



Cancer Data Analytics on the ISB- CGC platform

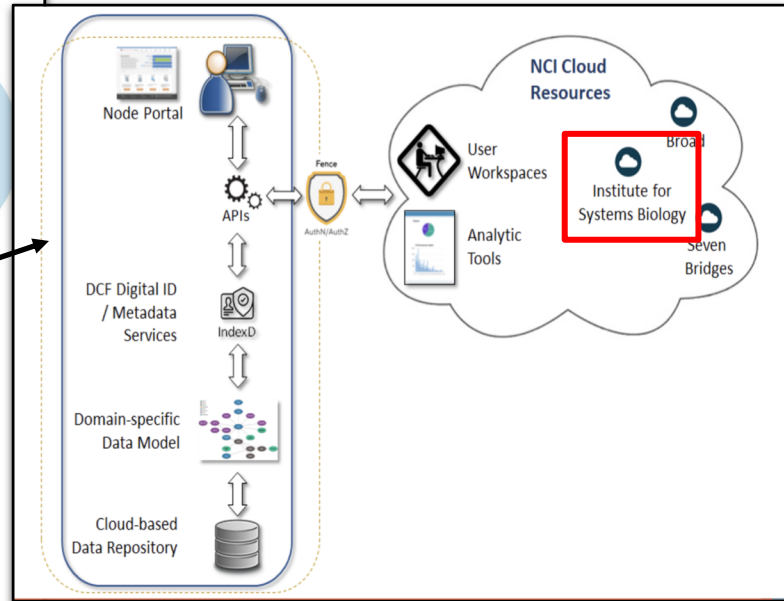
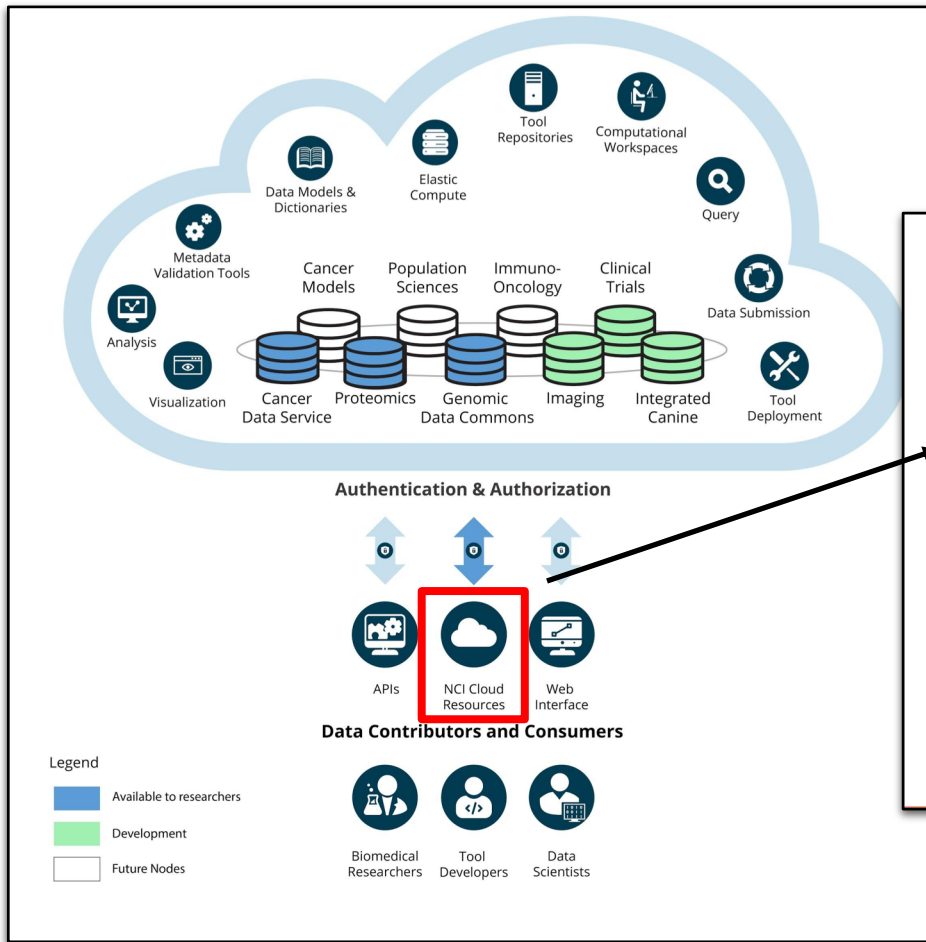
2020-04-10

Kawther Abdilleh, PhD
Fabian Seidl, PhD

Outline

- What is ISB-CGC?
- How do users find data to compute on using ISB-CGC?
- What analytical tools and applications are available through ISB-CGC?

NCI Cancer Research Data Commons Ecosystem

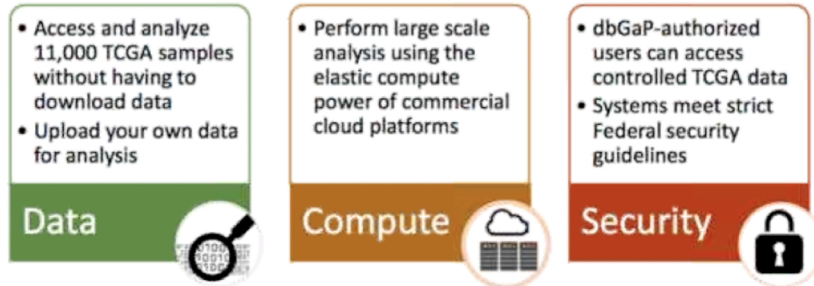


ISB-CGC is one of the NCI Cloud Resources

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

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 tools 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

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

Common problems of big data:

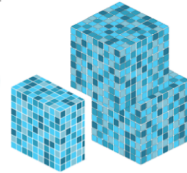
- Transfer speeds become bottlenecks with scaling data size
- Availability of data is tenuous
- Data discovery is onerous

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

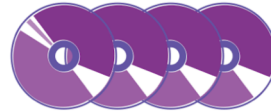
TCGA produced over

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

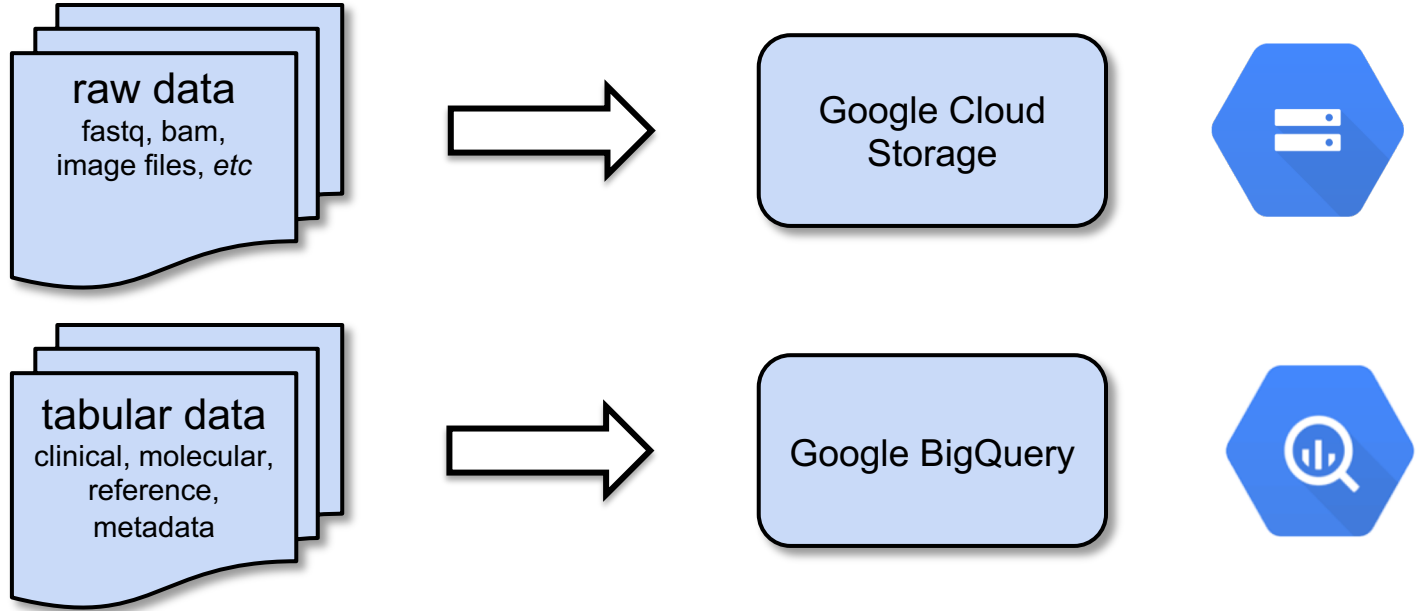
7

DIFFERENT
DATA TYPES



TCGA RESULTS & FINDINGS

How do users access data on ISB-CGC?

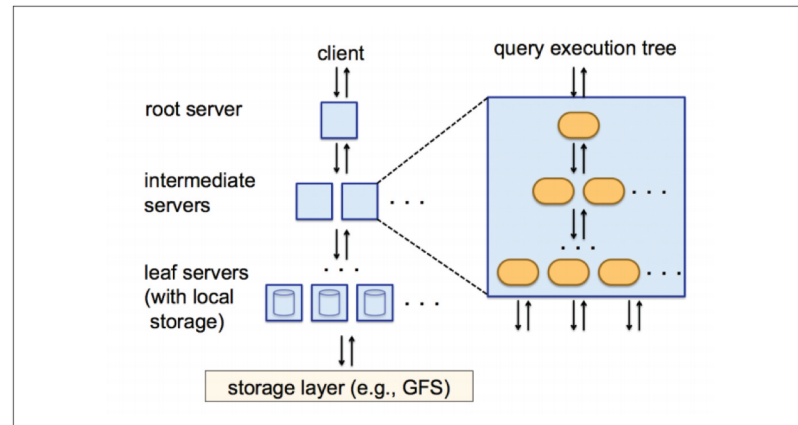


What is Google BigQuery and how does it enable –omics analyses?

- 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 cancer data analytics

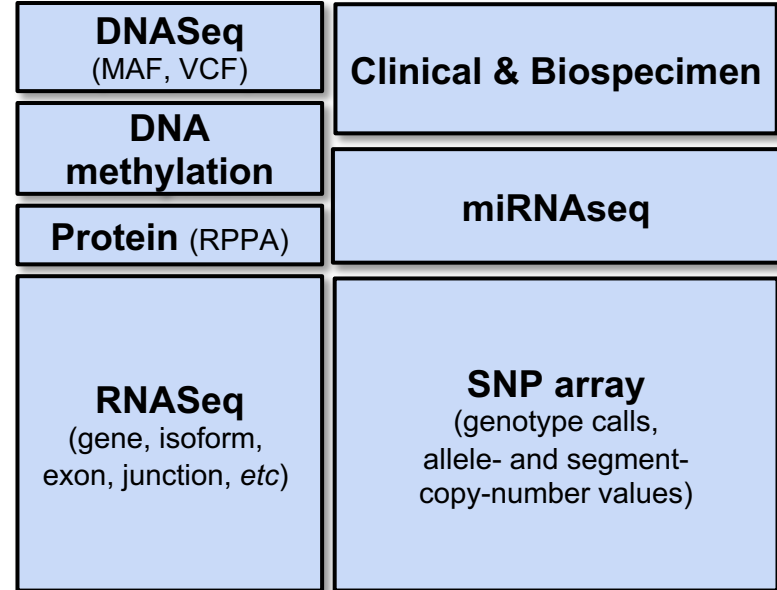
- Columnar database ideal for storing tabular data
- Query speed is automatically scaled by multiprocessing
- Powerful SQL language interface, including user defined functions
- Can join tables based on shared variables



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

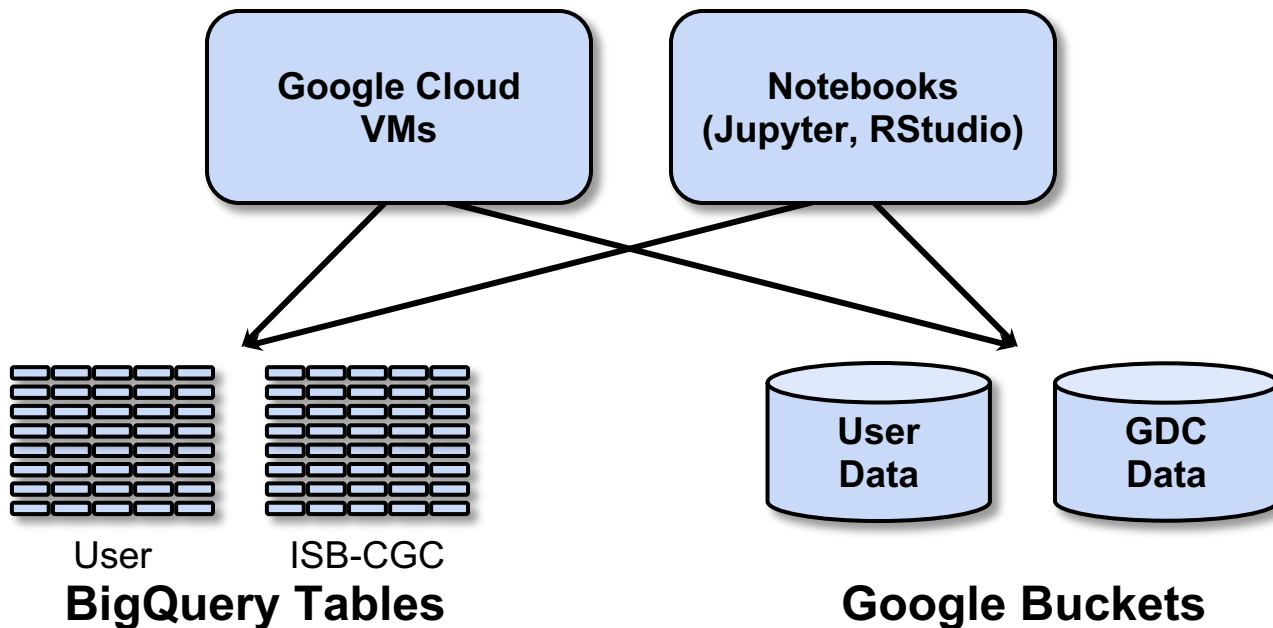


Over 300 open access BigQuery tables hosted by ISB-CGC

- 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

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...*

Interactive cancer 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...*

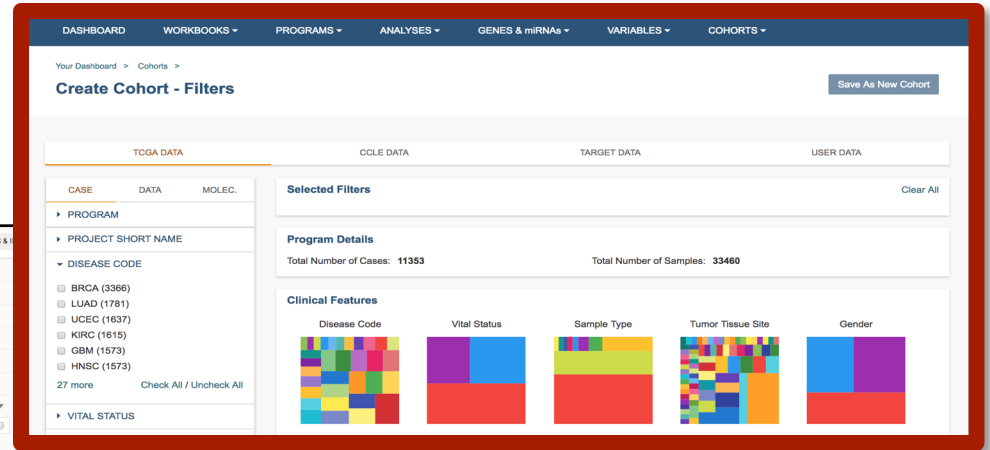
Direct Command line Access to Google virtual machines

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

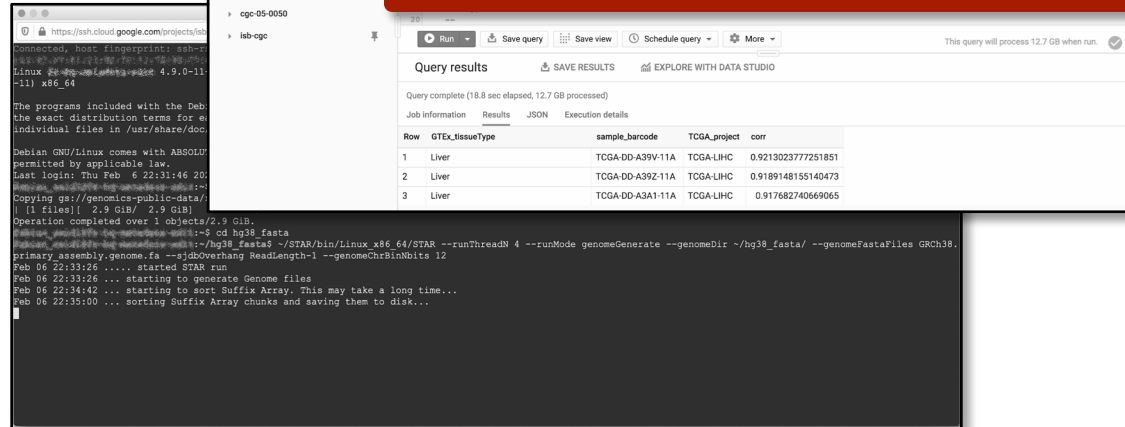
Three entry points for exploring cancer data on ISB-CGC

ISB-CGC web tools

Google BigQuery



Google VMs



```
Connected. Host fingerprint: ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAQgAAC3...
Linux 4.9.0-11-c11 x86_64

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

Debian GNU/Linux comes with ABSOLU
permitted by applicable law.
Last login: Thu Feb 6 22:31:46 20
root@cgc-05-0050:~# cat /dev/urand
Copying gs://genomics-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...
```

Interactive cohort-building using the ISB-CGC

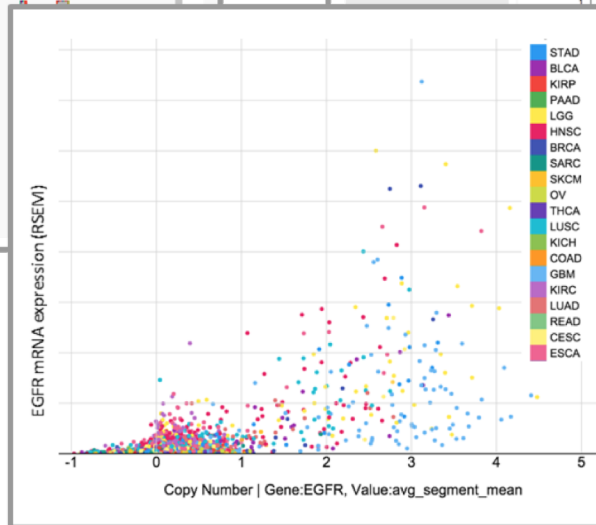
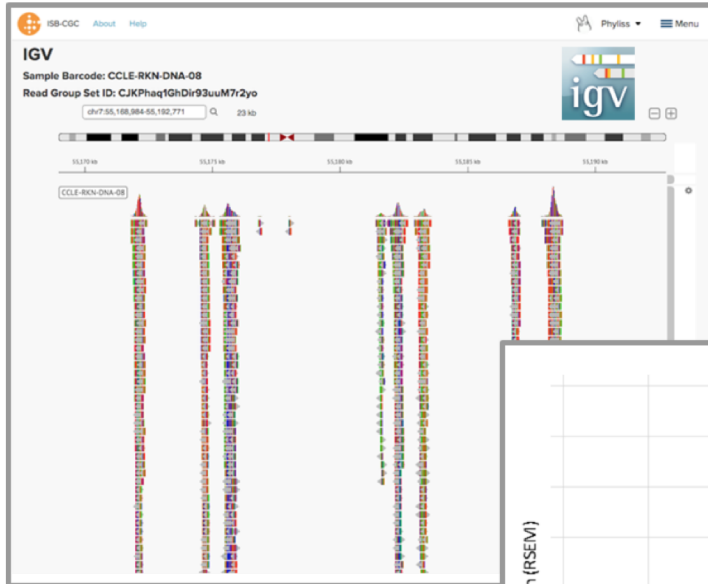
The interface displays the following components:

- Navigation Bar:** TCGA DATA, CCLE DATA, TARGET DATA, USER DATA
- Left Sidebar (Filter Categories):**
 - 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 (6,402 sample(s))
- Main Content Area:**
 - Selected Filters:** Clear All
 - Program Details:** Total Number of Cases: 11,353; Total Number of Samples: 33,460
 - Clinical Features:** Disease Code (Heatmap)

Filtered View (Bottom Screenshot):

- Selected Filters:** 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 (Visualizations)

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

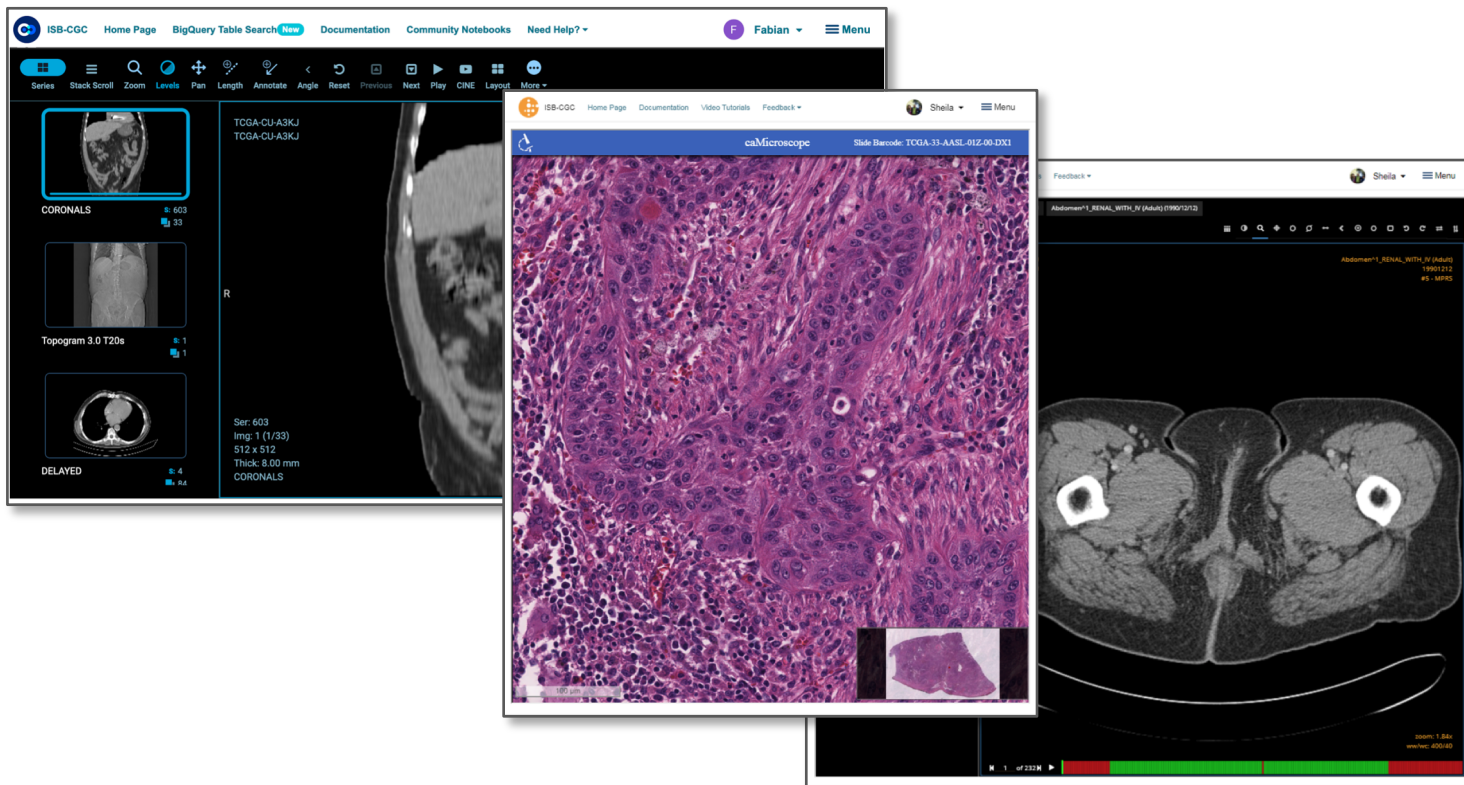
Show 25 entries Page Go Previous 1 2 3 ... 1588 Next

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: CURRENT

Name:

Program:

Category:
 CLINICAL BIOSPECIMEN DATA
 FILE METADATA
 GENOMIC REFERENCE DATABASE
 PROCESSED -OMICS 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 -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 - SCOMATIC 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 were used in 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: hg19](#) [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	3/26/2016		
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	CCLE	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

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



Three entry points for exploring cancer data on ISB-CGC

ISB-CGC WebApp

The screenshot shows the 'Create Cohort - Filters' page in the ISB-CGC WebApp. It features a navigation bar with 'DASHBOARD', 'WORKBOOKS', 'PROGRAMS', 'ANALYSES', 'GENES & miRNAs', 'VARIABLES', and 'COHORTS'. Below the navigation, there are tabs for 'TCGA DATA', 'CCLE DATA', 'TARGET DATA', and 'USER DATA'. A 'Selected Filters' section is visible, and a 'Gender' filter is shown as a grey square. A 'Save As New Cohort' button is in the top right corner.

Google BigQuery

The screenshot displays the Google BigQuery interface. On the left, there is a sidebar with 'Query history', 'Saved queries', 'Job history', 'Transfers', 'Scheduled queries', 'BI Engine', and 'Resources'. The main area is the 'Query editor' with a SQL query: `5 --
6 GTEX_top5K AS (
7 SELECT
8 gene_id,
9 gene_description,
10 STDEV(gene_exp) AS sigmaExp
11 FROM
12 [bigquery:GTEX_v7.gene_median_tpm]
13 GROUP BY
14 1,
15 2
16 ORDER BY
17 sigmaExp DESC
18 LIMIT
19 5000 ,
20 --`. Below the editor, the 'Query results' section shows a table with 3 rows and 5 columns: 'Row', 'GTEX_tissueType', 'sample_barcode', 'TCGA_project', and 'corr'. The results are:


Row	GTEX_tissueType	sample_barcode	TCGA_project	corr
1	Liver	TCGA-DD-A39V-11A	TCGA-LIHC	0.921302377251851
2	Liver	TCGA-DD-A39Z-11A	TCGA-LIHC	0.9189148155140473
3	Liver	TCGA-DD-A3A1-11A	TCGA-LIHC	0.917682740669065

Google VMs


The screenshot shows a terminal window with system boot logs. The logs include:

```
Connected, host fingerprint: ssh  
[...]  
Linux 4.9.0-111-x86_64  
The programs included with the D  
the exact distribution terms for  
individual files in /usr/share/d  
Debian GNU/Linux comes with ABSO  
permitted by applicable law.  
Last login: Thu Feb 6 22:31:46  
[...]  
[1 files] | 2.9 GiB / 2.9 GiB  
Operation completed over 1 objec  
[...]  
primary_assembly.genome.fa --sjdbOverhang ReadLength-1 --genomeChrBinNbits 12  
Feb 06 22:33:26 .... started STAR run  
Feb 06 22:33:26 ... starting to generate Genome files  
Feb 06 22:34:45 ... 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...
```



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!

BigQuery | FEATURES & INFO | SHORTCUTS | + COMPOSE NEW QUERY

Query editor | HIDE EDITOR | FULL SCREEN

```
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 --
```

Run | Save query | Save view | Schedule query | More

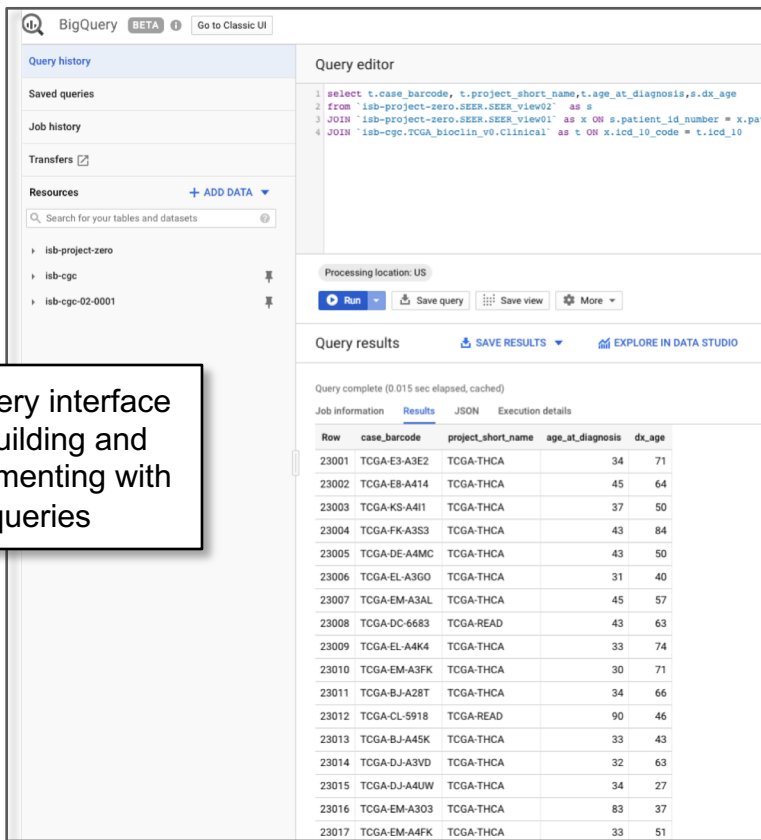
This query will process 12.7 GB when run. ✓

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



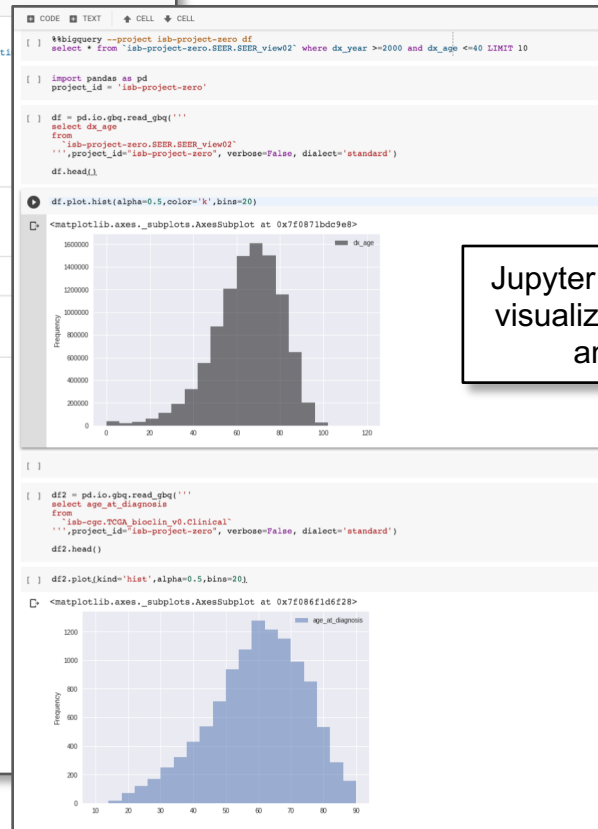
The screenshot shows the BigQuery interface. On the left, there's a sidebar with 'Query history', 'Saved queries', 'Job history', 'Transfers', and 'Resources'. The main area is the 'Query editor' with a SQL query:

```
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 x.patient_id_number = x.pat
4 JOIN `isb-cgc.TCGA_bioclin_v0.Clinical` as t ON x.tcd_10_code = t.tcd_10
```

Below the editor, it says 'Processing location: US' and has buttons for 'Run', 'Save query', 'Save view', and 'More'. The 'Query results' section shows a table with 20 rows of data:

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-A303	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 code cells. The first cell contains a BigQuery query and a histogram:

```
%%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
projct_id = 'isb-project-zero'
```

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

```
df.plot.hist(alpha=0.5, color='k', bins=20)
```

The histogram shows the frequency distribution of dx_age for the specified query. The x-axis is labeled 'dx_age' and ranges from 0 to 120. The y-axis is labeled 'Frequency' and ranges from 0 to 160,000. The distribution is unimodal and slightly right-skewed, peaking around 60-70.

The second cell contains another BigQuery query and a histogram:

```
df2 = pd.io.gbg.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)
```

The histogram shows the frequency distribution of age_at_diagnosis. The x-axis is labeled 'age_at_diagnosis' and ranges from 10 to 90. The y-axis is labeled 'Frequency' and ranges from 0 to 1200. The distribution is unimodal and slightly right-skewed, peaking around 60-70.

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, there's a navigation pane with 'Query editor' selected. The main area shows a SQL query editor with a query that filters for 'kirc_summed_sc_join'. Below the editor, a table of results is displayed with columns: Row, shor_name, sample_barcode, topq_gene_name, HTSeq_FPKM_IQ, topq_rna_fpkm, ht_seq, gc_gene_name, summed_value, and hc_rna_sum. The table contains 15 rows of data.

Built in syntax checking

Notebook (R or Python)

The screenshot shows a Jupyter Notebook interface. The first cell is a 'SETUP' cell with a comment '3 cells hidden'. The second cell is a 'BigQueries!' cell containing a BigQuery query. The query is: `sql = ''' SELECT COUNT(sample) AS n, Type FROM [big-cgc-02-0001.brca_single_cell_rna_gse75688_sample_info'] GROUP BY ... Type res0 = runQuery (bqClient, sql, dryRun=False) res0`. The third cell is an 'In runQuery ...' cell with a comment 'the results for this query were previously cached' and a table showing the result:

n	Type
0	515 SC
1	13 Bulk

. Below the table, it says 'We have 515 single cell RNA profiles, and 13 Bulk RNA profiles.' The fourth cell is another 'BigQueries!' cell with a query: `sql = ''' SELECT COUNT(sample) AS n, Type, [info], [info] FROM [big-cgc-02-0001.brca_single_cell_rna_gse75688_sample_info'] GROUP BY ... 2,3,4 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, there's a section 'BigQuery is NoOps—there is no infrastructure to manage and you don't need a database administrator—so you can focus on analyzing data to find meaningful insights, use familiar SQL, and take advantage of our pay-as-you-go model.' Below this, there are several navigation cards: 'Quickstarts' (Learn in 5 minutes), 'How-to guides' (Perform specific tasks), 'APIs & reference' (API, web UI, and command-line), 'Concepts' (Develop a deep understanding of BigQuery), 'Tutorials' (Walkthroughs of common applications), and 'Resources' (Pricing, quotas, release notes, and other resources).

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

The image shows a grid of Google Cloud Platform services and their free tier limits. Two callout boxes highlight specific services:

- BigQuery (Data Analytics):** 1 TB of queries per month. Fully managed, petabyte scale, analytics data warehouse. 1 TB of querying per month. 10 GB of storage.
- Compute Engine (Compute):** 1 F1-micro instance per month. Scalable, high-performance virtual machines. 1 f1-micro instance per month (US regions only—excluding Northern Virginia [us-east4]). 30 GB-months HDD. 5 GB-months snapshot in select regions. 1 GB network egress from North America to all region destinations per month (excluding China and Australia).

Category	Service	Free Tier Limit	Description
COMPUTE	Cloud Run	2 million Requests per month	Fully managed environment to run stateless containers.
DATABASE	Firestore	1 GB Storage	Scalable NoSQL, document database.
COMPUTE	Compute Engine	1 F1-micro instance per month	Scalable, high-performance virtual machines.
STORAGE	Cloud Storage	5 GB 30-day regional storage	Best-in-class performance, reliability, and pricing for all your storage needs.
DATA ANALYTICS	Pub/Sub	10 GB Messages per month	A global service for real-time and reliable messaging and streaming data.
COMPUTE	Cloud Functions	2 million Invocations per month	A serverless environment to build and connect cloud services with code.
COMPUTE	Google Kubernetes Engine Clusters	All size clusters	One-click container orchestration via Kubernetes clusters, managed by Google.
COMPUTE	App Engine	28 Instance hours per day	Platform for building scalable web applications and mobile back ends.
MANAGEMENT TOOLS	Stackdriver	50 GB Logs with 30-day retention	Monitoring, logging, and diagnostics for applications on Google Cloud and AWS.
DATA ANALYTICS	BigQuery	1 TB Queries per month	Fully managed, petabyte scale, analytics data warehouse.
AI AND MACHINE LEARNING	Vision AI	1,000 Units per month	Label detection, OCR, facial detection and more.
AI AND MACHINE LEARNING	Speech-to-Text	60 Minutes per month	Speech-to-text transcription – the same that powers Google's own products.

Contact us about setting up your own Google Cloud Platform Project!

feedback@isb-cgc.org

How To Get Started on ISB-CGC

The ISB-CGC provides both interactive (through a [web application](#)) and programmatic access to data hosted by institutes such as the Genomic Data Commons (GDC) of the National Cancer Institute (NCI), and the Wellcome Trust Sanger Institute, leveraging many aspects of the Google Cloud Platform. ISB-CGC hosts both open-access and controlled-access cancer genomics data from the NCI from a [variety of Programs and Data Sets](#). Some of these are controlled-access data [which require dbGaP authorization to access](#).

Click here to go to the [ISB-CGC Home Page](#).

Data Access and Google Cloud Project Setup

- A GCP project is required to make use of all of the data, tools, and Google Cloud functionality.
- Do you have a Google identity already (e.g. a GMail account)? Your institutional email may be a Google identity (if your institution uses Google Apps), or you may have a personal GMail address.
- If not, it only takes a minute to [create a Google identity](#). You can even link a non-GMail account (eg. scientist@nih.gov) as a Google identity by [this](#) method.
- Create your own GCP project and take advantage of a one-time [\\$300 Google Credit](#).
- If you have already used this one-time offer (or there is some other reason you cannot use it), please see the information here about how to request [ISB-CGC Cloud Credits](#).
- [Registering the GCP project](#)
- [Enable Required Google Cloud APIs](#)

Accessing and Analyzing Data via BigQuery

Google Cloud Life Sciences

Features

Cost-optimized compute

Google Cloud's Healthcare and Life Sciences team has optimized the most popular methods—like [GATK](#), [DeepVariant](#), and [Sentieon](#)—to run on GCP.

Fully integrated with GCP

Experience the power of Google Cloud's infrastructure with fast virtual machines, scalable storage, serverless data warehouses, and fully managed databases with GCP integration to tools like [Cloud Spanner](#) and [BigQuery](#).

Flexible machine sizes

Take advantage of [Compute Engine](#), our infrastructure as a service (IaaS), to run large-scale workloads on virtual machines and pay only for what you use.

Open and interoperable

Use the tools and workflows you already know and enjoy support for open industry standards like [Global Alliance for Genomics and Health](#).

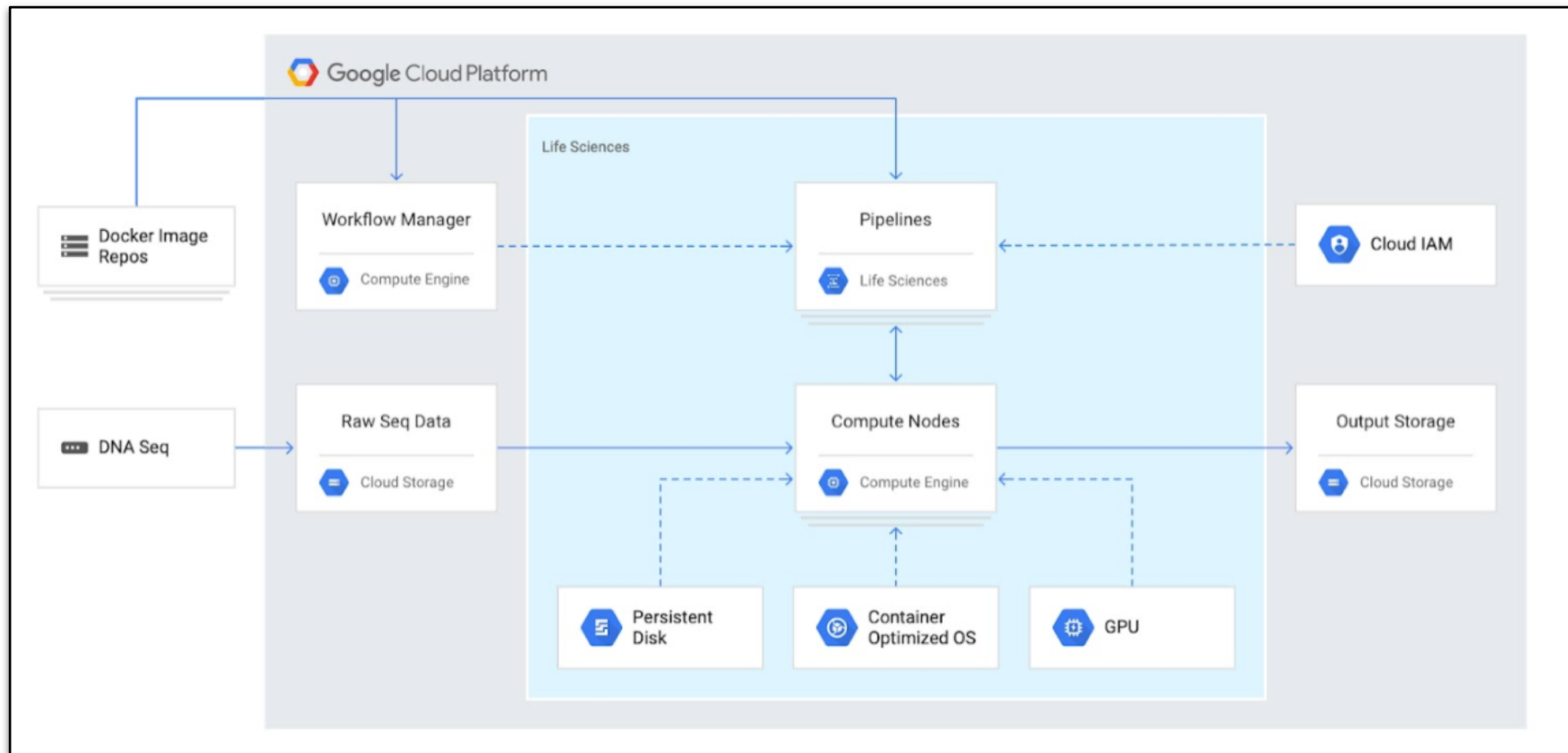
Built for batch processing

[Preemptible VMs](#) for affordable batch processing on fault-tolerant workloads to save you time and money.

Ready for AI / ML

Bring your data closer to [public datasets](#) and advanced analytics that come along with GCP.

Google Cloud Life Sciences



We have BigQuery AI/ML examples in our Query of the Month blog

July, 2018

First look: BigQuery ML

Exciting news! Google has just released a beta feature in BigQuery: Machine Learning. There are two available model types, linear and logistic. The first, linear regression, models a continuous variable given a selection of variables, both categorical (eg US postal code) and numerical (eg age or weight). The second, logistic regression, models a binary label given some variables. For example, for classification between two groups, which we have used extensively in this blog. All the groups we created had features like 'does or does not have a mutation in GATA3'. Even the logistic regression is regularized! We have both [L1 and L2 regularization available](#), which is similar to an implementation of elasticnet.

In the following examples, I'm going to be working in the BigQuery web interface, but it's also possible to train and apply these models using the command line tool (bq), the REST API, or a scripting language (R or python).

The introductory documentation can be found [here](#).

Something people are always concerned about: how much does it cost?! Well, from the docs, at this time (July 2018) pricing is still under development. But essentially it's similar to other queries. You're charged according to how much data is processed in training the model (first 10GB free). However, the actual data read for training is more than just the size of the table. It's not entirely clear at this point, but when I later report it.

August, 2018

Using BigQuery ML in a Shiny app.

Last month, we tried out the newly-released Google BigQuery ML. This month we'll continue to build examples, learn some new things, and build a [shiny web app](#).

One newsworthy bit of information, in case you missed it a few months ago, is that the R package used for interacting with BigQuery, `bigquery`, has undergone a major revision (hitting [version 1.0.0](#)), and many of the function calls have changed significantly. The returned object from making a BigQuery call (with function `bq_project_query`) is now a "tibble" rather than a data frame.

Working with BigQuery ML is quite a bit different than what we've done before. In the past, when working with BigQuery, we've computed different statistics, and we've even used those statistics for classification, but that work was all done in the SQL - including, for example, formulating a Z-score in SQL. Now, most of our work will go into preparing the training data table to be used when fitting the model.

When fitting models, we have two important parameters to think about: the L1 and L2 regularization rates. (There are other parameters, but we'll focus on these for the moment.) Both of these parameters effectively push the weights of less useful predictors towards zero. L2 (or euclidean norm) will push weights towards zero, but L1 regularization will make variable (gene) weights exactly zero. Using these regularizers can help us get an idea of which features (eg genes) are most useful in separating groups (eg cancer types).

Example workflows launched from cloud-based notebooks

ISB-CGC Community Notebooks¶

Title: How to create convert 10X bams to fastq files using dsub
Author: David L Gibbs
Created: 2019-08-07
Purpose: Demonstrate how to make fastq files from 10X bams
Notes:

How to use dsub to convert 10X bam files to fastqs

In this example, we'll be using DataBiosphere's dsub. dsub makes it easy to run a job without having to spin up and shut down a VM.

Using samtools to index, sort, and convert a bam file to a fastq file.

In this example, we'll be using the Google Genomics Pipelines API. The pipelines API makes it easy to run a job without having to spin up and shut down a VM.

How to find a tool using GA4GH Tool Repository Service (TRS)

The Global Alliance for Genomics and Health (GA4GH) is an international coalition, formed to enable the sharing of genomic and clinical data. One of the things that GA4GH makes available is the ability to find and use tools using their GA4GH Tool Registry Service (TRS) API. The GA4GH TRS API is a standard for listing and describing available tools (both stand-alone, Docker-based tools as well as workflows in CWL, WDL or Nextflow) in a given registry (like Dockstore, BioContainers, and Agora).

In this notebook, we'll explore how to use the GA4GH TRS API to find tools of interest from the Dockstore.

How to use a WES service

This notebook is designed to be a quick introduction to using a workflow execution service (WES) and is intended as a follow-up to a previous notebook on searching for tools using a tool registry service (TRS; How to find a tool using TRS [here](#)). This notebook must be run in an environment capable of running docker. Google Colab notebooks will be extremely difficult to use. It's advised that a Jupyter-lab environment is started using the Google Cloud Console, AI platform.

Software used:

wes-service, a client and server implementation of the GA4GH Workflow Execution Service 1.0.0 API.

Easily run batches of jobs in the cloud using dsub

Overview

`dsub` is a command-line tool that makes it easy to submit and run batch scripts in the cloud.

The `dsub` user experience is modeled after traditional high-performance computing job schedulers like Grid Engine and Slurm. You write a script and then submit it to a job scheduler from a shell prompt on your local machine.

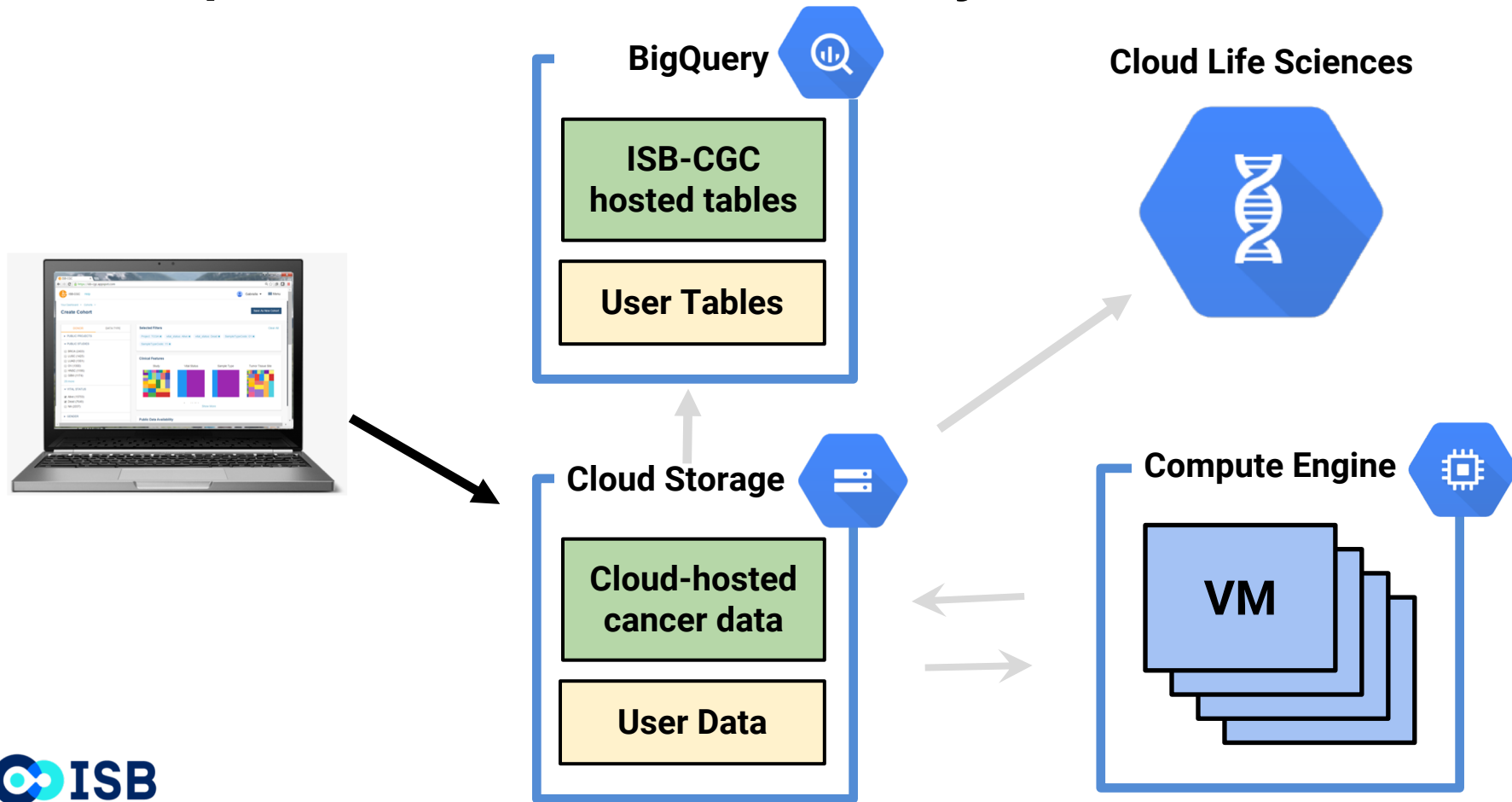
Today `dsub` supports Google Cloud as the backend batch job runner, along with a local provider for development and testing. With help from the community, we'd like to add other backends, such as a Grid Engine, Slurm, Amazon Batch, and Azure Batch.

Getting started

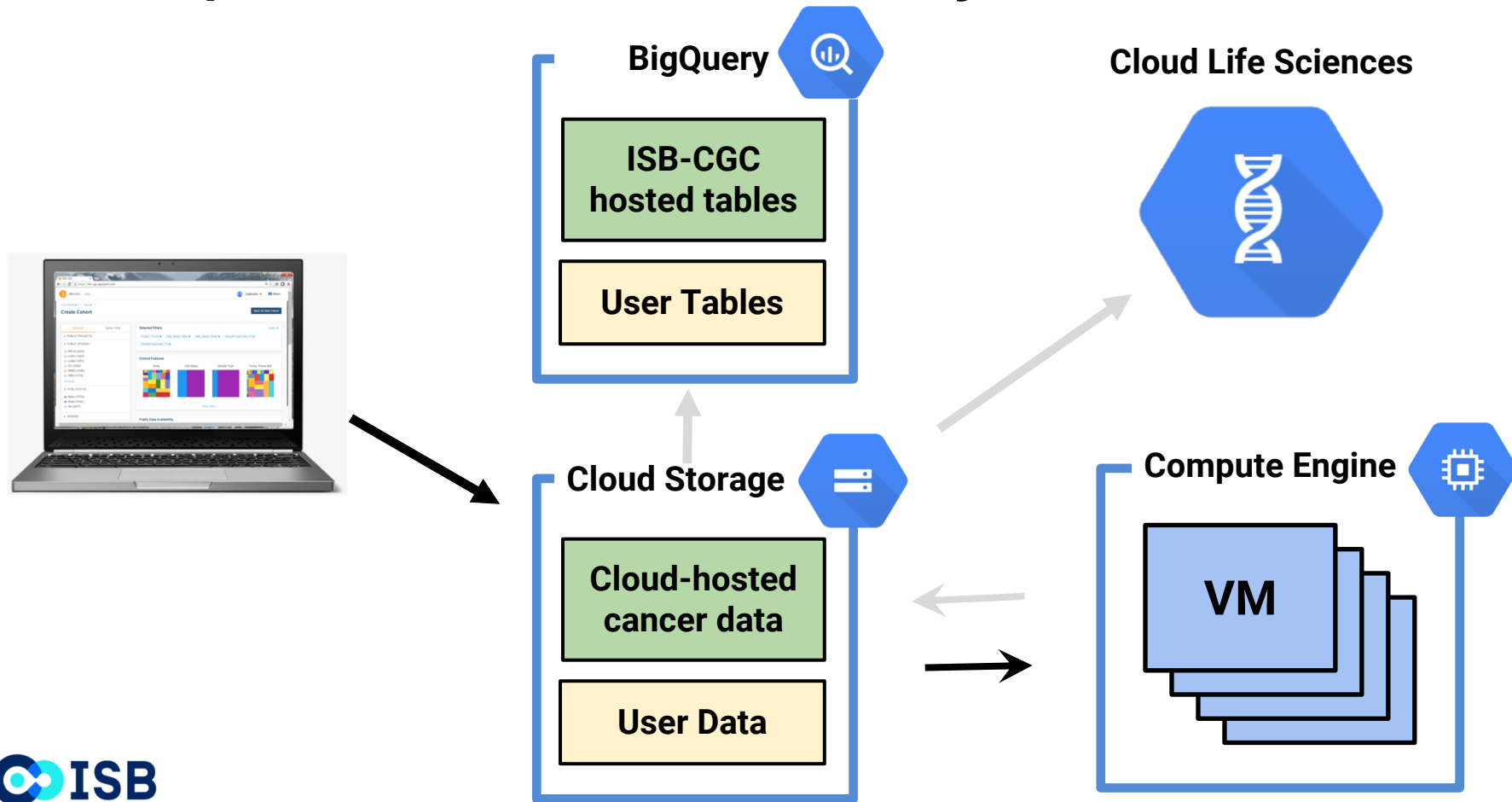
You can install `dsub` from [PyPI](#), or you can clone and install from [github](#).

```
dsub \  
  --name kallisto_quant \  
  --project ${GS_PROJECT} \  
  --zones 'us-*' \  
  --image "nareshr/kallisto:v0.43" \  
  --input "KALIDX=${GS_BUCKET}/Homo_sapiens.GRCh37.cdna.all.kal.idx" \  
  --input "FASTQ=${GS_BUCKET}/All_CCLE_customDB.fastq" \  
  --output-recursive "KALOUT=${GS_BUCKET}/output" \  
  --logging ${GS_BUCKET}/log \  
  --min-cores 8 \  
  --command 'kallisto quant -i ${KALIDX} -o ${KALOUT} -b 100 --single -l 180 -s 20 -t 8 ${FASTQ}' \  
  --wait
```

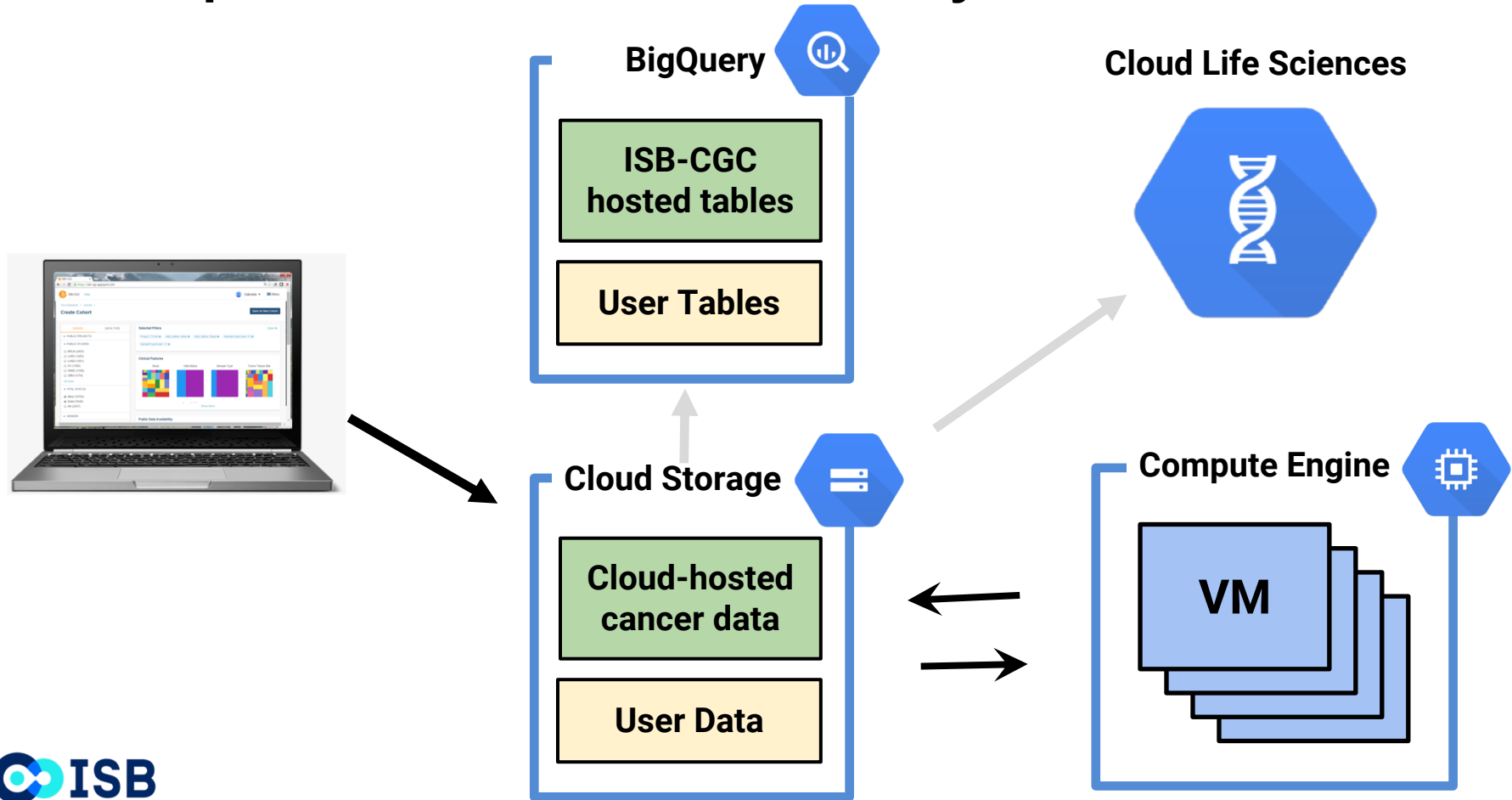
Example end-to-end workflow analysis on ISB-CGC



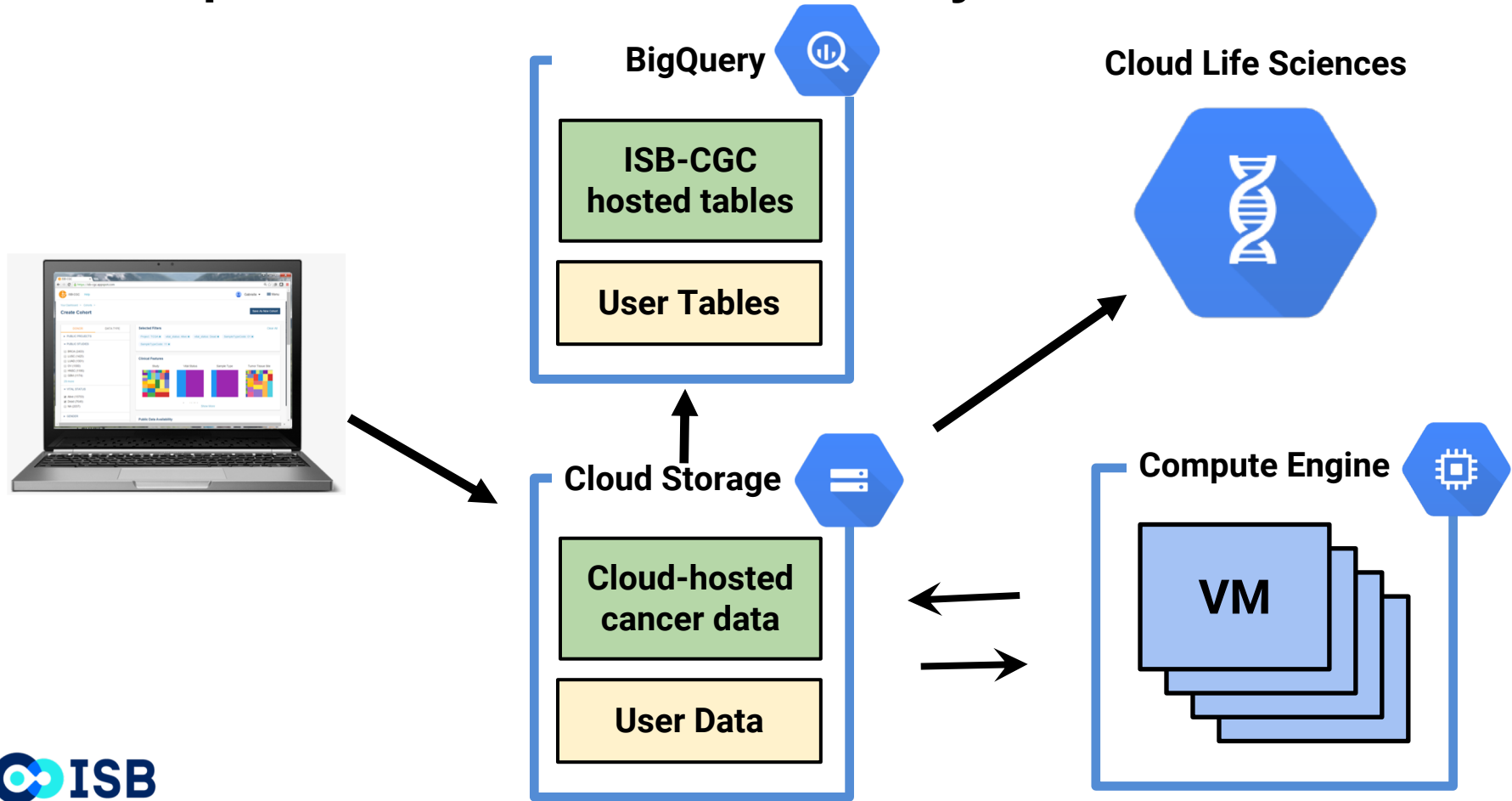
Example end-to-end workflow analysis on ISB-CGC



Example end-to-end workflow analysis on ISB-CGC



Example end-to-end workflow analysis on ISB-CGC



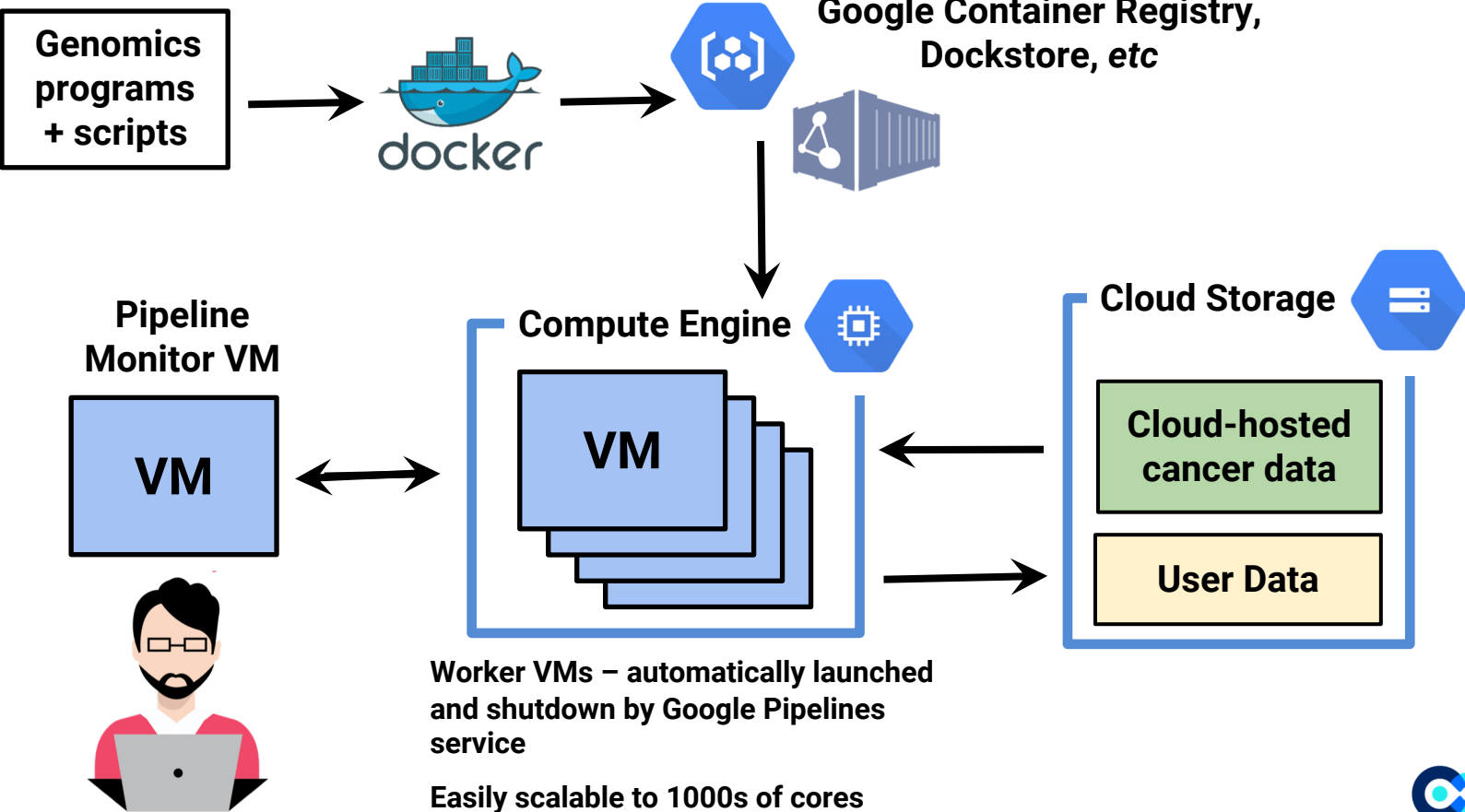
Example RNASeq quantification workflow using a Docker image

```
STAR_alignment_run.sh — Desktop Add License
1 wget https://api.gdc.cancer.gov/data/25aa497c-e615-4cb7-8751-71f744f9691f \
2 $path_to_reference/gencode.v22.annotation.gtf.gz
3 gunzip gencode.v22.annotation.gtf.gz
4
5 path_to_reference=/home/fseidl/genome
6 mkdir $path_to_reference/star_index_oh75
7
8 sudo docker run --rm -v $path_to_reference:/data -t broadinstitute/gtex_rnaseq:v8 \
9 /bin/bash -c "STAR \
10 --runMode genomeGenerate \
11 --genomeDir /data/star_index_oh75 \
12 --genomeFastaFiles /data/GRCh38.d1.vd1.fa \
13 --sjdbGTFfile /data/gencode.v22.annotation.gtf \
14 --sjdbOverhang 75 \
15 --runThreadN 4"
16
17 sudo docker run --rm -v $path_to_reference:/data -t broadinstitute/gtex_rnaseq:v8 \
18 /bin/bash -c "rsem-prepare-reference \
19 /data/GRCh38.d1.vd1.fa \
20 /data/rsem_reference \
21 --gtf /data/gencode.v22.annotation.gtf \
22 --num-threads 4"
23
24 path_to_data=/home/fseidl/bams
25 input_bam=HG00182.mapped.ILLUMINA.bwa.FIN.low_coverage.20120522.bam
26 sample_id=HG00182
27 mkdir $path_to_data
28
29 gsutil cp gs://genomics-public-data/1000-genomes/bam/$input_bam $path_to_data
30
31 sudo docker run --rm -v $path_to_data:/data -t broadinstitute/gtex_rnaseq \
32 /bin/bash -c "src/run_SamToFastq.py /data/$input_bam -p $sample_id -o /data"
33
34 # STAR alignment
35 sudo docker run --rm -v $path_to_data:/data -v $path_to_reference:/genome -t broadinstitute/gtex_rnaseq:v8 \
36 /bin/bash -c "src/run_STAR.py \
37 /genome/star_index_oh75 \
38 /data/${sample_id}_1.fastq.gz \
39 /data/${sample_id}_2.fastq.gz \
40 ${sample_id} \
41 --threads 4 \
42 --output_dir /tmp/star_out && mv /tmp/star_out /data/star_out"
43
```

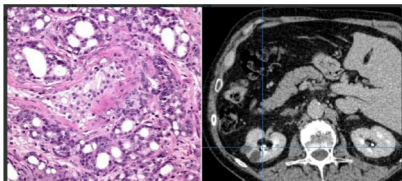


Row	HGNC_gene_symbol	gene_id	normalized_count
1	RELN	5649	5092.9882
2	DDX49	54555	1233.9143
3	OCSTAMP	128506	1.2346
4	RUFY4	285180	0.9481
5	SLC6A4	6532	1.0684
6	NOXA1	10811	14.8593
7	TAGLN2	8407	3785.6545
8	ZNF484	83744	40.8805
9	RNF217	154214	167.8584
10	RHOC	389	2670.3879
11	RNF219	79596	136.9329
12	MANEA	79694	797.0085
13	PALB2	79728	174.2287
14	MRPL35	51318	803.0303
15	IQSEC1	9922	2266.2474
16	FAM57B	83723	1.3774
17	CFLAR	8837	1108.9744
18	MAML2	84441	79.8898

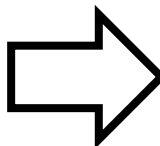
Advanced workflow execution on -omics data enabled by ISB-CGC



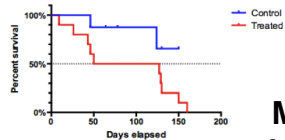
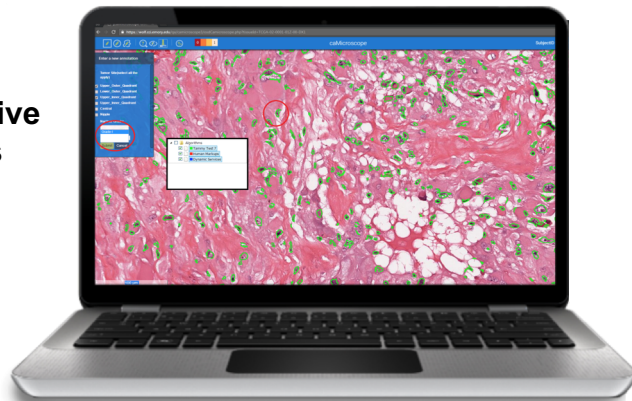
Performing deep learning and integrative analysis on pathology images



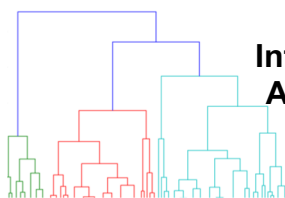
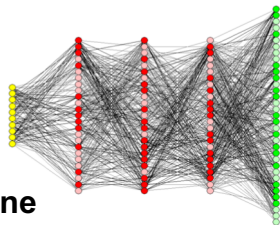
Pathology and Radiology images in Cloud Storage



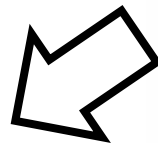
Interactive Tools



Machine Learning & Integrative Analyses

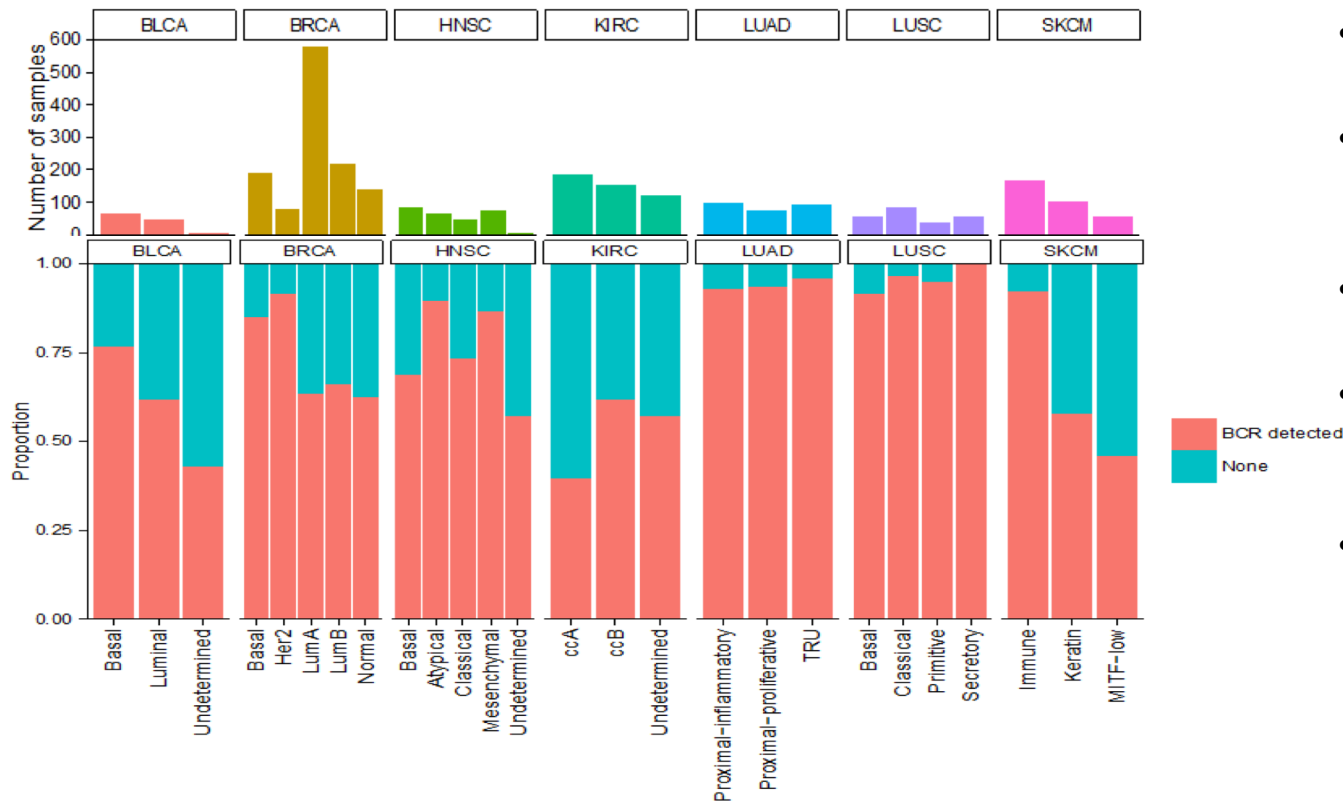


BIG QUERY



Automated Segmentation & Feature Extraction

Investigating B-Cell receptor expression and prognosis in tumor progression



- 10k samples
- – 1 day per sample per cluster node
- **12 weeks** @ UNC HPC
- Google Cloud Platform using >10k clusters
- **2 days** on GCP on ISB-CGC

Re-processing 70+ terabytes of raw RNA-Seq data using Google Cloud preemptible VMs

- Sambamba (sorting, indexing)
- Picard Tools (BAM to FASTQ)
- Trim Galore!
- Kallisto (pseudoalignment, quantification)
- Docker (orchestration, reproducibility)
- Ran pipeline using Preemptible VMs
- \$1,065.49 / 11,373 samples

Re-processing 70+ terabytes of raw RNA-Seq data using Google Cloud preemptible VMs

Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n1-standard-1	1	3.75GB	\$0.0475	\$0.0100
n1-standard-2	2	7.5GB	\$0.0950	\$0.0200
n1-standard-4	4	15GB	\$0.1900	\$0.0400
n1-standard-8	8	30GB	\$0.3800	\$0.0800
n1-standard-16	16	60GB	\$0.7600	\$0.1600
n1-standard-32	32	120GB	\$1.5200	\$0.3200
n1-standard-64	64	240GB	\$3.0400	\$0.6400
n1-standard-96 Skylake Platform only	96	360GB	\$4.5600	\$0.9600

- Affordable, short-lived compute instances
- Suitable for batch jobs and fault-tolerant workloads
- Same machine types and options as regular compute instances
- Last for up to 24 hours
- Similar to "spot instances" in AWS

Some more workflows done by 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: PMC5159871

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

PJ 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

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

101/160960

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

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

Resources to help you get started on ISB-CGC

- [ISB-CGC User Documentation](#)
- [Query of the Month Blog](#)
- [ISB-CGC Github Pipelines Repo](#)

For feedback/suggestions or need help getting starting, contact us at feedback@isb-cgc.org

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