



Cancer -omics analysis on the ISB-CGC platform

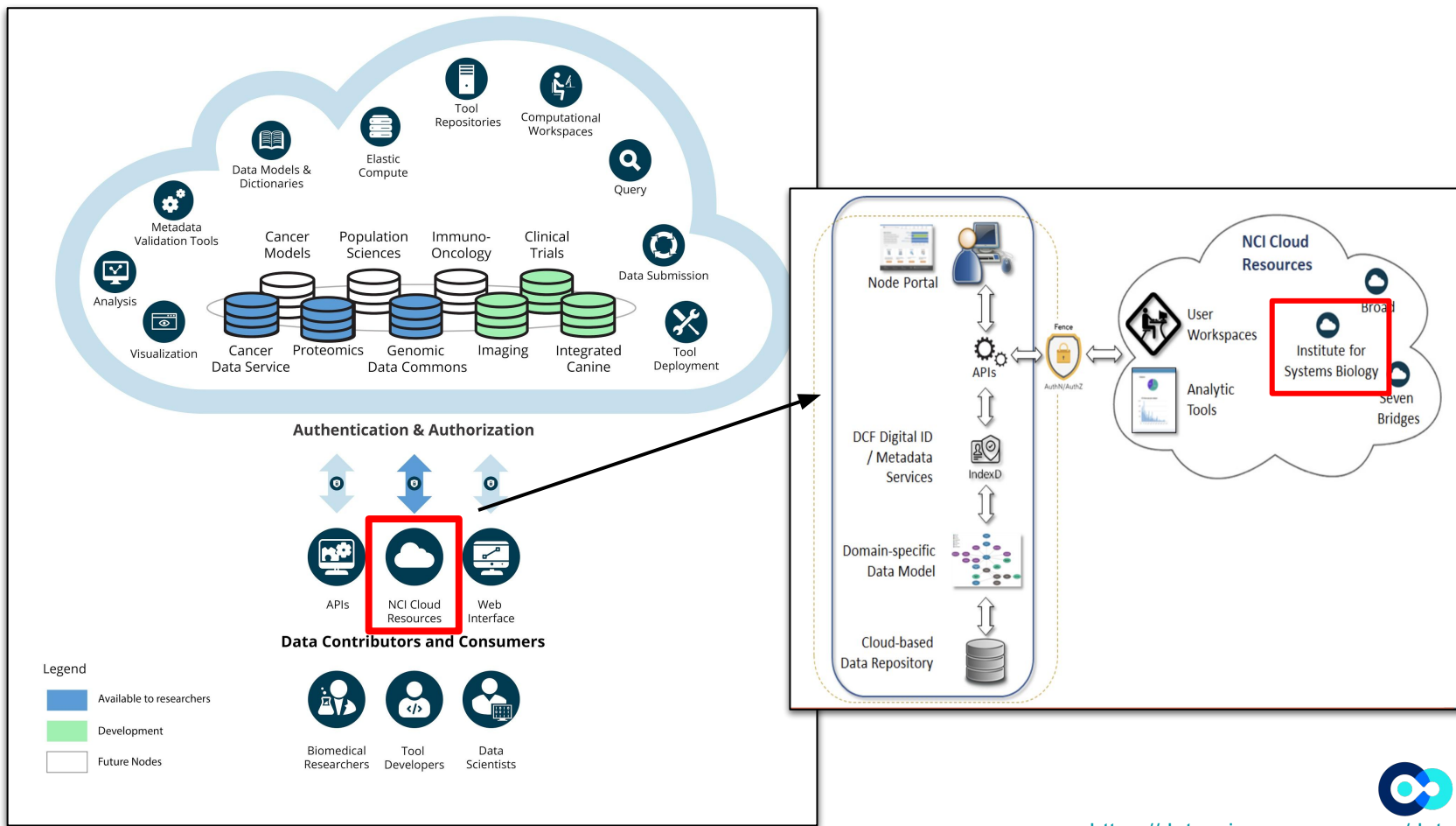
2020-03-17

Kawther Abdilleh - Bioinformatics Scientist
Fabian Seidl - Bioinformatics Scientist
Bill Longabaugh - Co-PI ISB-CGC

Outline

- What is ISB-CGC?
- How do users interact with the ISB-CGC platform?
- What resources can be used to interoperate with ISB-CGC?
- What are the policy and security restrictions that users need to know?

NCI Cancer Research Data Commons Ecosystem

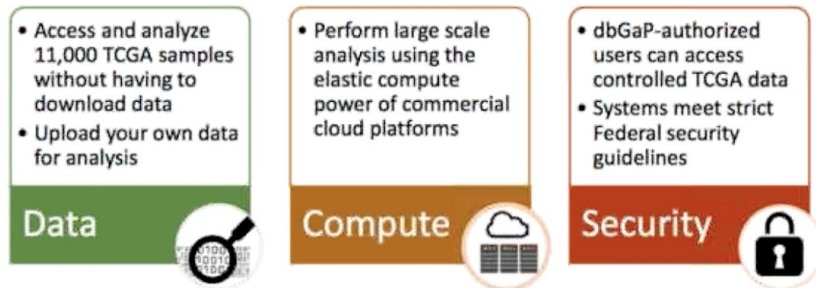


ISB-CGC is one of the NCI Cloud Resources

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

Provide:

- Access to large genomic data sets without need to download
- Access to popular pipelines and visualization tools
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses



 #NCICloud

ISB-CGC provides Data as a Service (DaaS) solutions to the rapid growth of cancer data

Common problems of big data:

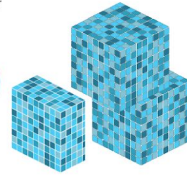
- Transfer speeds become bottlenecks with scaling data size
- Availability of data is tenuous
- Data discovery is onerous

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

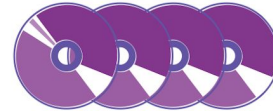
TCGA produced over

2.5
PETABYTES
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000
DVDs



TCGA data describes



33
DIFFERENT
TUMOR TYPES

...including

10
RARE
CANCERS

...based on paired tumor and normal tissue sets collected from



11,000
PATIENTS

...using

7

DIFFERENT
DATA TYPES



TCGA RESULTS & FINDINGS

Our mission at ISB-CGC

To make NCI multi-omics cancer data as well as high-performance compute resources available via the Google Cloud Platform through multiple modes:

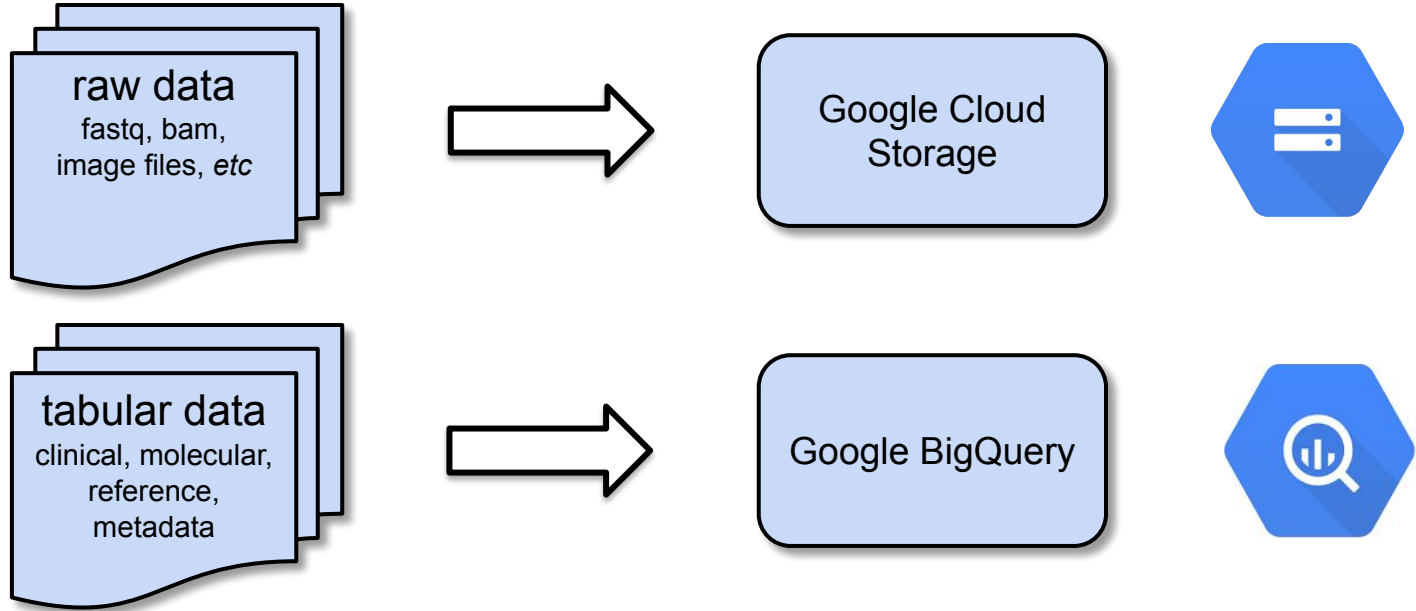
- Interactive web application for cohort building and data discovery
- Easily accessible and query-able tables for multivariate data analysis
- Advanced pipeline and workflow execution on Google Cloud virtual machines

<https://isb-cgc.org>

Our Approach at ISB-CGC

- Build an open platform for a broad range of users and use-cases
- Use existing systems to minimize development and maintenance costs
- Leverage the best existing Google tools and technologies
- Collaborate with the research community
- Provide a range of examples and tutorials

How do users access data on ISB-CGC?

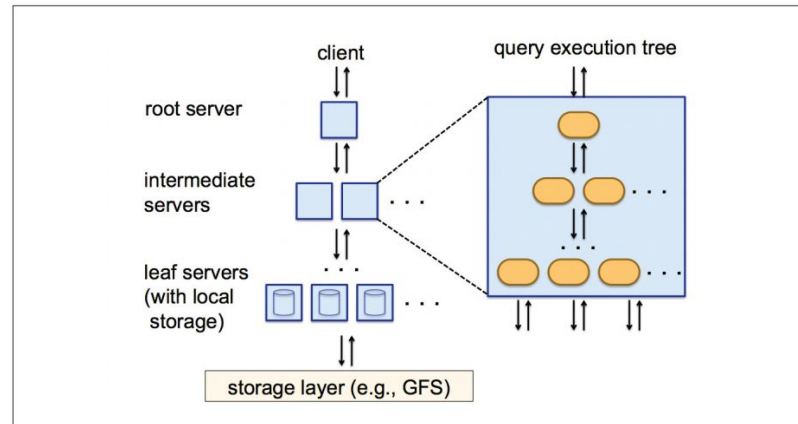


What is Google BigQuery and how does it enable –omics analyses?

- Cloud-based web service from Google Cloud used for handling and analyzing big data
- In the world of “omics”, it can facilitate high-throughput data analysis on the Cloud inexpensively in the following ways:
 - **Storage:**
 - Store the results from large-scale pipelines/workflows in centralized BigQuery tables
 - First **10 GB** of storage per month are free. **\$0.02 per GB** thereafter (e.g. store VCFs, MAFs, tab-delimited files)
 - **Analysis:**
 - Use standard SQL to query large -omics data, the first **TB** of query data is free a month. **\$5.00 per TB** of queries thereafter.
 - Preview or interrogate data without worrying about downloading data file by file
 - Seamlessly integrate BigQuery tables with commonly used data analysis tools including R and Jupyter notebooks

Attributes of Google BigQuery that make it ideal for use in research

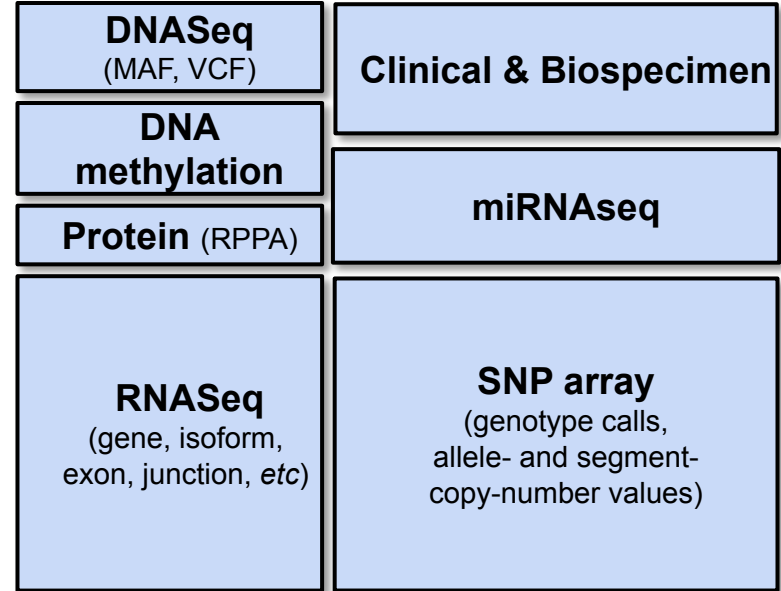
- Columnar database ideal for storing tabular data
- Query speed is automatically scaled by multiprocessing
- Powerful SQL language interface, including user defined functions
- Can join tables based on shared variables



Tree architecture of Dremel

ISB-CGC leverages Google BigQuery to improve accessibility of GDC -omics data

- >500,000 files for TCGA data alone are hosted by the GDC
- ISB-CGC combines data of a similar type into single BigQuery tables
 - For example: ~150 individual MAF files were combined to generate a single table
- Aggregate tables can be queried cheaply and quickly on the Google Cloud

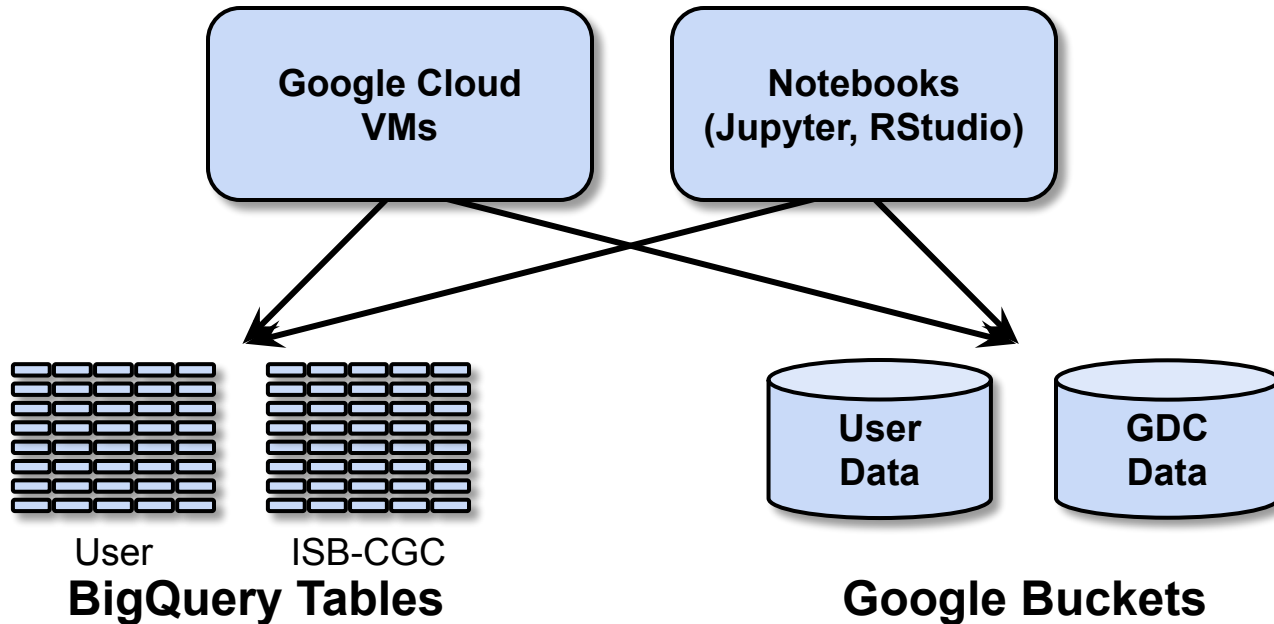


Over 300 open access BigQuery tables hosted by ISB-CGC

- Derived (analyzed) molecular datasets (**TCGA, TARGET, CCLE**)
 - Expression (RNA, protein), copy number, mutations, methylation, clinical, etc.
- Genomic reference tables
 - **PanCancer Atlas, COSMIC, ClinVar, cytoBand, dbSNP, Kaviar, Ensembl, Reactome, Gene Ontology, etc.**
- Metadata tables
 - Indexes of files, Google Cloud file paths, case ID, etc.

Multiple easy avenues for computing on data on ISB-CGC

ISB-CGC enables full command line access to analyze cloud hosted data via a collection of powerful tools and technologies along with the ability to install your own tools



Some example use-cases of the three entry points to ISB-CGC

Interactive web-based exploration

- Select a subset of TCGA samples based on clinical or molecular characteristics
- Compare one cohort to another
- Upload a small private dataset to analyze in conjunction with TCGA data
- *etc...*

Direct Command line Access to VMs

- Test new algorithm on hundreds or thousands of BAM or FASTQ files
- Run novel image segmentation method across whole-slide images
- *etc...*

Interactive big data exploration and analysis

- Interactive data exploration in BigQuery
- Use R or Python to perform custom multivariate analyses
- Develop and customize bioinformatics tools and pipelines
- *etc...*

We provide data exploration tools through our web app

ISB-CGC web tools

Google BigQuery

The screenshot shows the 'Create Cohort - Filters' interface. At the top, there are navigation tabs: DASHBOARD, WORKBOOKS, PROGRAMS, ANALYSES, GENES & miRNAs, VARIABLES, and COHORTS. Below the navigation, the page title is 'Create Cohort - Filters' with a 'Save As New Cohort' button. The main content area is divided into four sections: TCGA DATA, CCLE DATA, TARGET DATA, and USER DATA. The TCGA DATA section is active and contains a list of filters: PROGRAM, PROJECT SHORT NAME, DISEASE CODE (with checkboxes for BRCA (3366), LUAD (1781), UCEC (1637), KIRC (1615), GBM (1573), and HNSC (1573)), and VITAL STATUS. The right side of the interface shows 'Selected Filters' with a 'Clear All' button, 'Program Details' (Total Number of Cases: 11353, Total Number of Samples: 33460), and 'Clinical Features' with five small bar charts for Disease Code, Vital Status, Sample Type, Tumor Tissue Site, and Gender.

The screenshot shows the Google BigQuery interface. It includes a search bar, a 'Query history' section, and a 'Resources' section with a '+ ADD DATA' button. The main area shows a search for 'cgc-05-0050' and a list of results, including 'isb-cgc'.

The screenshot shows a terminal window with the following output:

```
Connected, host fingerprint: ssh-rsa
Linux 4.9.0-11-x86_64

The programs included with the deb:
the exact distribution terms for each
individual files in /usr/share/doc.

Debian GNU/Linux comes with ABSOLUTELY
permitted by applicable law.
Last login: Thu Feb 6 22:31:46 20
Copying g3://genome-public-data/
[1 files] 2.9 GiB/ 2.9 GiB
Operation completed over 1 objects/2.9 GiB.
Feb 06 22:33:26 ... started STAR run
Feb 06 22:33:26 ... starting to generate Genome files
Feb 06 22:34:46 ... starting to sort Suffix Array. This may take a long time...
Feb 06 22:35:00 ... sorting Suffix Array chunks and saving them to disk...
```

Google VMs

Building cohorts using the ISB-CGC web app

The image displays two overlapping screenshots of the ISB-CGC web application interface, illustrating the process of building a cohort.

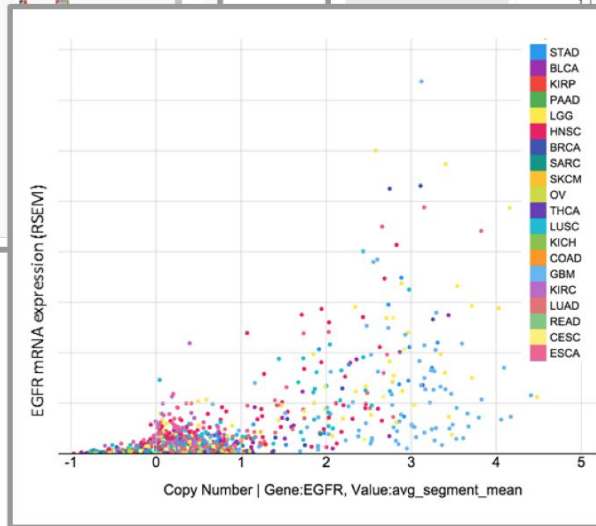
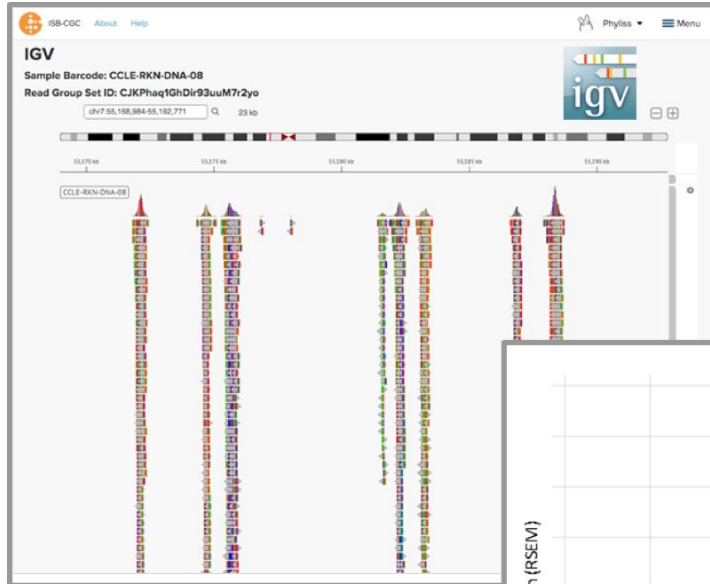
Top Screenshot (Initial State):

- Navigation:** TCGA DATA, CCLE DATA, TARGET DATA, USER DATA.
- Filters:** Selected Filters (Clear All).
- Program Details:** Total Number of Cases: 11,353; Total Number of Samples: 33,460.
- Clinical Features:** Disease Code (Heatmap).
- Left Panel (Filters):**
 - PROGRAM: PROJECT SHORT NAME
 - TCGA-BRCA (3,366 sample(s))
 - TCGA-LUAD (1,781 sample(s))
 - TCGA-UCEC (1,637 sample(s))
 - TCGA-KIRC (1,615 sample(s))
 - TCGA-GBM (1,573 sample(s))
 - TCGA-HNSC (1,573 sample(s))
 - DISEASE CODE
 - VITAL STATUS
 - GENDER
 - Female (12,175 sample(s))
 - Male (11,255 sample(s))
 - NA (10,030 sample(s))
 - AGE AT DIAGNOSIS
 - 10 to 39 (2,426 sample(s))
 - 40 to 49 (3,081 sample(s))
 - 50 to 59 (5,409 sample(s))

Bottom Screenshot (Filtered State):

- Navigation:** TCGA DATA, CCLE DATA, TARGET DATA, USER DATA.
- Filters:** Selected Filters (Clear All): Gender: Female x, Project Short Name: TCGA-BRCA x.
- Program Details:** Total Number of Cases: 1,085; Total Number of Samples: 2,269.
- Clinical Features:** Disease Code, Vital Status, Sample Type, Tumor Tissue Site, Gender (Heatmaps).
- Left Panel (Filters):**
 - PROGRAM: PROJECT SHORT NAME
 - TCGA-BRCA (2,269 sample(s))
 - TCGA-LUAD (635 sample(s))
 - TCGA-UCEC (1,117 sample(s))
 - TCGA-KIRC (390 sample(s))
 - TCGA-GBM (445 sample(s))
 - TCGA-HNSC (306 sample(s))
 - DISEASE CODE
 - VITAL STATUS
 - GENDER
 - Female (2,269 sample(s))
 - Male (24 sample(s))
 - NA (1,073 sample(s))
 - AGE AT DIAGNOSIS
 - 10 to 39 (181 sample(s))
 - 40 to 49 (467 sample(s))
 - 50 to 59 (621 sample(s))

ISB-CGC: Interactive Apps



Integrated visualization methods for Big Data

Integrated genome viewer
(view read pile-ups)

caMicroscope
(view histology)

OHIF
(view radiology)

All Files IGV Pathology Images Pathology Reports Radiology Images

Build
HG19

- ▶ CASE
- ▶ DATA TYPE
- ▶ DATA CATEGORY
- ▶ EXPERIMENTAL STRATEGY
- ▶ DATA FORMAT
- ▶ PLATFORM
- ▶ DISEASE CODE

File Listing

Showing 1 to 25 of 39692 entries

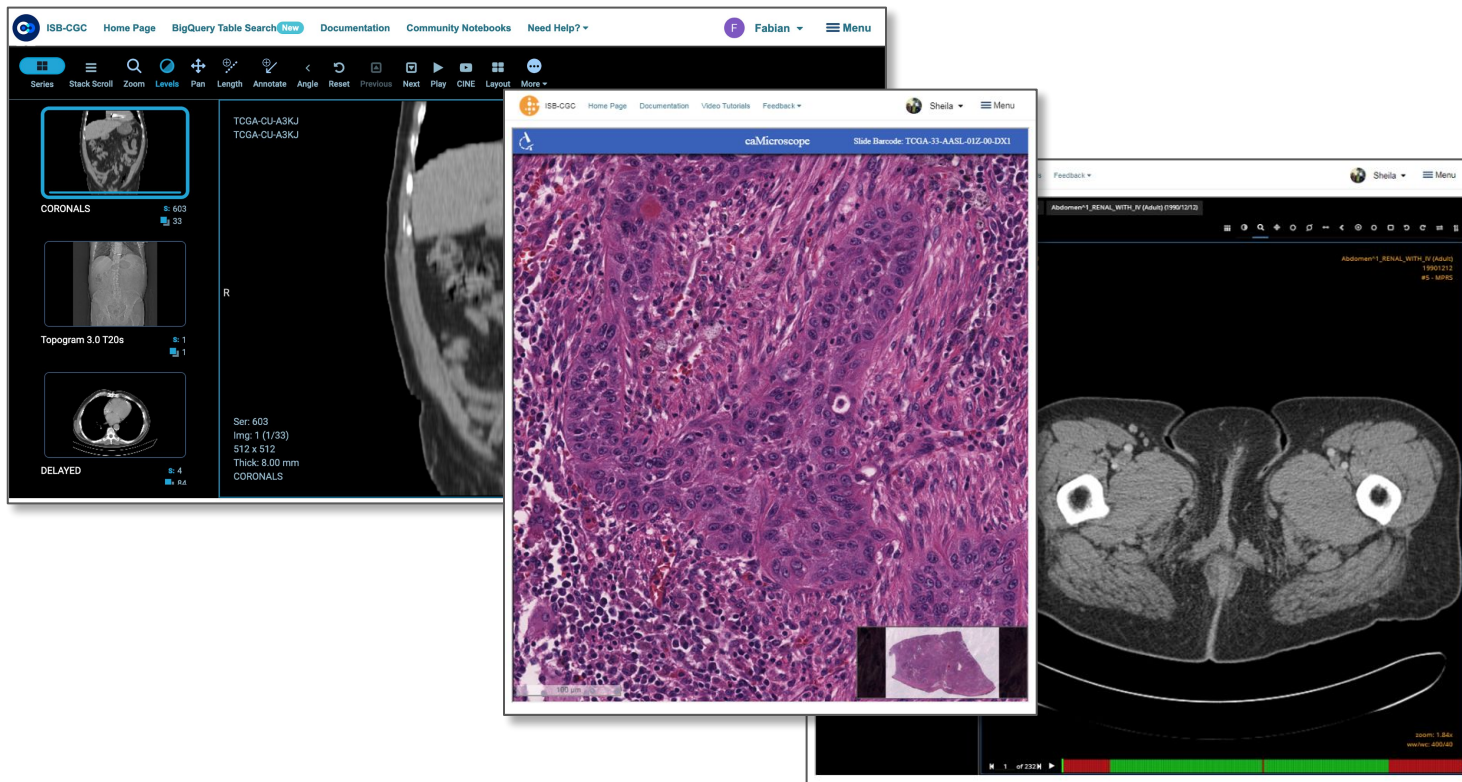
Show 25 entries Page Go Previous 1 2 3 ... 1588 Next

Choose Columns to Display

Program	Case Barcode	File Name	Disease Code	Exp. Strategy	Platform	Data Category	Data Type	Data Format	File Size
TCGA	TCGA-OL-A660	SWEDE_p_TCGAb322_2... [GDC ID: 0686da7c-d103-...	BRCA	Genotyping array	Affymetrix SNP Array 6.0	Simple nucleotide variation	Genotypes	TXT	20.9 MB
TCGA	TCGA-OL-A660	SWEDE_p_TCGAb322_2... [GDC ID: 4bd19f77-9aa7-4...	BRCA	Genotyping array	Affymetrix SNP Array 6.0	Simple nucleotide variation	Genotypes	TXT	20.9 MB
TCGA	TCGA-OL-A660	UNCID_2171596.c7f5714... [GDC ID: b677ea35-d758-...	BRCA	RNA-Seq	Illumina HiSeq	Raw sequencing data	Aligned reads	BAM	7.8 GB
TCGA	TCGA-OL-A660	c61047b5e4ae38963735fc... [GDC ID: 0a6db03e-748a-...	BRCA	WXS	Illumina HiSeq	Raw sequencing data	Aligned reads	BAM	4.9 GB
TCGA	TCGA-OL-A660	256cd674e76be0f163766b... [GDC ID: 72a31a7e-99df-4...	BRCA	WXS	Illumina HiSeq	Raw sequencing	Aligned reads	BAM	7.2 GB

CSV BigQuery GCS

ISB-CGC: Interactive image viewers



The ISB-CGC BigQuery Table Search UI

BigQuery Table Search

Explore and learn more about available ISB-CGC BigQuery tables with this search feature.
Find tables of interest based on category, reference genome build, data type and free-form text search.

ISB-CGC BigQuery Documentation [↗](#) ISB-CGC BigQuery Access Info [↗](#) Google BigQuery Console [↗](#) About BigQuery [↗](#) Release Notes [↗](#)

Status:

Name:

Program:

Category:

- CLINICAL BIOSPECIMEN DATA [🔗](#)
- FILE METADATA [🔗](#)
- GENOMIC REFERENCE DATABASE [🔗](#)
- PROCESSED -OMICS DATA [🔗](#)

Reference Genome:

Source:

Data Type:

Experimental Strategy:

[Reset All Filters](#)

[+ Show More Filters](#)

Show entries

[Columns](#) [CSV Download](#) Search:

Name	Program	Category	Source	Data Type	Status	Rows	Created	Preview	Open
CCLE 2016 - AFFYU133 MICROARRAY	CCLE	PROCESSED -OMICS DATA	BROAD	GENE EXPRESSION	CURRENT	17,525,476	2/26/2016	Preview	Open
CCLE 2016 - COPY NUMBER SEGMENTS	CCLE	PROCESSED -OMICS DATA	BROAD	COPY NUMBER SEGMENT	CURRENT	760,192	2/27/2016	Preview	Open
CCLE 2016 - FASTQC METRICS	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,249	3/28/2016	Preview	Open
CCLE 2016 - FILE METADATA	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,915	3/29/2016	Preview	Open
CCLE 2016 - SAMPLE INFORMATION	CCLE	PROCESSED -OMICS DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	929	2/26/2016	Preview	Open
CCLE 2016 - SOMATIC MUTATION	CCLE	PROCESSED -OMICS DATA	BROAD	SOMATIC MUTATIONS	CURRENT	116,708	2/26/2016	Preview	Open
CCLE BIOSPECIMEN V0	CCLE	CLINICAL BIOSPECIMEN DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	954	4/4/2019	Preview	Open
CCLE CLINICAL V1	CCLE	CLINICAL BIOSPECIMEN DATA	BROAD	CLINICAL DATA	CURRENT	950	6/21/2019	Preview	Open
CCLE HG19 METADATA RELEASE 14	CCLE	FILE METADATA	BROAD	FILE METADATA	CURRENT	1,273	3/7/2019	Preview	Open
CLINVAR 20180401 GRCH37		GENOMIC REFERENCE DATABASE	CLINVAR	SOMATIC MUTATIONS	CURRENT	354,471	4/17/2018	Preview	Open

Showing 1 to 10 of 214 entries (filtered from 327 total entries)

[Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) ... [22](#) [Next](#)

Have feedback or corrections? Please email us at feedback@isb-cgc.org.

More information on a table at the click of a button!

BigQuery Table Search

Explore and learn more about available ISB-CGC BigQuery tables with this search feature.
Find tables of interest based on category, reference genome build, data type and free-form text search.

ISB-CGC BigQuery Documentation | ISB-CGC BigQuery Access Info | Google BigQuery Console | About BigQuery | Release Notes

Status: CURRENT

Name:

Program:

Category:
 CLINICAL BIOSPECIMEN DATA
 FILE METADATA
 GENOMIC REFERENCE DATABASE
 PROCESSED -OMIC DATA

Reference Genome: ALL

Source:

Data Type:

Experimental Strategy:

[Reset All Filters](#)

[+ Show More Filters](#)

Show 10 entries

Columns | CSV Download | Search:

Name	Program	Category	Source	Data Type	Status	Rows	Created	Preview	Open
CCLE 2016 - AFFYU133 MICROARRAY	CCLE	PROCESSED -OMIC DATA	BROAD	GENE EXPRESSION	CURRENT	17,525,476	2/26/2016		
CCLE 2016 - COPY NUMBER SEGMENTS	CCLE	PROCESSED -OMIC DATA	BROAD	COPY NUMBER SEGMENT	CURRENT	760,192	2/27/2016		

Full ID: [isb-cgc.ccle_201602_alpha.Copy_Number_Segments](#) | COPY | OPEN

Dataset ID: [ccle_201602_alpha](#)

Table ID: [Copy_Number_Segments](#)

Description: Data was extracted from an older CCLE dataset from Google Genomics on February 2016. Copy number segment data are made available here.

Schema:

Field Name	Type	Mode	Description
CCLE_name	STRING	NULLABLE	Cell line primary name, appended with a short name for the location of the cancer: e.g. TC71_BONE_HUPT4_PANCREAS, etc
Cell_line_primary_name	STRING	NULLABLE	The cell line primary name: e.g. TC71, NCI-60, etc
Platform	STRING	NULLABLE	Platform used to generate these data (Genome_Wide_SNP_6)
Chromosome	STRING	NULLABLE	Chromosome, possible values: chr1-22, and chrX
Start	INTEGER	NULLABLE	Start position
End	INTEGER	NULLABLE	End position
Num_Probes	INTEGER	NULLABLE	The num_probes field specifies the number of probes on the SNP chip that went into estimating the mean copy number for this segment
Segment_Mean	FLOAT	NULLABLE	Provides the log2(CN2) mean value estimate

Labels: [access: open](#) [data_type: copy_number_segment](#) [program: ccle](#) [reference_genome: 0_hg18](#) [source: broad](#) [category: processed_omics_data](#) [status: current](#)

Name	Program	Category	Source	Data Type	Status	Rows	Created	Preview	Open
CCLE 2016 - FASTQC METRICS	CCLE	PROCESSED -OMIC DATA	BROAD	FILE METADATA	CURRENT	1,249	3/28/2016		
CCLE 2016 - FILE METADATA	CCLE	PROCESSED -OMIC DATA	BROAD	FILE METADATA	CURRENT	1,915	3/29/2016		
CCLE 2016 - SAMPLE INFORMATION	CCLE	PROCESSED -OMIC DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	929	2/26/2016		
CCLE 2016 - SOMATIC MUTATION	CCLE	PROCESSED -OMIC DATA	BROAD	SOMATIC MUTATIONS	CURRENT	116,708	2/26/2016		
CCLE BIOSPECIMEN V0	CCLE	CLINICAL BIOSPECIMEN DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	954	4/4/2019		
CCLE CLINICAL V1	CCLE	CLINICAL BIOSPECIMEN DATA	BROAD	CLINICAL DATA	CURRENT	950	6/21/2019		
CCLE HG19 METADATA RELEASE 14	CCLE	FILE METADATA	BROAD	FILE METADATA	CURRENT	1,273	3/7/2019		
CLINVAR 20180401 GRCH37		GENOMIC REFERENCE DATABASE	CLINVAR	SOMATIC MUTATIONS	CURRENT	354,471	4/17/2018		




Showing 1 to 10 of 214 entries (filtered from 327 total entries)

Previous | 1 | 2 | 3 | 4 | 5 | ... | 22 | Next

Have feedback or corrections? Please email us at feedback@isb-cgc.org.

Benefits of the ISB-CGC BigQuery Table Search

- No login required!
- Allows users to browse and learn more about available ISB-CGC BigQuery tables
- Each table has been curated to include detailed table and field descriptions as well as table labels
- Identify table(s) of interest by filtering (e.g. by reference genome build, data type, category) or via free-form text search
- Get a snapshot of table contents by previewing the first few (~10) lines
- Found a table you're interested in? Simply click on the "open" button to jump directly to the GCP BigQuery Console.

	CCLE CLINICAL V1	CCLC	CLINICAL BIOSPECIMEN DATA	BROAD	CLINICAL DATA	CURRENT	950	6/21/2019		
--	------------------	------	---------------------------------	-------	---------------	---------	-----	-----------	---	---

Mitelman database available through ISB-CGC

Manually curated open access database with critical information about chromosome aberrations and gene fusions in cancer. These data are also available through BigQuery.

Home

Search

Cases Cytogenetics

Gene Fusions

Clinical Associations

Recurrent Chromosome Aberrations

References

User Guide

About

Contact

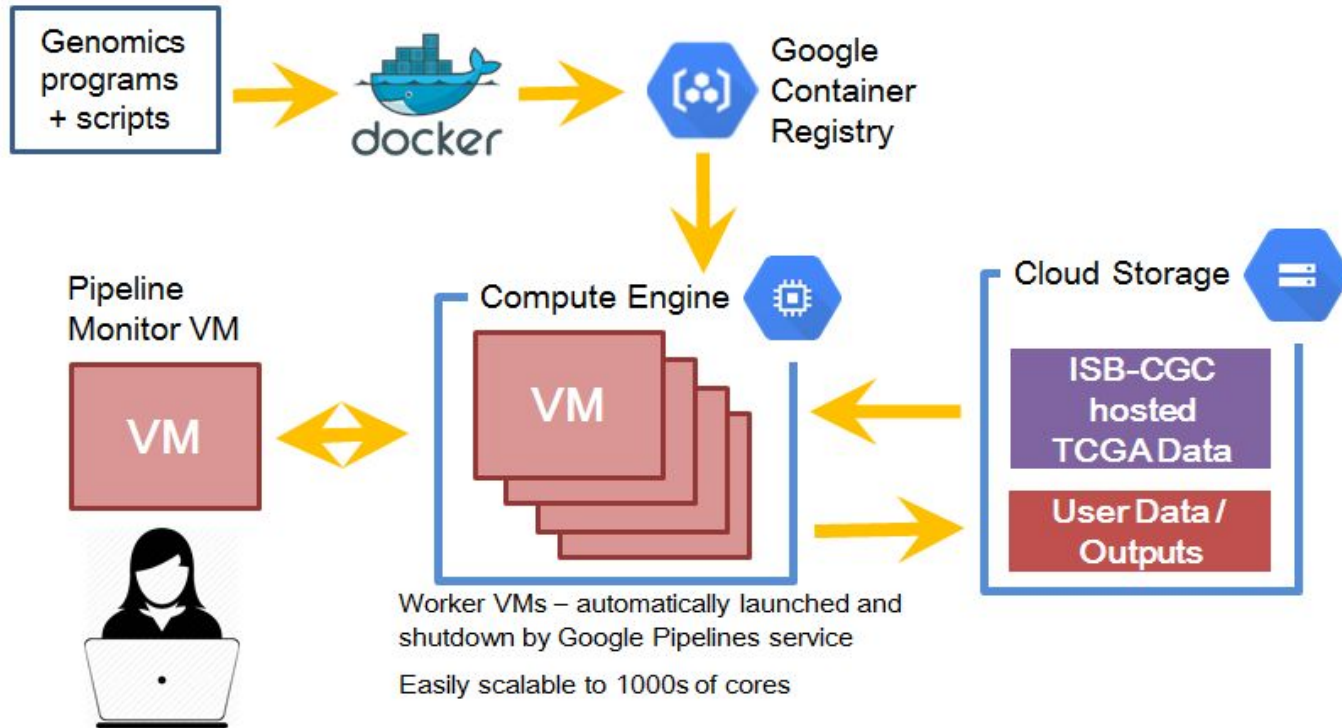
Mitelman Database
Chromosome Aberrations and Gene Fusions in Cancer

This site has been funded by:
 National Cancer Institute
 Swedish Cancer Society
 Swedish Childhood Cancer Foundation

This website is built and maintained by the ISB-CGC cloud project. * Photo credits: JJ Ying on Unsplash



VMs enable advanced bioinformatic workflows



Some workflows we've enabled for ISB-CGC end-users

Multiple PanCancer Atlas projects, including:

- Germline-variant calling
- Fusion gene analysis
- T-cell and B-cell receptor analysis
- viral DNA screening
- MYC pathway analysis (BQ)
- 8-oxoG filtering (MC3 project)

Other end-user projects include:

- SMC-RNA Dream challenge (supporting both the organizers and many participants)
- tumor-specific alternative polyadenylation
- ML algorithm evaluation & benchmarking
- RNA seq alignment to novel transcriptome(s)
- mRNA expression quantitation
- targeted de-novo assembly
- structural variations (WGS + SNP6 data)
- metagenomics / cancer analysis
- statistical meta-analysis of miRNAs in cancer
- code/tutorial development
- GDC hg38 TCGA miRNA QC (w/ BCGSC)



Sci Rep. 2016; 6: 39259.
Published online 2016 Dec 16. doi: 10.1038/srep39259

PMCID: PMC5158871

A cloud-based workflow to quantify transcript-expression levels in public cancer compendia

P.J. Tatlow¹ and Stephen R. Piccolo^{1,2}



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

Pan-cancer analysis reveals complex tumor-specific alternative polyadenylation

Human Mutation
Variation, Informatics, and Disease



Explore this journal >

INFORMATICS

Detection of homozygous deletions in tumor-suppressor genes ranging from dozen to hundreds nucleotides in cancer models

Lun-Ching Chang, Suleyman Vural, Dmitriy Sonkin

First published: 23 August 2017 Full publication history

DOI: 10.1002/humu.23308 View/save citation

rrren, Ewan A. Gibb, Daniel MacMillan, Johnathan Wong, Readman
ammond, Catherine A. Ennis, Abigail Hahn, Sheila Reynolds, Inanc

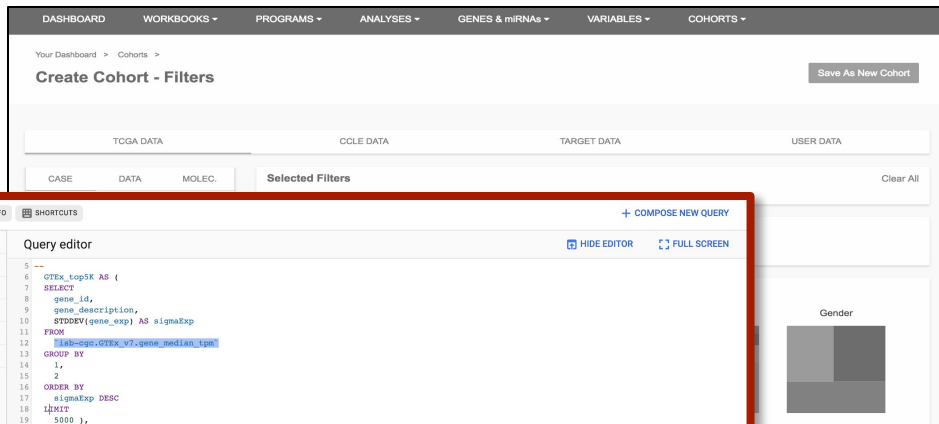
101/160960

nd has not been peer-reviewed [what does this mean?].

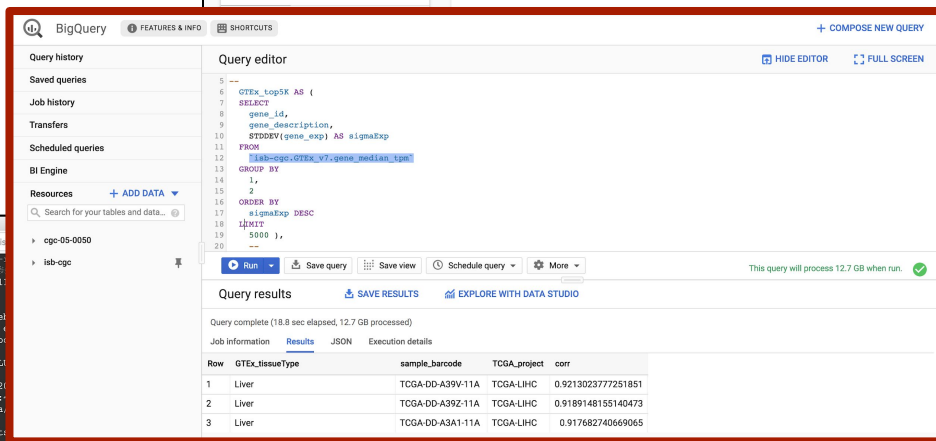
*with many other manuscripts
and grants currently in
progress or submitted*

Three entry points for exploring cancer data on ISB-CGC

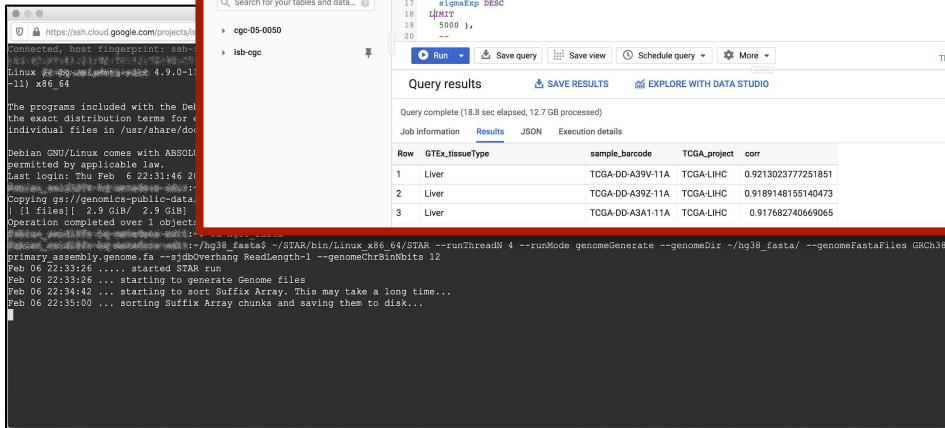
ISB-CGC WebApp



Google BigQuery



Google VMs



BigQuery integrates with a variety of commonly used analysis tools



bigquery and
bigQueryR



googleAuthR



Pre-built VM images

IP[y]:

IPython



Cloud notebooks
and workspaces.

Cloud Datalab



Analyze correlation between TCGA samples & GTEx tissue types quickly and cheaply!

```
5 --
6 GTEX_top5K AS (
7 SELECT
8   gene_id,
9   gene_description,
10  STDDEV(gene_exp) AS sigmaExp
11 FROM
12   `isb-cgc.GTEX_v7.gene_median_tpm`
13 GROUP BY
14   1,
15   2
16 ORDER BY
17   sigmaExp DESC
18 LIMIT
19   5000 ),
20 --
```

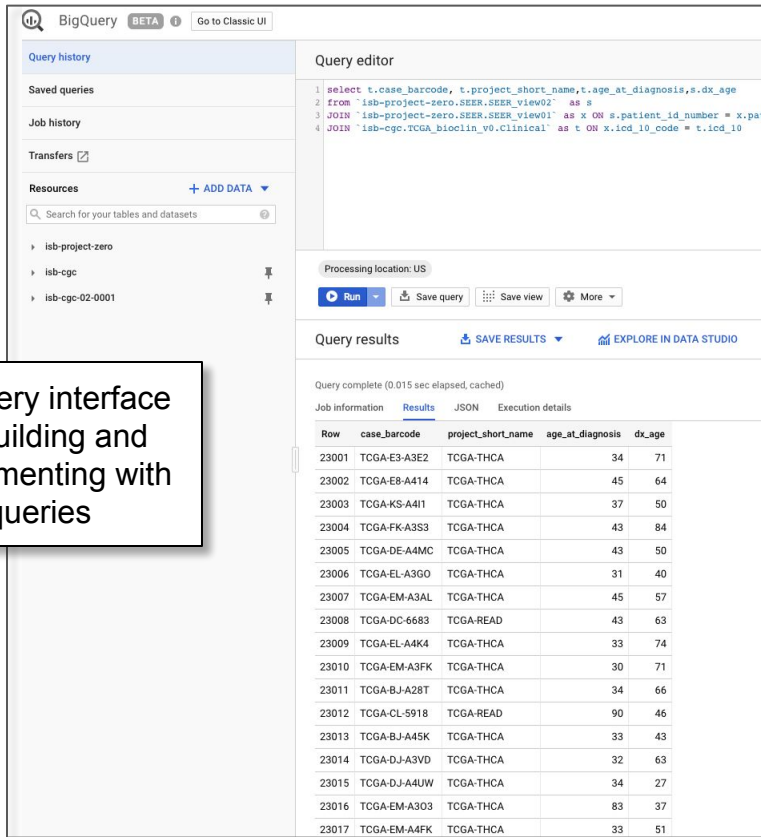
Query results

Query complete (18.8 sec elapsed, 12.7 GB processed)

Job information Results JSON Execution details

Row	GTEx_tissueType	sample_barcode	TCGA_project	corr
1	Liver	TCGA-DD-A39V-11A	TCGA-LIHC	0.9213023777251851
2	Liver	TCGA-DD-A39Z-11A	TCGA-LIHC	0.9189148155140473
3	Liver	TCGA-DD-A3A1-11A	TCGA-LIHC	0.917682740669065

Tables can be joined in BigQuery using SQL to draw connections amongst data



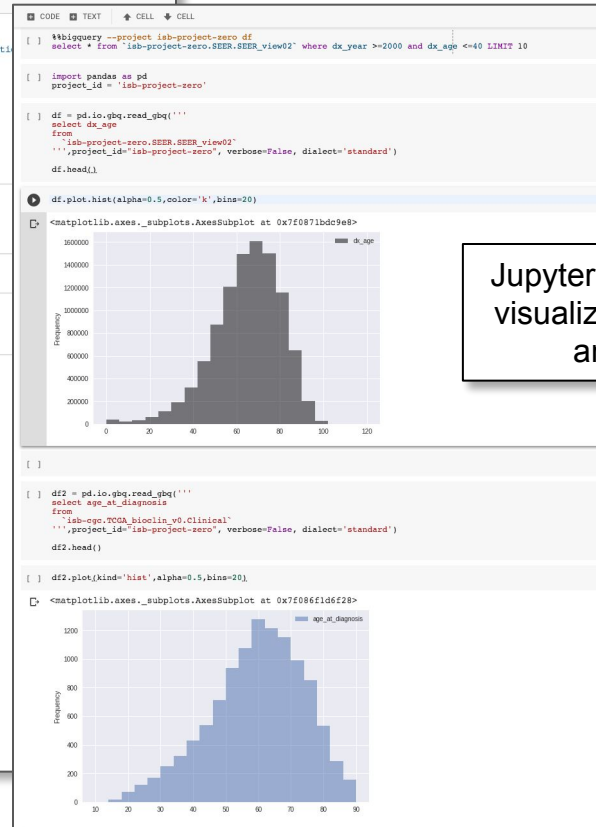
The screenshot shows the BigQuery interface. On the left, there is a sidebar with 'Query history', 'Saved queries', 'Job history', 'Transfers', and 'Resources'. The main area is the 'Query editor' with a SQL query:

```
1 select t.case_barcode, t.project_short_name, t.age_at_diagnosis, s.dx_age
2 from `isb-project-zero.SEER_SEER_view02` as s
3 JOIN `isb-project-zero.SEER_SEER_view01` as x ON s.patient_id_number = x.pat
4 JOIN `isb-cgc.TCGA_bioclin_v0.clinical` as t ON x.lcd_10_code = t.lcd_10
```

Below the editor, there are buttons for 'Run', 'Save query', 'Save view', and 'More'. The 'Query results' section shows a table with 17 rows and 5 columns: Row, case_barcode, project_short_name, age_at_diagnosis, and dx_age.

Row	case_barcode	project_short_name	age_at_diagnosis	dx_age
23001	TCGA-E3-A3E2	TCGA-THCA	34	71
23002	TCGA-E8-A414	TCGA-THCA	45	64
23003	TCGA-KS-A411	TCGA-THCA	37	50
23004	TCGA-FK-A3S3	TCGA-THCA	43	84
23005	TCGA-DE-A4MC	TCGA-THCA	43	50
23006	TCGA-EL-A3G0	TCGA-THCA	31	40
23007	TCGA-EM-A3AL	TCGA-THCA	45	57
23008	TCGA-DC-6683	TCGA-READ	43	63
23009	TCGA-EL-A4K4	TCGA-THCA	33	74
23010	TCGA-EM-A3FK	TCGA-THCA	30	71
23011	TCGA-BJ-A28T	TCGA-THCA	34	66
23012	TCGA-CL-5918	TCGA-READ	90	46
23013	TCGA-BJ-A45K	TCGA-THCA	33	43
23014	TCGA-DJ-A3VD	TCGA-THCA	32	63
23015	TCGA-DJ-A4UW	TCGA-THCA	34	27
23016	TCGA-EM-A3O3	TCGA-THCA	83	37
23017	TCGA-EM-A4FK	TCGA-THCA	33	51

BigQuery interface for building and experimenting with queries



The screenshot shows a Jupyter notebook with two cells. The first cell contains SQL code to read data from BigQuery and a histogram of 'dx_age'.

```
%%bigquery --project isb-project-zero df
select * from `isb-project-zero.SEER_SEER_view02` where dx_year >= 2000 and dx_age <= 40 LIMIT 10

import pandas as pd
project_id = 'isb-project-zero'

df = pd.io.gbig.read_gbq("""
select dx_age
from
  isb-project-zero.SEER_SEER_view02
  ...project_id='isb-project-zero', verbose=False, dialect='standard'
""")
df.head()
```

The second cell contains more SQL code and a histogram of 'age_at_diagnosis'.

```
df2 = pd.io.gbig.read_gbq("""
select age_at_diagnosis
from
  isb-cgc.TCGA_bioclin_v0.clinical
  ...project_id='isb-project-zero', verbose=False, dialect='standard'
""")
df2.head()

df2.plot(kind='hist', alpha=0.5, bins=20)
```

Both histograms show frequency on the y-axis and age on the x-axis. The first histogram is for 'dx_age' and the second is for 'age_at_diagnosis'.

Jupyter notebook to visualize and share analysis

Use Google BigQuery to easily connect your research to public datasets

ISB-CGC and Other
Public Datasets



Private User Data
and Derived Results

A typical work setup across multiple browser tabs

Google web interface

The screenshot shows the Google Cloud Platform BigQuery interface. On the left is a navigation sidebar with sections like 'Query history', 'Saved queries', 'Job history', 'Taskruns', 'Scheduled queries', 'BI Engine', and 'Resources'. The main area is the 'Query editor' for a query named 'kirc_summed_sc_join'. Below the editor is a table of results with columns: Row, Short name, Sample barcode, Tissue, Topo gene name, HTSeq_FPKM_UQ, Topo gene FPKM, Cell ID, Sc gene name, summed values, and Sc gene sum. The table contains 15 rows of data.

Row	Short name	Sample barcode	Tissue	Topo gene name	HTSeq_FPKM_UQ	Topo gene FPKM	Cell ID	Sc gene name	summed values	Sc gene sum
1	TCGA-KIRC	TCGA-MMA-A50-D1A	BCCND1D	3442.526458	8212	8035.34	BCCND1D	0.0	1	1
2	TCGA-KIRC	TCGA-MMA-A50-D1A	COTL1	84341.73275	18937	8035.34	COTL1	0.0	1	1
3	TCGA-KIRC	TCGA-MMA-A50-D1A	IFBID3	49049.918242	13297	8035.34	IFBID3	0.0	1	1
4	TCGA-KIRC	TCGA-MMA-A50-D1A	C17orf18	12648.309455	12480	8035.34	C17orf18	0.0	1	1
5	TCGA-KIRC	TCGA-MMA-A50-D1A	EBAG9	14768.72216	13000	8035.34	EBAG9	0.0	1	1
6	TCGA-KIRC	TCGA-MMA-A50-D1A	GLEC2L	209.1578853	2581	8035.34	GLEC2L	0.0	1	1
7	TCGA-KIRC	TCGA-MMA-A50-D1A	RP11-139G13.13	0.0	1	8035.34	RP11-139G13.13	0.0	1	1
8	TCGA-KIRC	TCGA-MMA-A50-D1A	SNRPD	385.20341577	2917	8035.34	SNRPD	0.0	1	1
9	TCGA-KIRC	TCGA-MMA-A50-D1A	C17orf17	11286.5074348	6228	8035.34	C17orf17	0.0	1	1
10	TCGA-KIRC	TCGA-MMA-A50-D1A	MAGEC1	0.0	1	8035.34	MAGEC1	0.0	1	1
11	TCGA-KIRC	TCGA-MMA-A50-D1A	CD5L	4688.8778437	5029	8035.34	CD5L	0.0	1	1
12	TCGA-KIRC	TCGA-MMA-A50-D1A	ECTZL	485.82399418	3411	8035.34	ECTZL	0.08	14017	14025
13	TCGA-KIRC	TCGA-MMA-A50-D1A	CTSL	102356.60719	1919	8035.34	CTSL	2.7	14623	14630
14	TCGA-KIRC	TCGA-MMA-A50-D1A	ZNF554	35685.889472	8292	8035.34	ZNF554	19.13	16279	16286
15	TCGA-KIRC	TCGA-MMA-A50-D1A	CDK6RAP1	18835.512716	12887	8035.34	CDK6RAP1	151.0	18890	18897

Built in syntax checking

Notebook (R or Python)

The screenshot shows a Jupyter Notebook interface with a 'SETUP' section and a 'BigQueries!' section. The code in the notebook uses BigQuery's runQuery function to execute a query that counts samples by type. The output shows the results of the query, including the number of samples for each type.

```
runQuery(' bqClient, sql, dryRun=False )
res0
```

```
runQuery(' bqClient, sql, dryRun=False )
res0
```

```
runQuery(' bqClient, sql, dryRun=False )
res0
```

Integrate with notebooks to generate your own publication quality visuals

Searchable web docs

The screenshot shows the Google BigQuery documentation page. The page title is 'Google BigQuery documentation'. Below the title is a description of BigQuery as Google's fully managed, petabyte scale, low cost analytics data warehouse. There are several navigation cards for 'Quickstarts', 'How-to guides', 'APIs & reference', 'Concepts', 'Tutorials', and 'Resources'. The 'Quickstarts' card says 'Learn in 5 minutes'. The 'How-to guides' card says 'Perform specific tasks'. The 'APIs & reference' card says 'API, web UI, and command-line'. The 'Concepts' card says 'Develop a deep understanding of BigQuery'. The 'Tutorials' card says 'Walkthroughs of common applications'. The 'Resources' card says 'Pricing, quotas, release notes, and other resources'.

What you need to know to interoperate with ISB-CGC

- Thin layer on top of Google Cloud Platform - full access to all Google tools and technologies
 - Can run any type of workflow
 - Come in with own GCP or AWS
 - ISB-CGC APIs + any and all Google APIs
- Authentication & Authorization (A&A) once using *service accounts*
- Store and compute on data in BigQuery
 - BigQuery metadata tables of manifests of GDC data (find out URLs for files to compute on)
 - Compiled Derived data in BQ (including reference tables)
 - No waiting in queue
 - Access to sudo in your VMs
 - Highly scalable in cores and RAM - use only what you need
 - Data backups automatically managed
 - Easily manage access to your data by other groups

Managing security and permissions via service accounts

- Service Account is the Authentication and Authorization method that researchers' computers run under, works for all members of Google Cloud Project (shared "Workspace")
- Applications assume the identity of the service account to call Google APIs, so that the users aren't directly involved
- ISB-CGC users create a Google Cloud Platform (GCP) project that comes automatically configured with a "Compute Engine default service account"
- Users must register their service accounts with ISB-CGC to access controlled-data
- Service accounts allow management of controlled data in
 - Files
 - Directories (even mimicked in object storage)
 - Data Structures
- Researchers with validated Service Account use all Google cloud resources natively and seamlessly, very familiar environment

Google Cloud Platform Free Tier lets you compute without entering a credit card!

<p>DATA ANALYTICS</p> <p>BigQuery</p> <p>1 TB</p> <p>Queries per month</p> <p>Fully managed, petabyte scale, analytics data warehouse.</p> <hr/> <p>1 TB of querying per month</p> <hr/> <p>10 GB of storage</p> <p>↑</p>	<table border="1"><tr><td data-bbox="639 238 852 436"><p>COMPUTE</p><p>Cloud Run</p><p>2 million</p><p>Requests per month</p><p>A fully managed environment to run stateless containers.</p><p>↓</p></td><td data-bbox="852 238 1064 436"><p>DATABASE</p><p>Firestore</p><p>1 GB</p><p>Storage</p><p>Scalable NoSQL, document database.</p><p>↓</p></td><td data-bbox="1064 238 1290 436"><p>COMPUTE</p><p>Compute Engine</p><p>1</p><p>F1-micro instance per month</p><p>Scalable, high-performance virtual machines.</p><p>↓</p></td></tr><tr><td data-bbox="639 436 852 635"><p>STORAGE</p><p>Cloud Storage</p><p>5 GB</p><p>Months regional storage</p><p>Best-in-class performance, reliability, and pricing for all your storage needs.</p><p>↓</p></td><td data-bbox="852 436 1064 635"><p>DATA ANALYTICS</p><p>Pub/Sub</p><p>10 GB</p><p>Messages per month</p><p>A global service for real-time and reliable messaging and streaming data.</p><p>↓</p></td><td data-bbox="1064 436 1290 635"><p>COMPUTE</p><p>Cloud Functions</p><p>2 million</p><p>Invocations per month</p><p>A serverless environment to build and connect cloud services with code.</p><p>↓</p></td></tr><tr><td data-bbox="639 635 852 834"><p>COMPUTE</p><p>Google Kubernetes Engine</p><p>Clusters</p><p>All size clusters</p><p>One-click container orchestration via Kubernetes clusters, managed by Google.</p><p>↓</p></td><td data-bbox="852 635 1064 834"><p>COMPUTE</p><p>App Engine</p><p>28</p><p>Instance hours per day</p><p>Platform for building scalable web applications and mobile back ends.</p><p>↓</p></td><td data-bbox="1064 635 1290 834"><p>MANAGEMENT TOOLS</p><p>Stackdriver</p><p>50 GB</p><p>Logs with 30-day retention</p><p>Monitoring, logging, and diagnostics for applications on Google Cloud and AWS.</p><p>↓</p></td></tr><tr><td data-bbox="639 834 852 1048"><p>DATA ANALYTICS</p><p>BigQuery</p><p>1 TB</p><p>Queries per month</p><p>Fully managed, petabyte scale, analytics data warehouse.</p><p>↓</p></td><td data-bbox="852 834 1064 1048"><p>AI AND MACHINE LEARNING</p><p>Vision AI</p><p>1,000</p><p>Units per month</p><p>Label detection, OCR, facial detection and more.</p><p>↓</p></td><td data-bbox="1064 834 1290 1048"><p>AI AND MACHINE LEARNING</p><p>Speech-to-Text</p><p>60</p><p>Minutes per month</p><p>Speech-to-text transcription – the same that powers Google's own products.</p><p>↓</p></td></tr></table>	<p>COMPUTE</p> <p>Cloud Run</p> <p>2 million</p> <p>Requests per month</p> <p>A fully managed environment to run stateless containers.</p> <p>↓</p>	<p>DATABASE</p> <p>Firestore</p> <p>1 GB</p> <p>Storage</p> <p>Scalable NoSQL, document database.</p> <p>↓</p>	<p>COMPUTE</p> <p>Compute Engine</p> <p>1</p> <p>F1-micro instance per month</p> <p>Scalable, high-performance virtual machines.</p> <p>↓</p>	<p>STORAGE</p> <p>Cloud Storage</p> <p>5 GB</p> <p>Months regional storage</p> <p>Best-in-class performance, reliability, and pricing for all your storage needs.</p> <p>↓</p>	<p>DATA ANALYTICS</p> <p>Pub/Sub</p> <p>10 GB</p> <p>Messages per month</p> <p>A global service for real-time and reliable messaging and streaming data.</p> <p>↓</p>	<p>COMPUTE</p> <p>Cloud Functions</p> <p>2 million</p> <p>Invocations per month</p> <p>A serverless environment to build and connect cloud services with code.</p> <p>↓</p>	<p>COMPUTE</p> <p>Google Kubernetes Engine</p> <p>Clusters</p> <p>All size clusters</p> <p>One-click container orchestration via Kubernetes clusters, managed by Google.</p> <p>↓</p>	<p>COMPUTE</p> <p>App Engine</p> <p>28</p> <p>Instance hours per day</p> <p>Platform for building scalable web applications and mobile back ends.</p> <p>↓</p>	<p>MANAGEMENT TOOLS</p> <p>Stackdriver</p> <p>50 GB</p> <p>Logs with 30-day retention</p> <p>Monitoring, logging, and diagnostics for applications on Google Cloud and AWS.</p> <p>↓</p>	<p>DATA ANALYTICS</p> <p>BigQuery</p> <p>1 TB</p> <p>Queries per month</p> <p>Fully managed, petabyte scale, analytics data warehouse.</p> <p>↓</p>	<p>AI AND MACHINE LEARNING</p> <p>Vision AI</p> <p>1,000</p> <p>Units per month</p> <p>Label detection, OCR, facial detection and more.</p> <p>↓</p>	<p>AI AND MACHINE LEARNING</p> <p>Speech-to-Text</p> <p>60</p> <p>Minutes per month</p> <p>Speech-to-text transcription – the same that powers Google's own products.</p> <p>↓</p>	<p>COMPUTE</p> <p>Compute Engine</p> <p>1</p> <p>F1-micro instance per month</p> <p>Scalable, high-performance virtual machines.</p> <hr/> <p>1 f1-micro instance per month (US regions only—excluding Northern Virginia [us-east4])</p> <hr/> <p>30 GB-months HDD</p> <hr/> <p>5 GB-months snapshot in select regions</p> <hr/> <p>1 GB network egress from North America to all region destinations per month (excluding China and Australia)</p> <p>↑</p>
<p>COMPUTE</p> <p>Cloud Run</p> <p>2 million</p> <p>Requests per month</p> <p>A fully managed environment to run stateless containers.</p> <p>↓</p>	<p>DATABASE</p> <p>Firestore</p> <p>1 GB</p> <p>Storage</p> <p>Scalable NoSQL, document database.</p> <p>↓</p>	<p>COMPUTE</p> <p>Compute Engine</p> <p>1</p> <p>F1-micro instance per month</p> <p>Scalable, high-performance virtual machines.</p> <p>↓</p>												
<p>STORAGE</p> <p>Cloud Storage</p> <p>5 GB</p> <p>Months regional storage</p> <p>Best-in-class performance, reliability, and pricing for all your storage needs.</p> <p>↓</p>	<p>DATA ANALYTICS</p> <p>Pub/Sub</p> <p>10 GB</p> <p>Messages per month</p> <p>A global service for real-time and reliable messaging and streaming data.</p> <p>↓</p>	<p>COMPUTE</p> <p>Cloud Functions</p> <p>2 million</p> <p>Invocations per month</p> <p>A serverless environment to build and connect cloud services with code.</p> <p>↓</p>												
<p>COMPUTE</p> <p>Google Kubernetes Engine</p> <p>Clusters</p> <p>All size clusters</p> <p>One-click container orchestration via Kubernetes clusters, managed by Google.</p> <p>↓</p>	<p>COMPUTE</p> <p>App Engine</p> <p>28</p> <p>Instance hours per day</p> <p>Platform for building scalable web applications and mobile back ends.</p> <p>↓</p>	<p>MANAGEMENT TOOLS</p> <p>Stackdriver</p> <p>50 GB</p> <p>Logs with 30-day retention</p> <p>Monitoring, logging, and diagnostics for applications on Google Cloud and AWS.</p> <p>↓</p>												
<p>DATA ANALYTICS</p> <p>BigQuery</p> <p>1 TB</p> <p>Queries per month</p> <p>Fully managed, petabyte scale, analytics data warehouse.</p> <p>↓</p>	<p>AI AND MACHINE LEARNING</p> <p>Vision AI</p> <p>1,000</p> <p>Units per month</p> <p>Label detection, OCR, facial detection and more.</p> <p>↓</p>	<p>AI AND MACHINE LEARNING</p> <p>Speech-to-Text</p> <p>60</p> <p>Minutes per month</p> <p>Speech-to-text transcription – the same that powers Google's own products.</p> <p>↓</p>												

Some example typical ISB-CGC use-cases...

- 1) Fire up VMs to run pipelines using any workflow language of your choice
- 2) Build cohorts on the web-app and download file manifests with locations of files to use for analyses
- 3) The ISB-CGC Gold Standard Use-Case (featured in our demo)
 - a) Use BigQuery to identify useful public data
 - b) Transition to notebooks to perform multivariate analysis
 - c) Leverage public data analysis tools (i.e., bioconductor)
 - d) Combine your own data with public data seamlessly
 - e) Generate beautiful figures

Questions?

ISB-CGC Team



Bill Longabaugh
Suzanne Paquette
David Gibbs
Jennifer Dougherty
Bill Clifford
Elaine Lee
Lauren Hagen
Boris Aguilar
Mi Tian
Lauren Wolfe
Ilya Shmulevich



David Pot
Madelyn Reyes
Kawther Abdilleh
Ron Taylor
Fabian Seidl
Deena Bleich
Mark Backus
Derrick Moore
Owais Shahzada