* and other characterizations, and associated publications. See also Advanced Sample Search	1583 233 Sample Sources 7580 Characterizations 1738 Physico-chemical 2829 In Vitro 171 In Vivo 2842 Other
Search Publications Search for information on nanotechnology publications including peer reviewed articles, reviews, and other types of reports related to the use of nanotechnology in biomedicine.	2404

USER ACTIONS

No account is required to browse publicly available data.

LOGIN ID

PASSWORD

Login

FEATURES

caNanoLab provides access to information on:

- Nanotechnology Protocols
- Nanomaterial Composition
- Nanomaterial Characterizations (physico-chemical, in vitro, in vivo)
- Nanomaterial Publications

For additional information, see the caNanoLab FAQ or User's Guide

HOW TO

- How do I obtain a new caNanoLab Login ID and Password?
- How do I reset my current caNanoLab password?
- How do I save data to caNanoLab?
- How do I find nanotechnology protocols?
- How do I find nanotechnology publications?
- How can I search for nanomaterials, composition annotations, and characterizations?
- How can I add nanomaterial characterizations?
- Where can I get definitions for nanotechnology concepts?
- How do I incorporate caNanoLab into a data sharing plan?

WHAT'S NEW

Release 3.1.9 of caNanoLab is not available for download yet, but you can download the prior version of caNanoLab (3.0) from the wiki home page.

For information on caNanoLab releases, refer to the caNanoLab Release Notes.

KEEPING UP WITH caNanoLab

- Stay connected and provide feedback through the caNanoLab User Hub.
- What's New in caNanoLab?

CONTACT US | PRIVACY NOTICE | POLICIES | DISCLAIMER

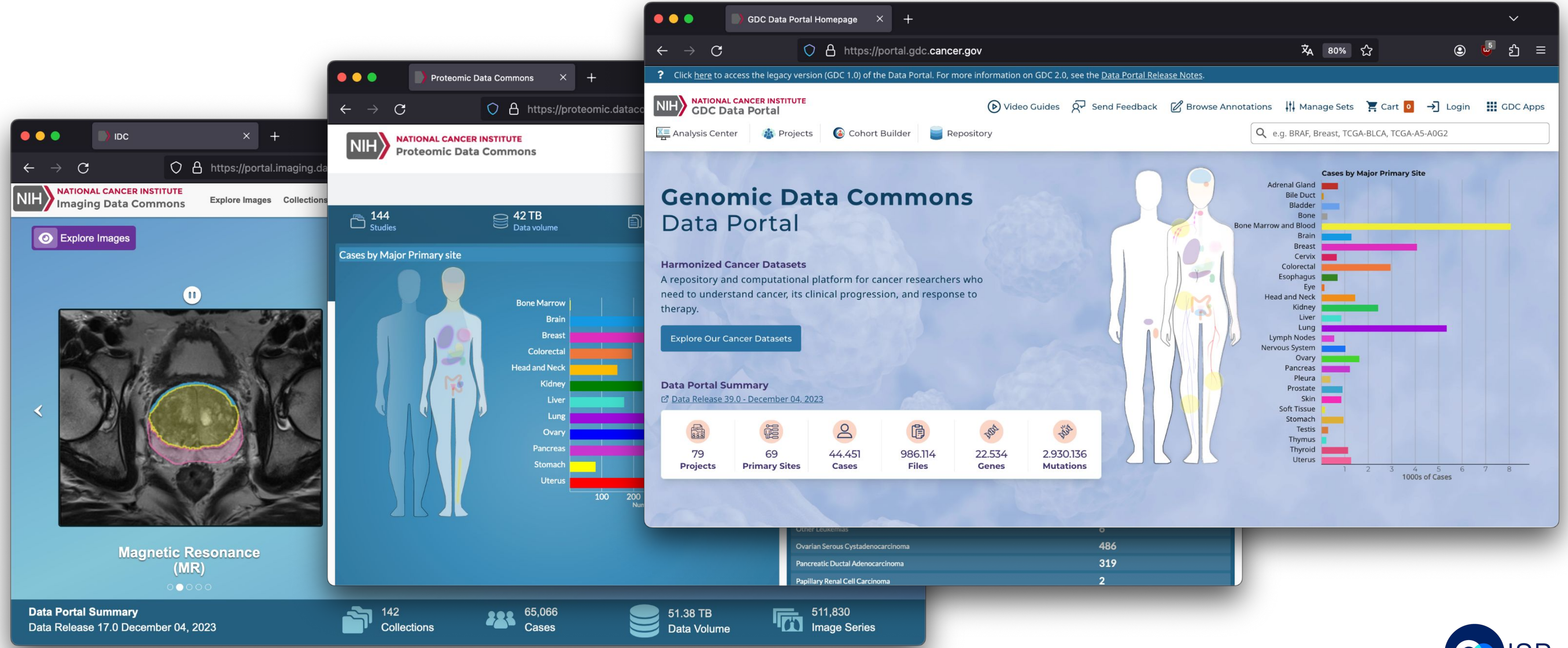
Department of Health and Human Services | National Institutes of Health | National Cancer Institute | USA.gov | Vulnerability Disclosure

NIH...Turning Discovery Into Health®

caNanoLab Release 3.1.9 Build cananolab-3.1.9-0f58a02

https://cananolab.cancer.gov

The three Data Commons host and control access to different types of cancer data



IDC

https://portal.imaging.da

NATIONAL CANCER INSTITUTE
Imaging Data Commons

Explore Images Collections

Explore Images

Magnetic Resonance (MR)

Data Portal Summary
Data Release 17.0 December 04, 2023

142 Collections	65,066 Cases	51.38 TB Data Volume	511,830 Image Series

Proteomic Data Commons

https://proteomic.datac

NATIONAL CANCER INSTITUTE
Proteomic Data Commons

144 Studies	42 TB Data volume

Cases by Major Primary site

Other Leukemias	6
Ovarian Serous Cystadenocarcinoma	486
Pancreatic Ductal Adenocarcinoma	319
Papillary Renal Cell Carcinoma	2

GDC Data Portal Homepage

https://portal.gdc.cancer.gov

Click here to access the legacy version (GDC 1.0) of the Data Portal. For more information on GDC 2.0, see the Data Portal Release Notes.

NATIONAL CANCER INSTITUTE
GDC Data Portal

Video Guides · Send Feedback · Browse Annotations · Manage Sets · Cart 0 · Login · GDC Apps

Analysis Center · Projects · Cohort Builder · Repository

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

Genomic Data Commons Data Portal

Harmonized Cancer Datasets

A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

Explore Our Cancer Datasets

Data Portal Summary
Data Release 39.0 - December 04, 2023

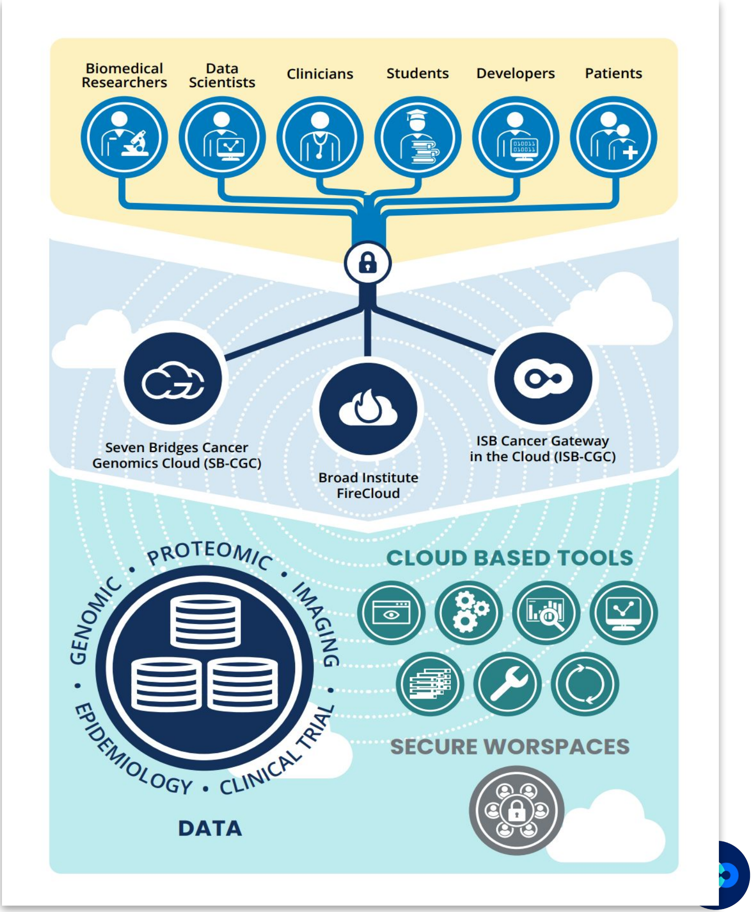
79 Projects	69 Primary Sites	44.451 Cases	986.114 Files	22.534 Genes	2.930.136 Mutations

Cases by Major Primary Site

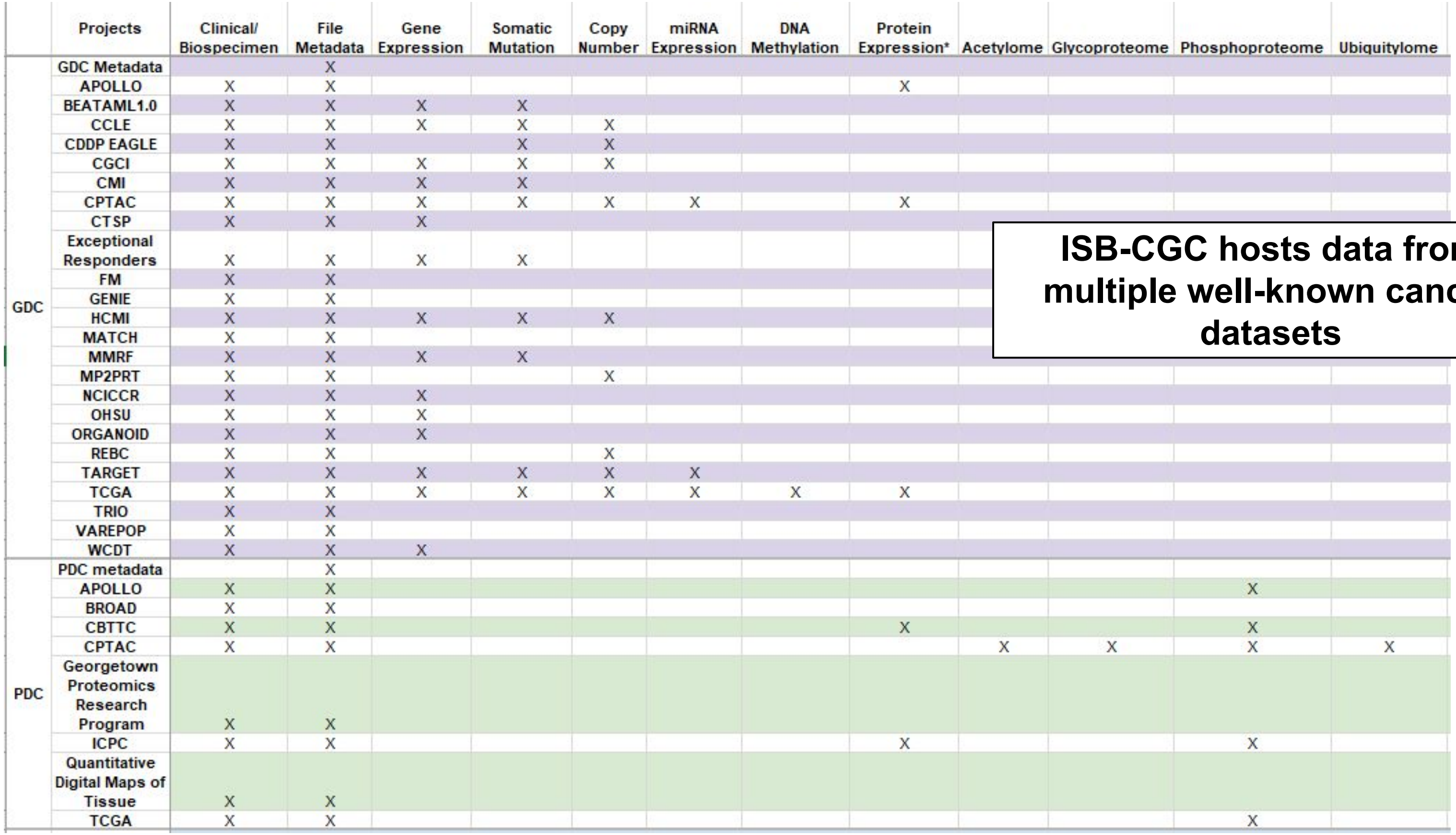
ISB-CGC

ISB-CGC's approach to enabling data science in the cloud

- Moving Excel files into the cloud
- Derived molecular data available for query as you need, updated frequently
- Tooling examples provided to enable data mining and Machine Learning of your data
- Sharing of results with those you choose
- Maximum flexibility of scripting and compute for those who desire it


Biomedical Researchers
Data Scientists
Clinicians
Students
Developers
Patients
Seven Bridges Cancer Genomics Cloud (SB-CGC)
Broad Institute FireCloud
ISB Cancer Gateway in the Cloud (ISB-CGC)
PROTEOMIC • IMAGING • CLINICAL TRIAL • EPIDEMIOLOGY • GENOMIC •
DATA
CLOUD BASED TOOLS
SECURE WORSPACES

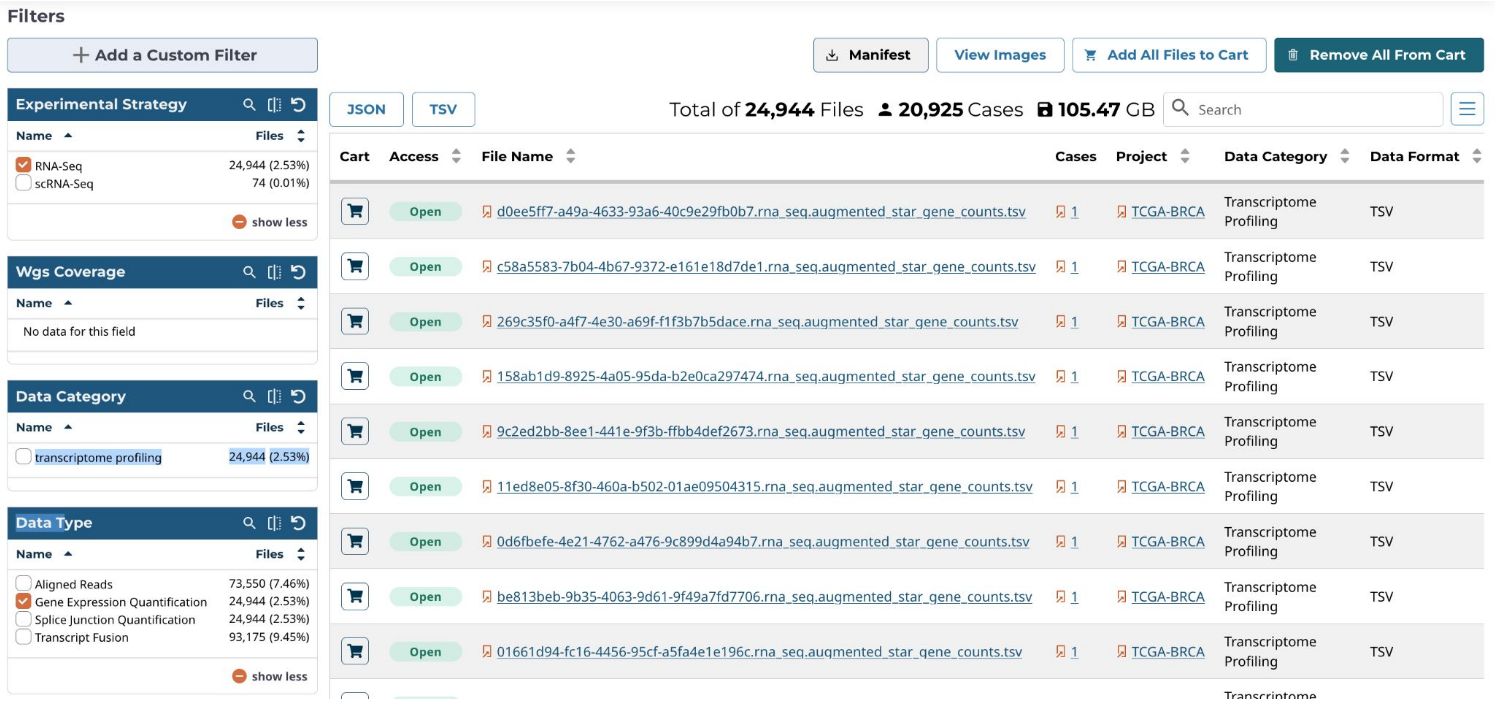
ISB-CGC Focuses on Derived Data via BigQuery



	Projects	Clinical/ Biospecimen	File Metadata	Gene Expression	Somatic Mutation	Copy Number	miRNA Expression	DNA Methylation	Protein Expression*	Acetylome	Glycoproteome	Phosphoproteome	Ubiquitylome
GDC	GDC Metadata		X										
	APOLLO	X	X						X				
	BEATAML1.0	X	X	X	X								
	CCLE	X	X	X	X	X							
	CDDP EAGLE	X	X		X	X							
	CGCI	X	X	X	X	X							
	CMI	X	X	X	X								
	CPTAC	X	X	X	X	X	X		X				
	CTSP	X	X	X									
	Exceptional Responders	X	X	X	X								
	FM	X	X										
	GENIE	X	X										
	HCMI	X	X	X	X	X							
	MATCH	X	X										
	MMRF	X	X	X	X								
	MP2PRT	X	X			X							
	NCICCR	X	X	X									
	OHSU	X	X	X									
	ORGANOID	X	X	X									
	REBC	X	X			X							
	TARGET	X	X	X	X	X	X						
	TCGA	X	X	X	X	X	X	X	X				
	TRIO	X	X										
	VAREPOP	X	X										
	WCDT	X	X	X									
PDC	PDC metadata		X										
	APOLLO	X	X									X	
	BROAD	X	X										
	CBTTC	X	X						X			X	
	CPTAC	X	X							X	X	X	X
	Georgetown Proteomics Research Program	X	X										
	ICPC	X	X						X			X	
	Quantitative Digital Maps of Tissue	X	X										
	TCGA	X	X									X	

ISB-CGC hosts data from multiple well-known cancer datasets

Data wrangling can be onerous, for example GDC has 24,944 individual transcriptome files for just TCGA



Experimental Strategy

Name	Files
☑ RNA-Seq	24,944 (2.53%)
☐ scRNA-Seq	74 (0.01%)

Wgs Coverage

Name	Files
No data for this field	

Data Category

Name	Files
☐ transcriptome profiling	24,944 (2.53%)

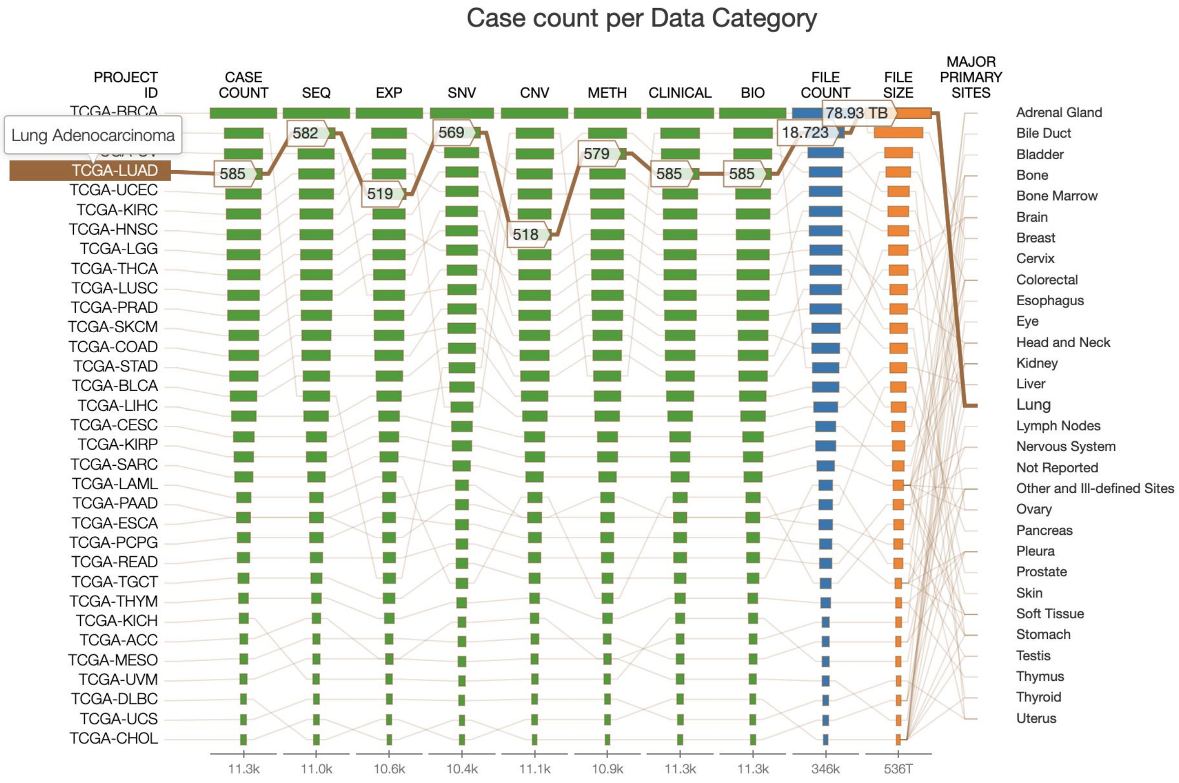
Data Type

Name	Files
☐ Aligned Reads	73,550 (7.46%)
☑ Gene Expression Quantification	24,944 (2.53%)
☐ Splice Junction Quantification	24,944 (2.53%)
☐ Transcript Fusion	93,175 (9.45%)

Total of **24,944** Files **20,925** Cases **105.47** GB

Cart	Access	File Name	Cases	Project	Data Category	Data Format
	Open	d0ee5ff7-a49a-4633-93a6-40c9e29fb0b7.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	c58a5583-7b04-4b67-9372-e161e18d7de1.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	269c35f0-a4f7-4e30-a69f-f1f3b7b5dace.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	158ab1d9-8925-4a05-95da-b2e0ca297474.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	9c2ed2bb-8ee1-441e-9f3b-ffbb4def2673.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	11ed8e05-8f30-460a-b502-01ae09504315.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	0d6fbefe-4e21-4762-a476-9c899d4a94b7.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	be813beb-9b35-4063-9d61-9f49a7fd7706.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
	Open	01661d94-fc16-4456-95cf-a5fa4e1e196c.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-BRCA	Transcriptome Profiling	TSV
					Transcriptome	

The data commons host a wealth of data from 20 cancer initiative programs

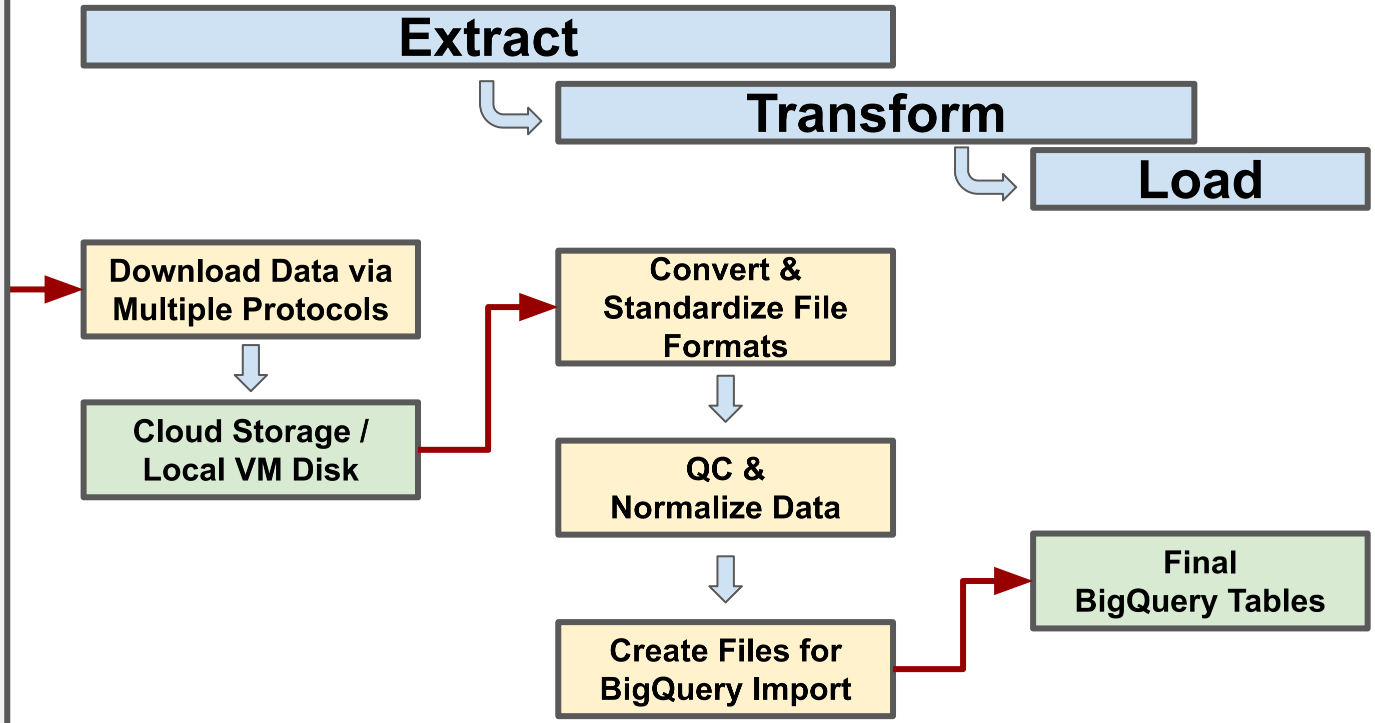


ISB-CGC runs ETL pipelines to reduce the processing barrier of entry

BEATAML	OHSU
CCLE	ORGANOID
CGCI	TARGET
CMI	TCGA
CPTAC	VAREPOP
CTSP	WCDT
FM	CBTN
GENIE	CPTAC
HCMI	ICPC
MMRF	Targetome
NCICCR	Reactome
Pan-Cancer Atlas	
Georgetown Proteomics Research Program	
Quantitative Digital Maps of Tissue Biopsies	

500k+ Files of Heterogeneous Data

WGS	DNASeq WXS
RNASeq (gene, isoform, exon, junction)	
SNP Array (CEL)	
DNASeq (MAF, VCF)	Clinical & Biospecimen
DNA Methylation	miRNASeq
Protein (RPPA)	SNP Array



Extract

Transform

Load

Download Data via Multiple Protocols

Cloud Storage / Local VM Disk

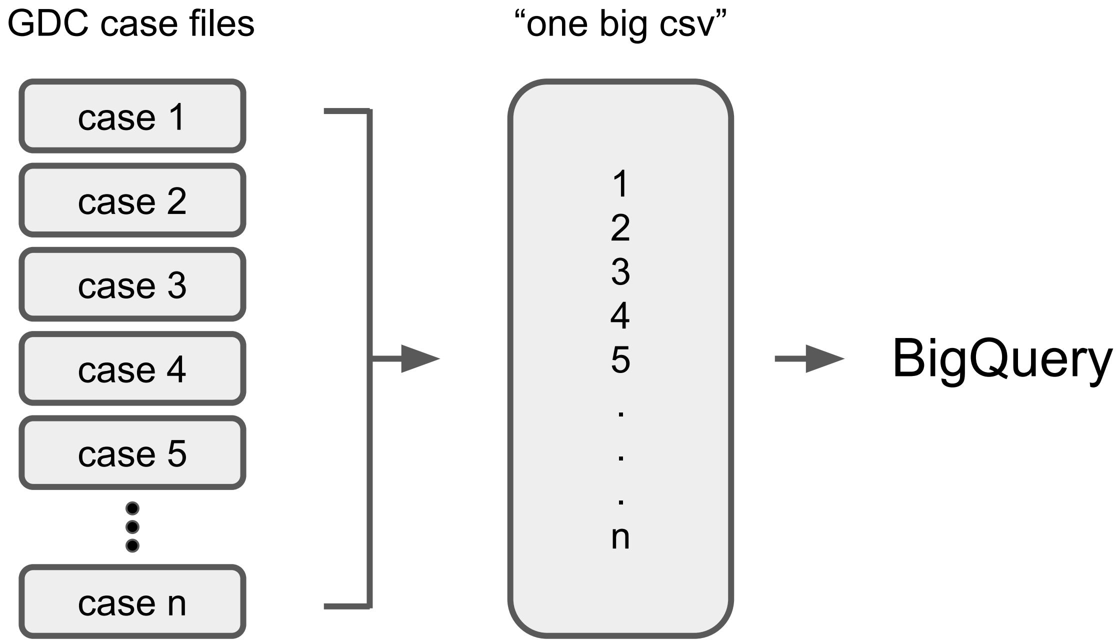
Convert & Standardize File Formats

QC & Normalize Data

Create Files for BigQuery Import

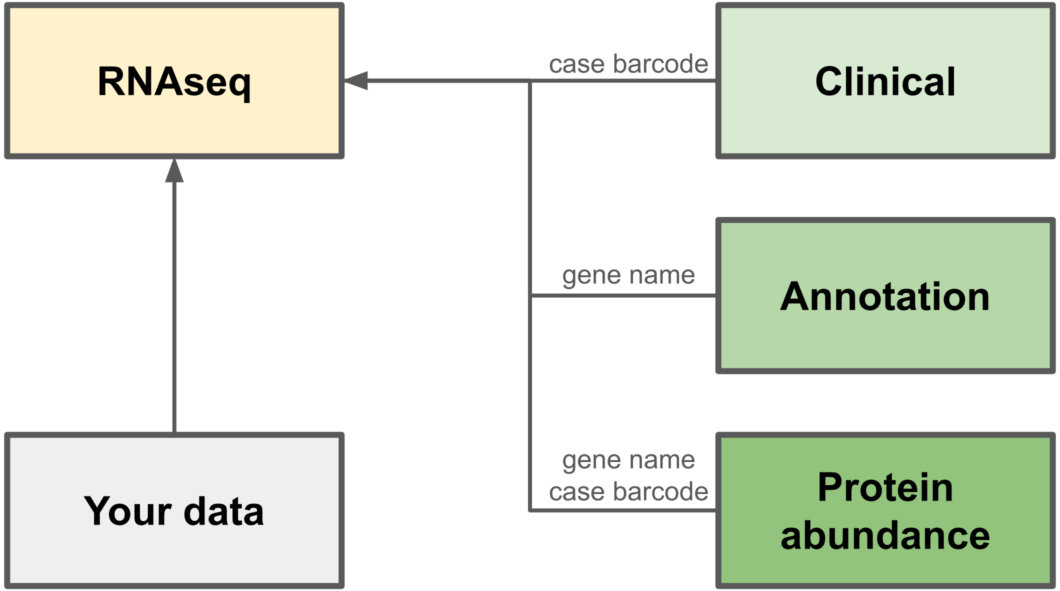
Final BigQuery Tables

The Google Cloud offers tools to simply host derived data by concatenating these files into a single BQ table



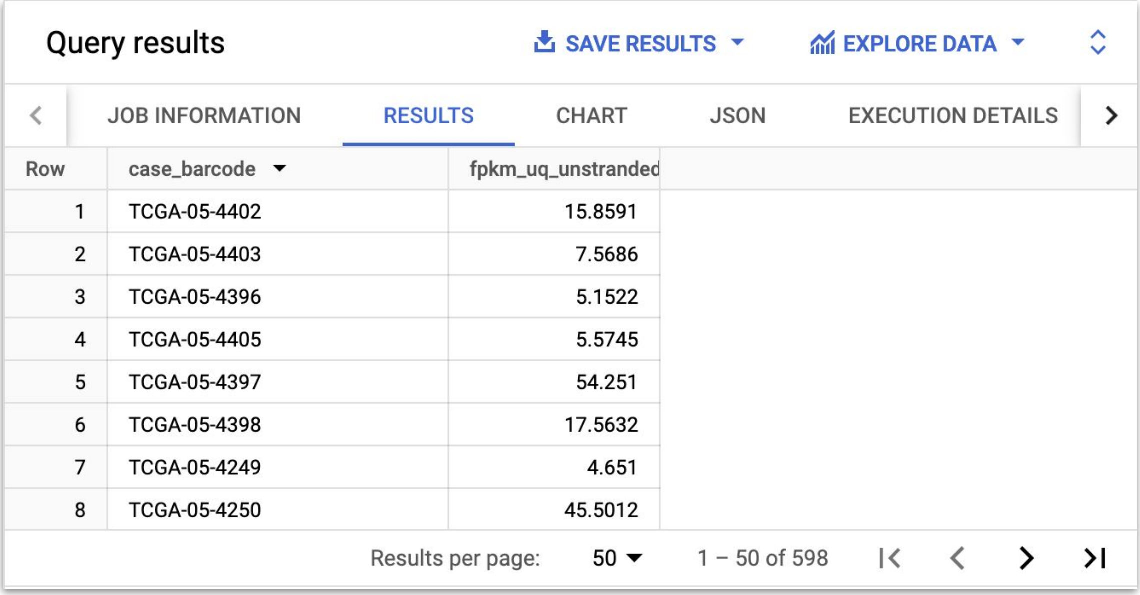
ISB-CGC

BigQuery enables simple and efficient links between data types



ISB-CGC

BigQuery enables simple and efficient linking between tables



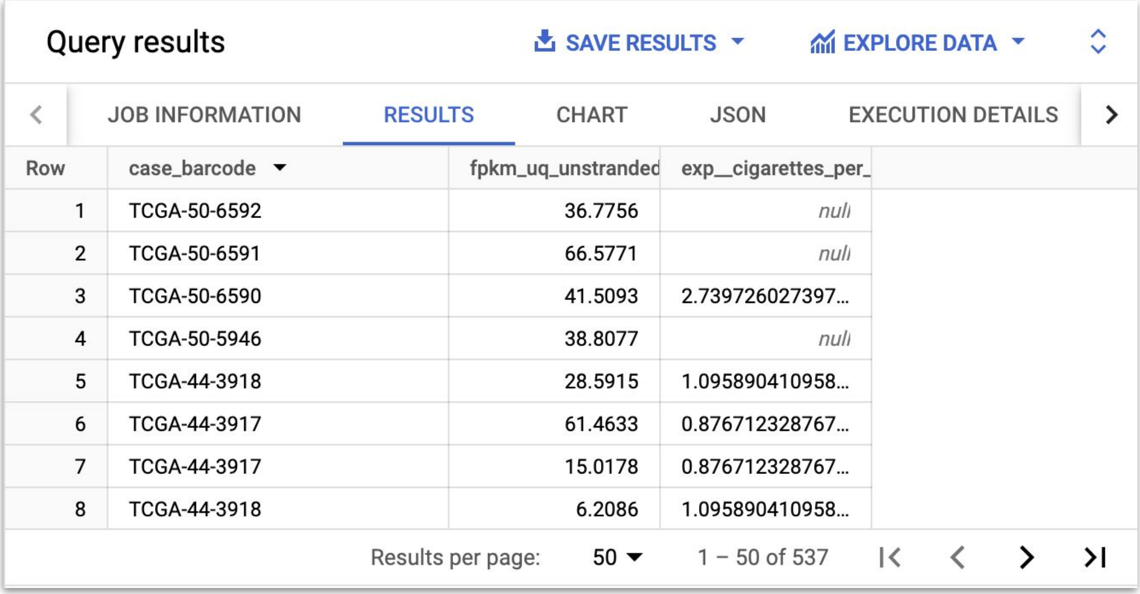
Row	case_barcode	fpkm_uq_unstranded
1	TCGA-05-4402	15.8591
2	TCGA-05-4403	7.5686
3	TCGA-05-4396	5.1522
4	TCGA-05-4405	5.5745
5	TCGA-05-4397	54.251
6	TCGA-05-4398	17.5632
7	TCGA-05-4249	4.651
8	TCGA-05-4250	45.5012

Results per page: 50 1 – 50 of 598

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current` rna
WHERE <conditionals>
```

ISB-CGC

BigQuery enables simple and efficient linking between tables

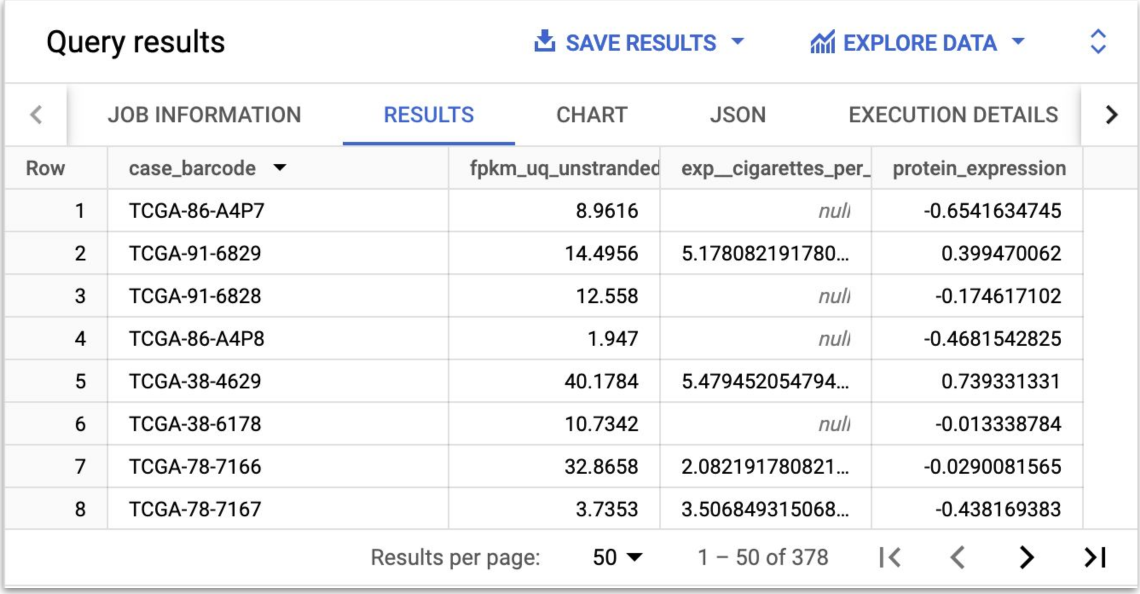


Row	case_barcode	fpkm_uq_unstranded	exp__cigarettes_per_
1	TCGA-50-6592	36.7756	*null*
2	TCGA-50-6591	66.5771	*null*
3	TCGA-50-6590	41.5093	2.739726027397...
4	TCGA-50-5946	38.8077	*null*
5	TCGA-44-3918	28.5915	1.095890410958...
6	TCGA-44-3917	61.4633	0.876712328767...
7	TCGA-44-3917	15.0178	0.876712328767...
8	TCGA-44-3918	6.2086	1.095890410958...

Results per page: 50 1 – 50 of 537

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current` rna
JOIN `isb-cgc-bq.TCGA.clinical_gdc_current` clin
WHERE <conditionals>
```

BigQuery enables simple and efficient linking between tables

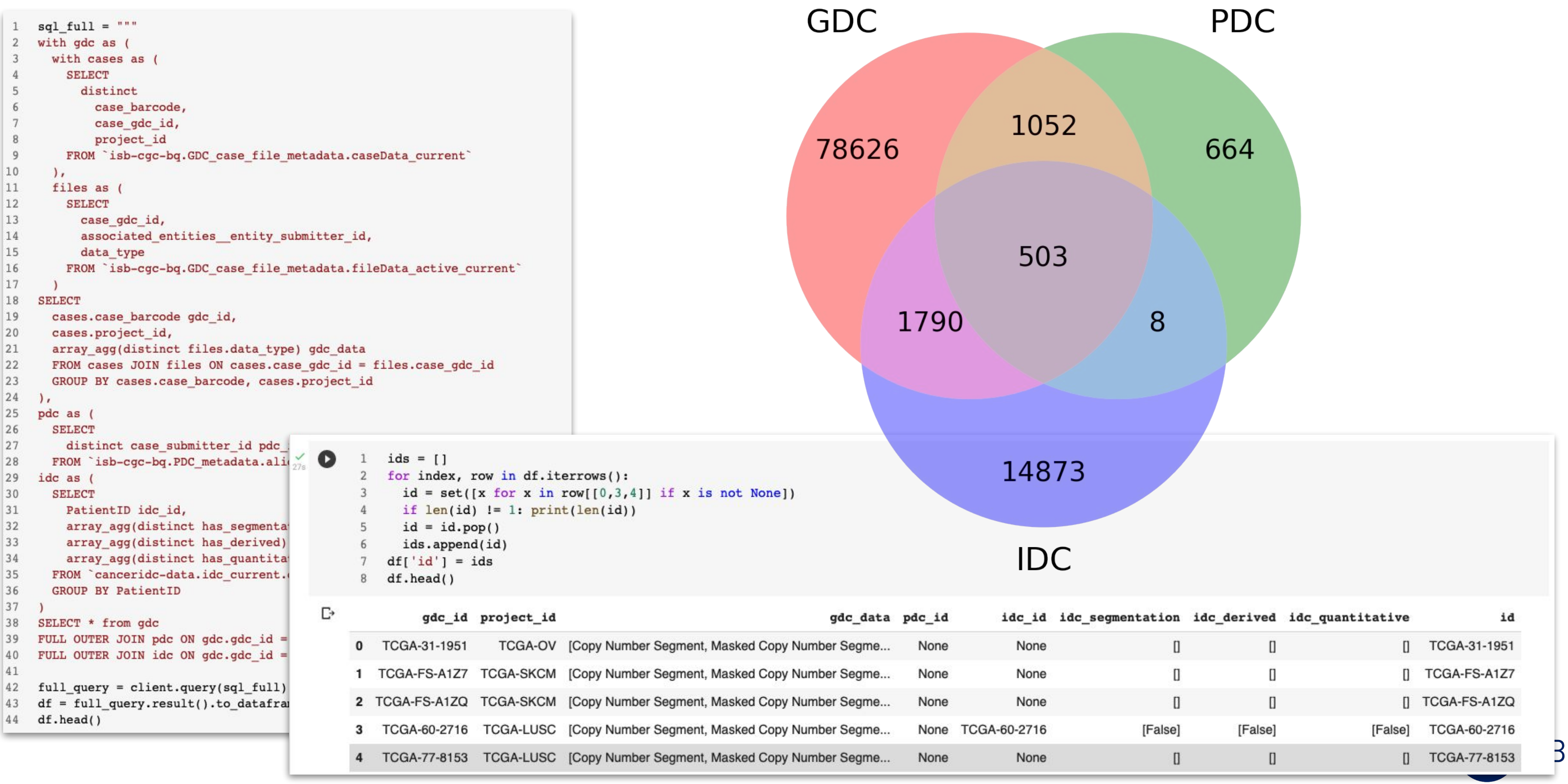


Row	case_barcode	fpkm_uq_unstranded	exp__cigarettes_per_	protein_expression
1	TCGA-86-A4P7	8.9616	*null*	-0.6541634745
2	TCGA-91-6829	14.4956	5.178082191780...	0.399470062
3	TCGA-91-6828	12.558	*null*	-0.174617102
4	TCGA-86-A4P8	1.947	*null*	-0.4681542825
5	TCGA-38-4629	40.1784	5.479452054794...	0.739331331
6	TCGA-38-6178	10.7342	*null*	-0.013338784
7	TCGA-78-7166	32.8658	2.082191780821...	-0.0290081565
8	TCGA-78-7167	3.7353	3.506849315068...	-0.438169383

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current` rna
JOIN `isb-cgc-bq.TCGA.clinical_gdc_current` clin
JOIN `isb-cgc-bq.TCGA.protein_expression_hg38_gdc_current` prot
WHERE <conditionals>
```

ISB-CGC

There is a substantial overlap of data from the same cases across the Commons



```
sql_full = """
with gdc as (
  with cases as (
    SELECT
      distinct
        case_barcode,
        case_gdc_id,
        project_id
    FROM `isb-cgc-bq.GDC_case_file_metadata.caseData_current`
  ),
  files as (
    SELECT
      case_gdc_id,
      associated_entities__entity_submitter_id,
      data_type
    FROM `isb-cgc-bq.GDC_case_file_metadata.fileData_active_current`
  )
SELECT
  cases.case_barcode gdc_id,
  cases.project_id,
  array_agg(distinct files.data_type) gdc_data
  FROM cases JOIN files ON cases.case_gdc_id = files.case_gdc_id
  GROUP BY cases.case_barcode, cases.project_id
),
pdc as (
  SELECT
    distinct case_submitter_id pdc_
  FROM `isb-cgc-bq.PDC_metadata.ali
idc as (
  SELECT
    PatientID idc_id,
    array_agg(distinct has_segmenta
    array_agg(distinct has_derived)
    array_agg(distinct has_quantita
  FROM `canceridc-data.idc_current.
  GROUP BY PatientID
)
SELECT * from gdc
FULL OUTER JOIN pdc ON gdc.gdc_id =
FULL OUTER JOIN idc ON gdc.gdc_id =

full_query = client.query(sql_full)
df = full_query.result().to_datafra
df.head()
```

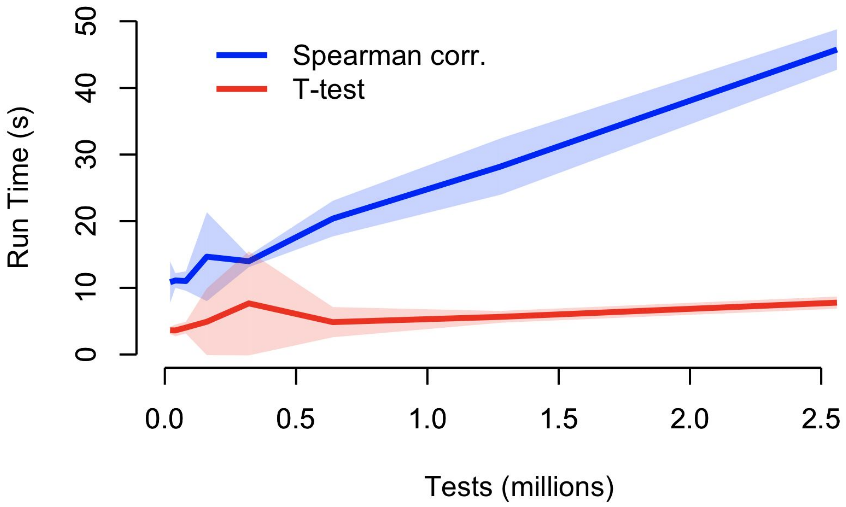
```
ids = []
for index, row in df.iterrows():
  id = set([x for x in row[[0,3,4]] if x is not None])
  if len(id) != 1: print(len(id))
  id = id.pop()
  ids.append(id)
df['id'] = ids
df.head()
```

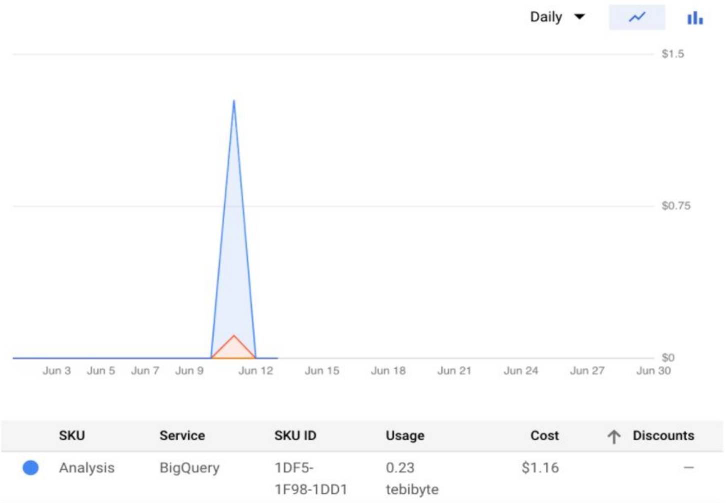
```
         gdc_id  project_id                                          gdc_data  pdc_id        idc_id  idc_segmentation  idc_derived  idc_quantitative            id
0  TCGA-31-1951     TCGA-OV  [Copy Number Segment, Masked Copy Number Segme...    None          None                []           []                []  TCGA-31-1951
1  TCGA-FS-A1Z7   TCGA-SKCM  [Copy Number Segment, Masked Copy Number Segme...    None          None                []           []                []  TCGA-FS-A1Z7
2  TCGA-FS-A1ZQ   TCGA-SKCM  [Copy Number Segment, Masked Copy Number Segme...    None          None                []           []                []  TCGA-FS-A1ZQ
3  TCGA-60-2716   TCGA-LUSC  [Copy Number Segment, Masked Copy Number Segme...    None  TCGA-60-2716           [False]      [False]           [False]  TCGA-60-2716
4  TCGA-77-8153   TCGA-LUSC  [Copy Number Segment, Masked Copy Number Segme...    None          None                []           []                []  TCGA-77-8153
```

BigQuery is a powerful statistical tool that can run hundreds of millions of tests in seconds

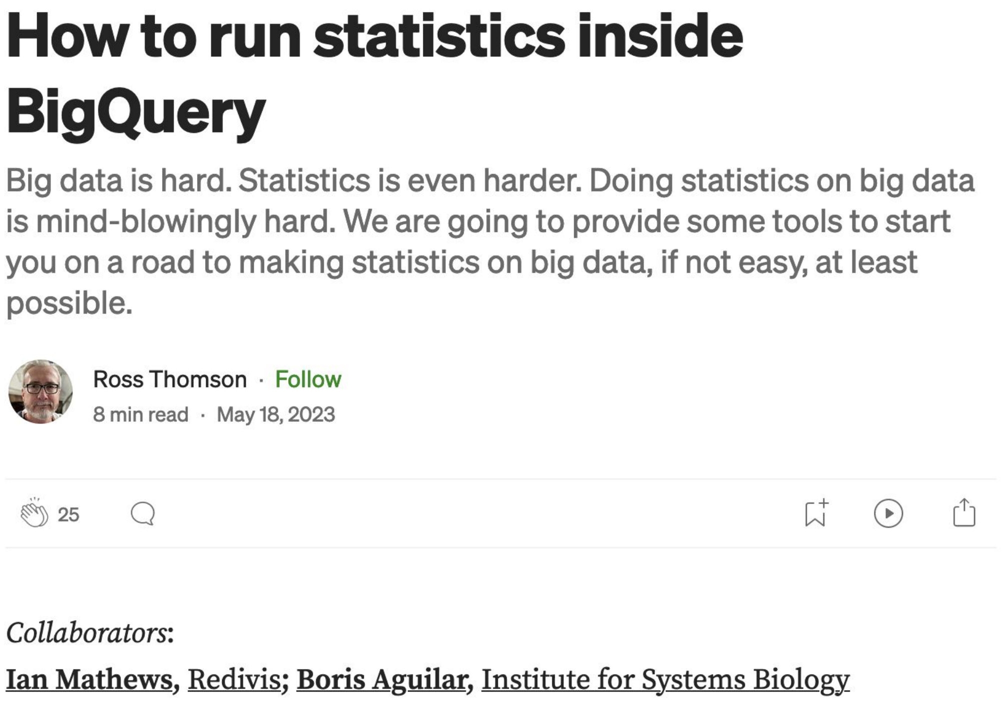
Testing BigQuery compute time with statistical tests

- Millions of tests in 40 seconds
- 6.6 billion correlations for $1.16





	SKU	Service	SKU ID	Usage	Cost	↑ Discounts
●	Analysis	BigQuery	1DF5-1F98-1DD1	0.23 tebibyte	$1.16	–



How to run statistics inside BigQuery

Big data is hard. Statistics is even harder. Doing statistics on big data is mind-blowingly hard. We are going to provide some tools to start you on a road to making statistics on big data, if not easy, at least possible.

Ross Thomson · Follow

8 min read · May 18, 2023

Collaborators:

Ian Mathews, Redivis; **Boris Aguilar**, Institute for Systems Biology

https://medium.com/@jrossthomson/how-to-run-statistics-inside-bigquery-95c0c6864f23

I’ll show hands on navigation of working in the Google Cloud Console

- BigQuery Search Tool
- Google Cloud Console
  - VMs and pipelines
  - Navigating BigQuery
- Create a project
- Upload a small table
- Introductory exploration
- Notebooks section

isb-cgc.org

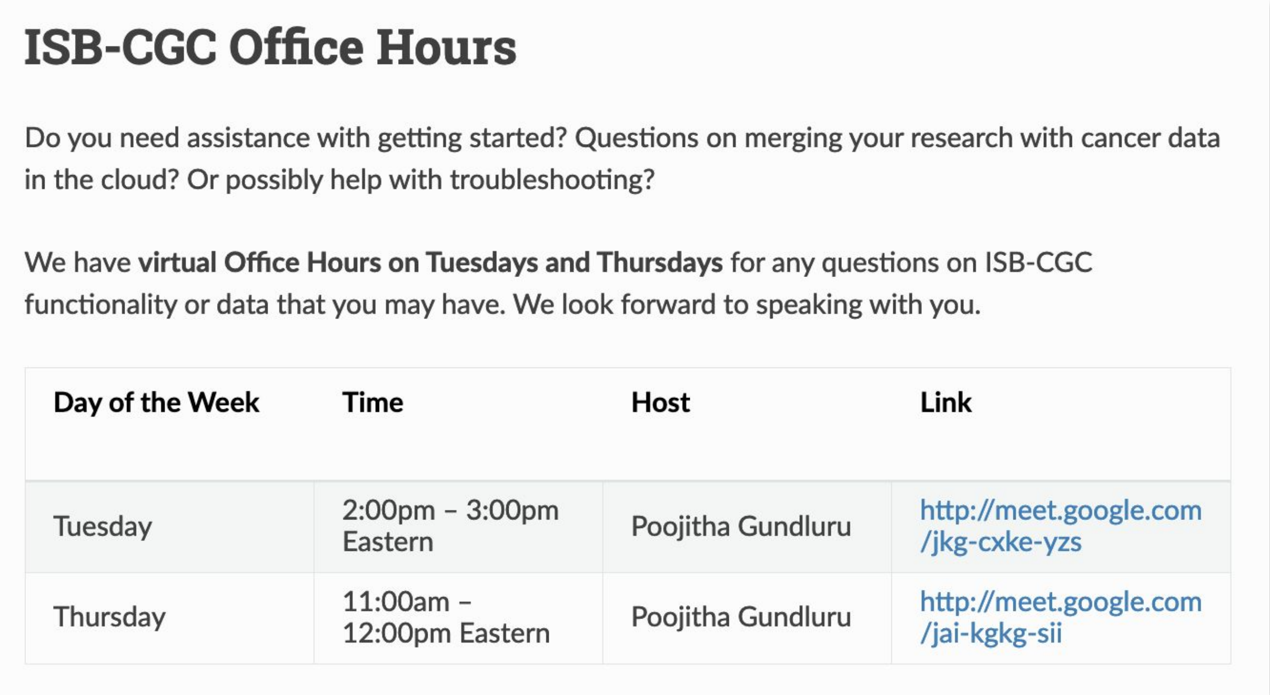
ISB-CGC

The benefits of working with ISB-CGC in the cloud

- Multiple specialized databases such as Mittleman
- Easy exploration of existing GDC and PDC data
- Access Virtual Machines and controlled data for customized pipelines
- BigQuery as a tool with scaling Excel functionality
    - Affordable storage and sharing of tabular data
    - Data exploration and quick statistics
    - Derived data from well known reference NCI datasets and annotations
    - Fast links between diverse data types
    - Advanced statistical analyses using Python, R, and Bioconductor
    - Rapidly able to expand to Machine Learning

✉ **feedback@isb-cgc.org**

@isb-cgc



ISB-CGC Office Hours

Do you need assistance with getting started? Questions on merging your research with cancer data in the cloud? Or possibly help with troubleshooting?

We have **virtual Office Hours on Tuesdays and Thursdays** for any questions on ISB-CGC functionality or data that you may have. We look forward to speaking with you.

Day of the Week	Time	Host	Link
Tuesday	2:00pm – 3:00pm Eastern	Poojitha Gundluru	http://meet.google.com/jkg-cxke-yzs
Thursday	11:00am – 12:00pm Eastern	Poojitha Gundluru	http://meet.google.com/jai-kgkg-sii

The ISB-CGC team

ISB

Bill Longabaugh

Suzanne Paquette

Bill Clifford

Elaine Lee

Mi Tian

Lauren Hagen

Boris Aguilar

Lauren Wolfe

Ilya Shmulevich

GENERAL DYNAMICS
Information Technology

David Pot

Danna Huffman

Fabian Seidl

Jacob Wilson

Poojitha Gundluru

Prema Venkatesan

Deena Bleich