Cancer Data Analytics on the ISB-CGC platform

April 9, 2021

www.isb-cgc.org

Agenda for Today's Presentation

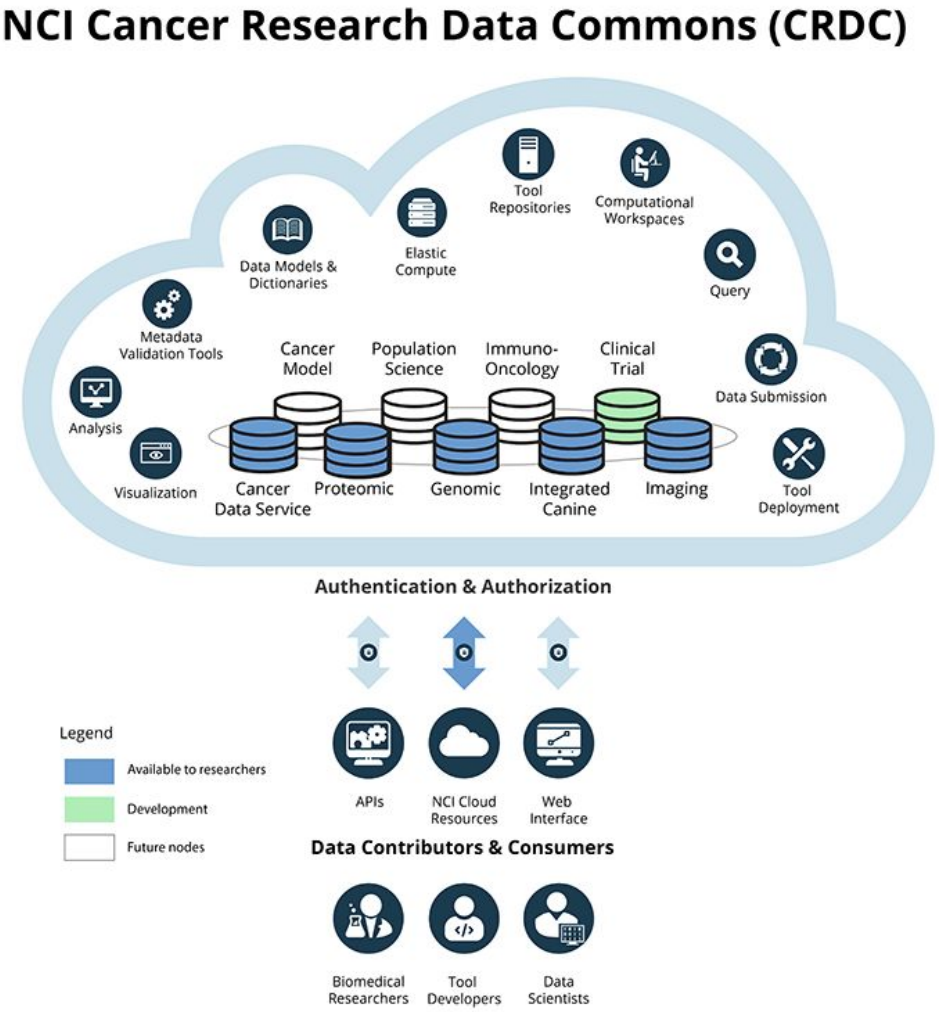
- An introduction to the ISB-CGC role in the CRDC
- The data analysis avenues accessible via the ISB-CGC
  - Web tools
  - Google VMs
  - Google BigQuery
- Examples of analyses run with our tools

ISB-CGC

The CRDC Today

The CRDC is a **cloud-based data** science infrastructure that provides **secure access** to large, **comprehensive**, and expanding collections of **cancer research data**. Users can explore and use analytical and visualization tools for data analysis in the cloud.

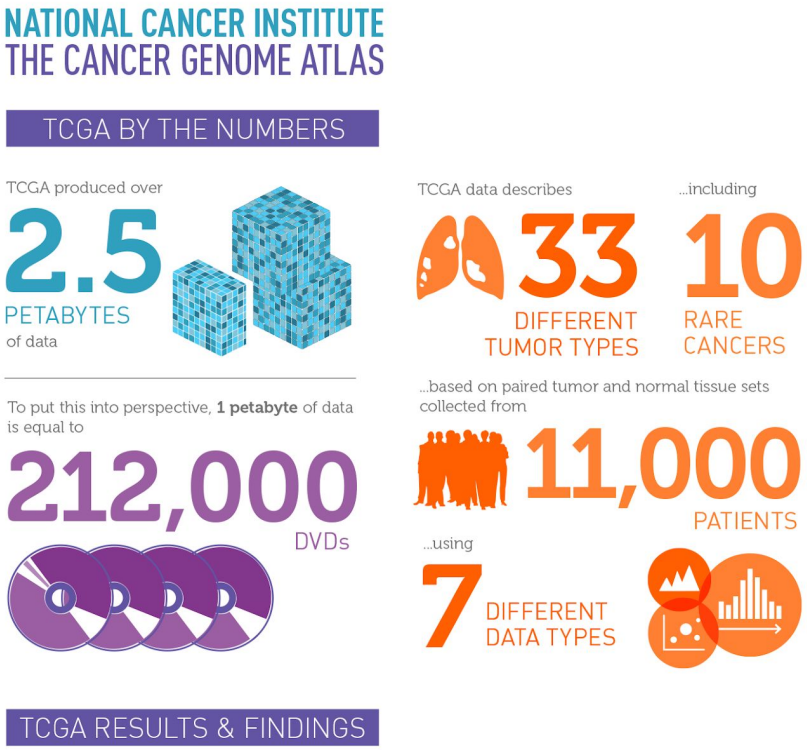
The Cloud Resources have >3,000 on average active users per month from >500 institutions around the world



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

Common problems of big data:

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



The ISB-CGC landing page provides access to our resources

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

Get started today! Contact us about setting up your own Google Cloud Platform Project with free cloud credits

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

Data 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

Resources

Interactive Patient Cohort Exploration

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

Programmatic Access

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

ISB-CGC is a component of the NCI Cancer Research Data Commons and has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

PRIVACY POLICY

www.isb-cgc.org

ISB-CGC

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

- Intentionally a “thin” layer on top of Google Cloud
  - Easy access to Google Tools
  - Direct access to NCI datasets
    - Raw GC bucket files
    - structured “BigQuery” tables for interactive data mining)

ISB-CGC

The data analysis avenues accessible via the ISB-CGC

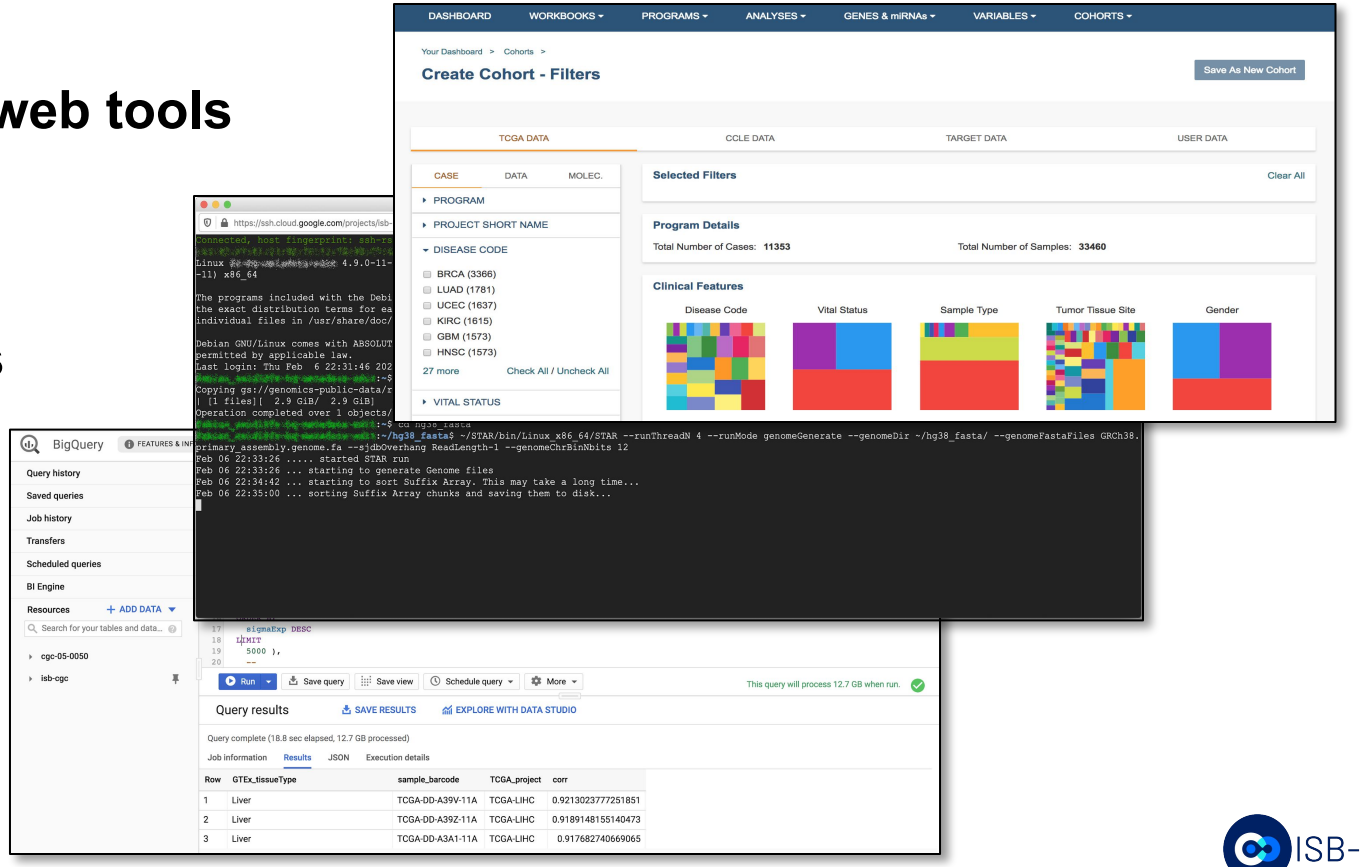
ISB-CGC

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC web tools

Google VMs

BigQuery



ISB-CGC

Some example use-cases of each major entry point

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
- Directly run statistical tests on data in BigQuery using custom functions
- *etc*…

Direct Command line Access to Google virtual machines

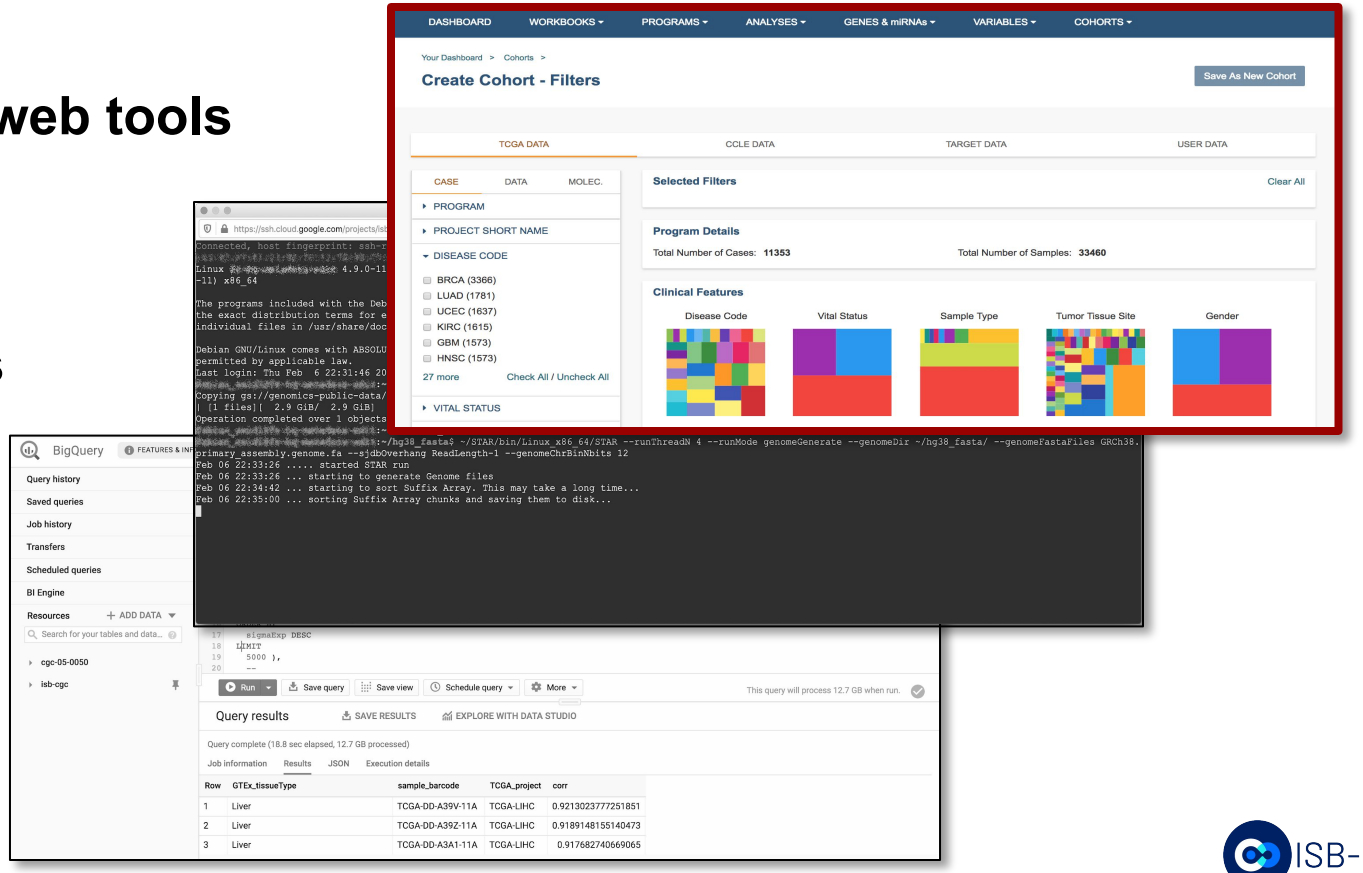
- Test new algorithm on hundreds or thousands of BAM or FASTQ files
- Run novel image segmentation method across whole-slide images
- *etc*…

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC web tools

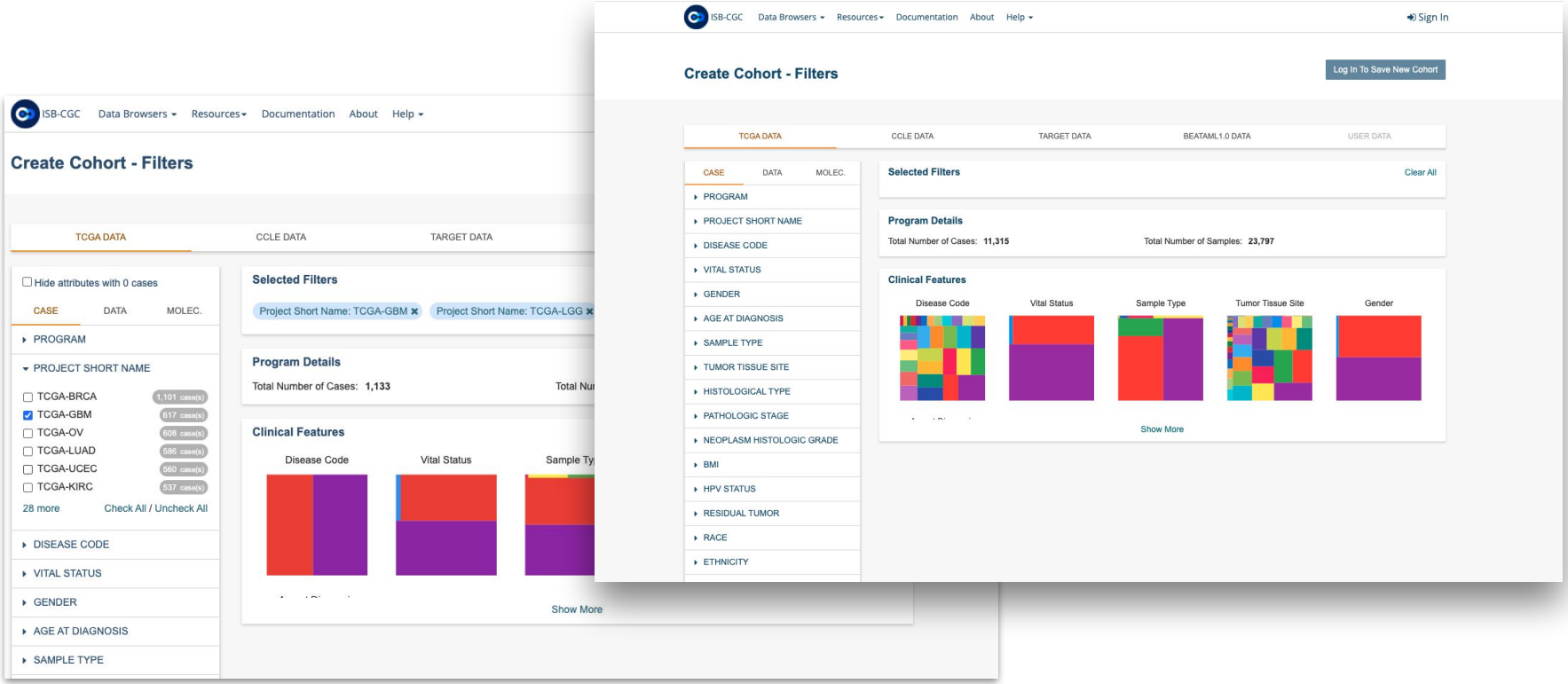
Google VMs

BigQuery



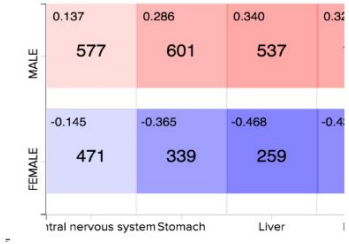
ISB-CGC

Interactive cohort-building using the ISB-CGC



ISB-CGC

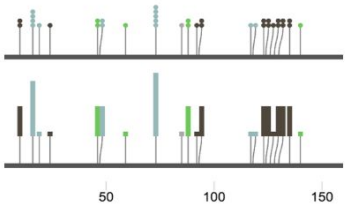
ISB-CGC has several interactive apps to generate quick plots



Cubby Hole Plot

Used to plot two categorical features. Boxes are colored by their related p-values.

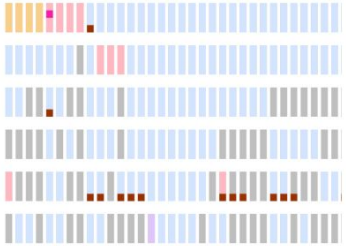
Start A New Workbook Cubby Hole Plot



SeqPeek

This visualization shows where somatic mutations have been observed on a linear representation of a specific protein. Each horizontal strip represents the protein, with data from different tumor types (aka cohorts or studies) shown stacked one on top of the other.

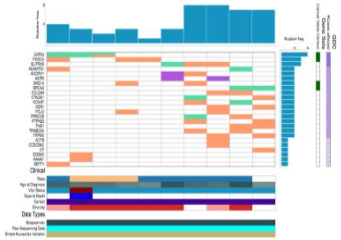
Start A New Workbook SeqPeek



OncoPrint

Used to plot multiple genomic alteration (somatic mutation) events across a set of samples by heatmap. OncoPrint is developed and provided by cBioPortal.

Start A New Workbook OncoPrint



OncoGrid

Used to view multiple genomic alteration (somatic mutation) events, clinical data, available files across a set of cases by interactive heatmap. OncoGrid library is developed at Ontario Institute for Cancer Research (OICR).

Start A New Workbook OncoGrid

ISB-CGC

File browser to browse the data hosted in cloud buckets

Integrated genome viewer (view read pile-ups)

caMicroscope (view histology)

OHIF (view radiology)

Your Dash

File

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

CSV | BigQuery | GCS

Showing 1 to 25 of 39692 entries

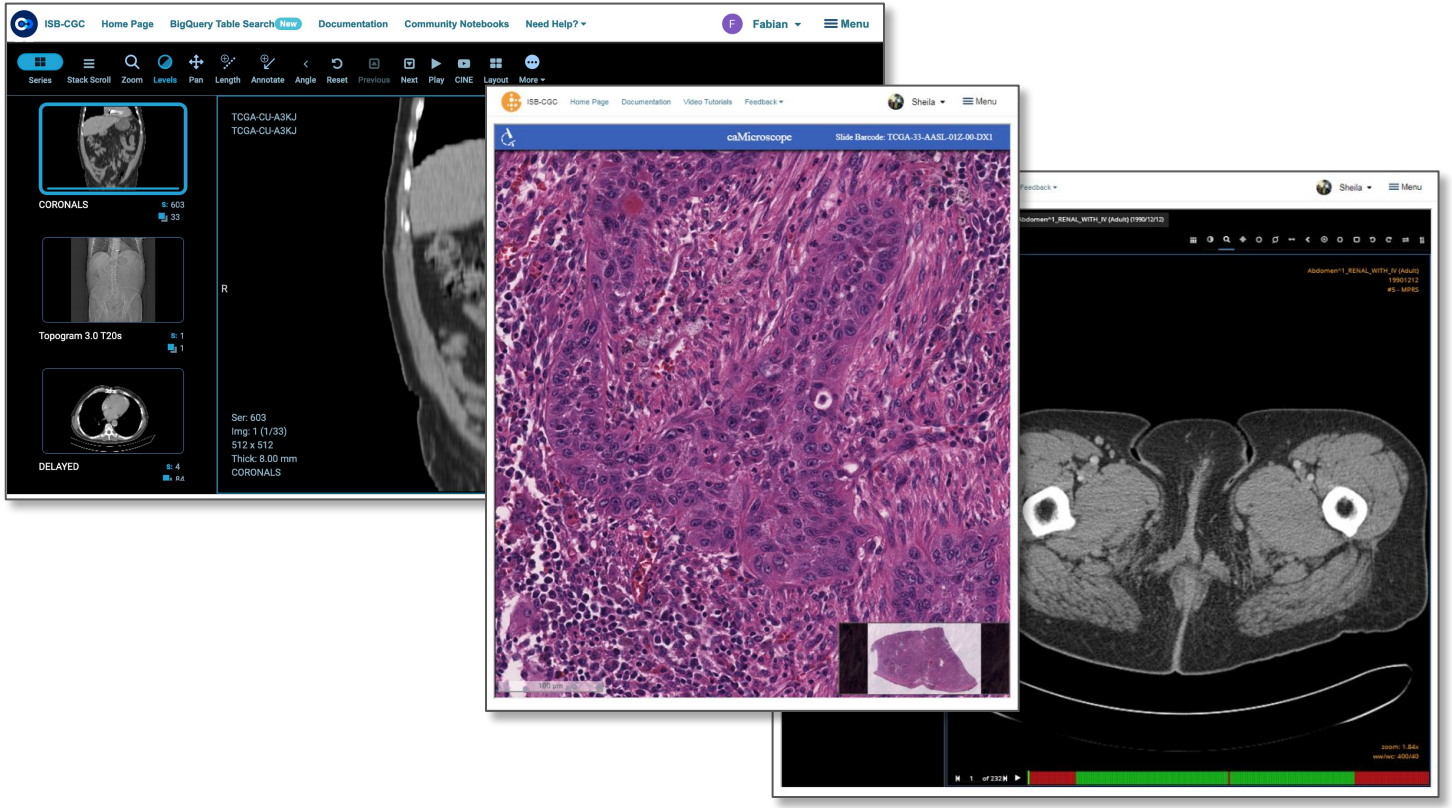
Show 25 **entries** **Page** Go Previous 1 2 3 ... 1588 Next

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-A66O	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-A66O	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-A66O	UNCID_2171596.c7f5714… [GDC ID: b677ea35-d758-…	BRCA	RNA-Seq	Illumina HiSeq	Raw sequencing data	Aligned reads	BAM	7.8 GB
TCGA	TCGA-OL-A66O	c61047b5e4ae38963735fc… [GDC ID: 0a6db03e-748a-…	BRCA	WXS	Illumina HiSeq	Raw sequencing data	Aligned reads	BAM	4.9 GB
TCGA	TCGA-OL-A66O	256cd674e76be0f163766b… [GDC ID: 72a31a7e-99df-4…	BRCA	WXS	Illumina HiSeq	Raw sequencing	Aligned reads	BAM	7.2 GB

ISB-CGC

Interactive image viewers allows for browsing of cancer imaging data



ISB-CGC

Mitelman database is also hosted by ISB-CGC

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

Also ongoing work towards hosting the TP53 database soon

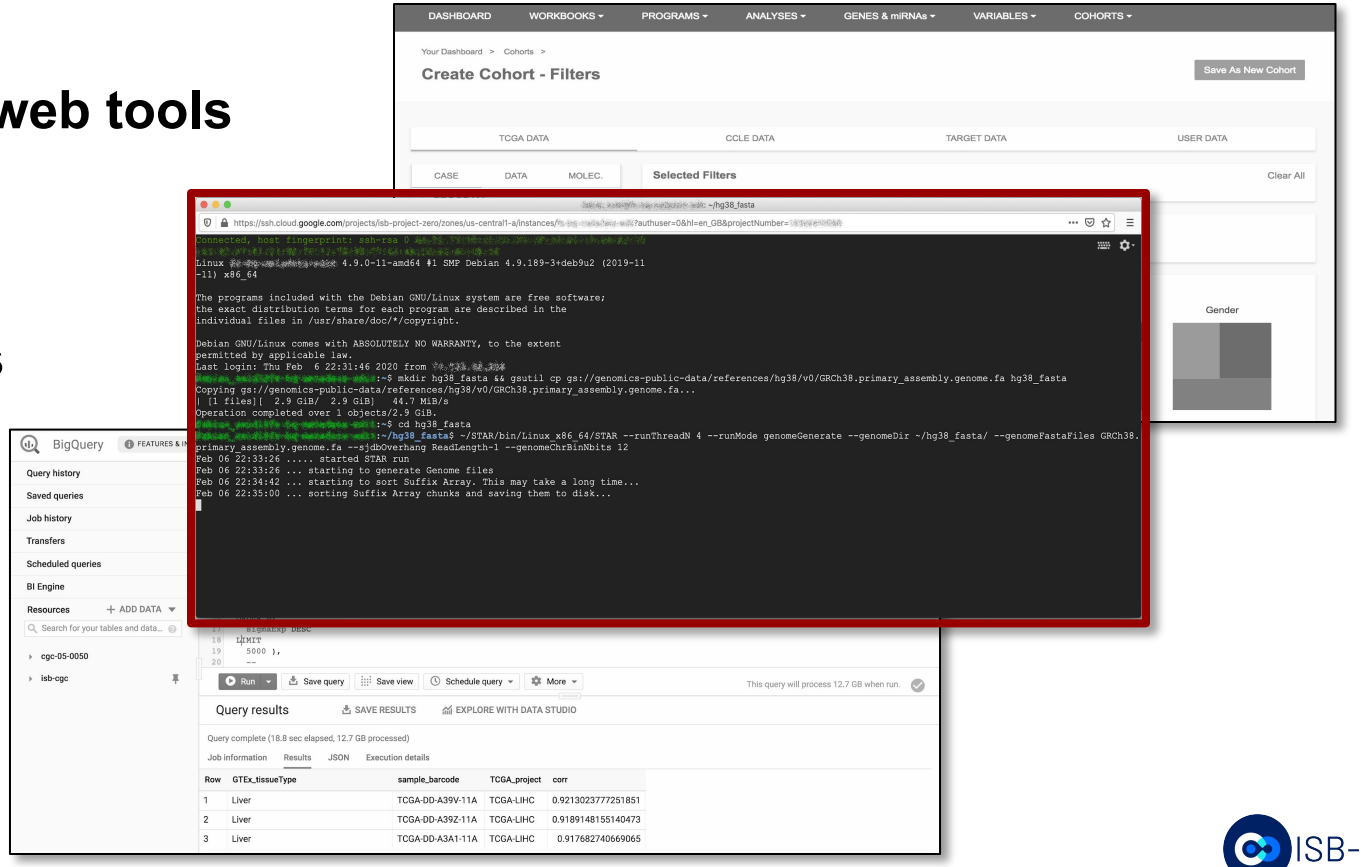
ISB-CGC

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC web tools

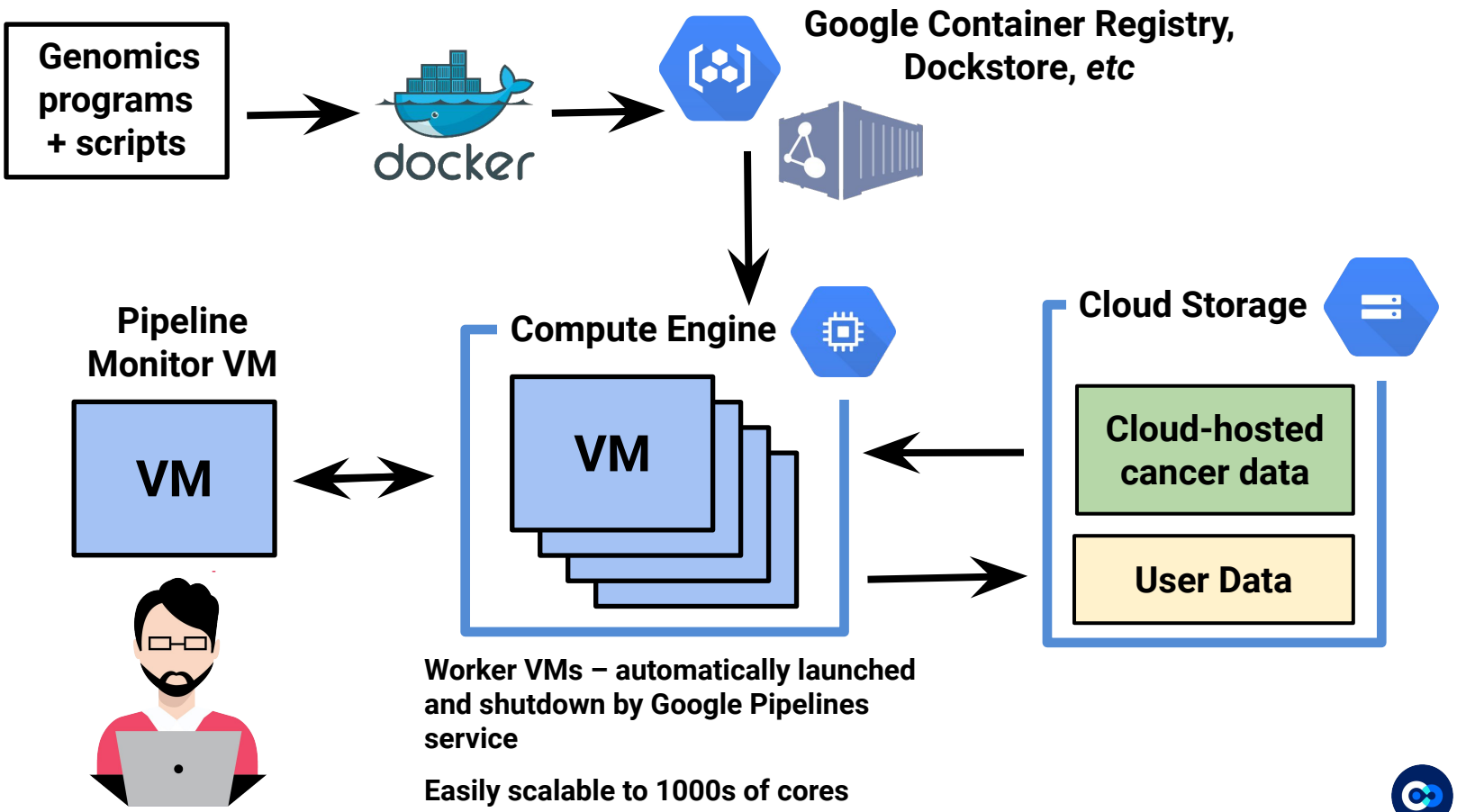
Google VMs

BigQuery



ISB-CGC

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



ISB-CGC

Google Cloud Platform Free Tier lets you perform small scale computes with no cost!

COMPUTE
Cloud Run
2 million
Requests per month
A 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 m

STORAGE
Cloud Storage
5 GB
Months 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 an 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 applications on Google Cloud and AW

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 sa powers Google's own products.

COMPUTE

Compute Engine

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

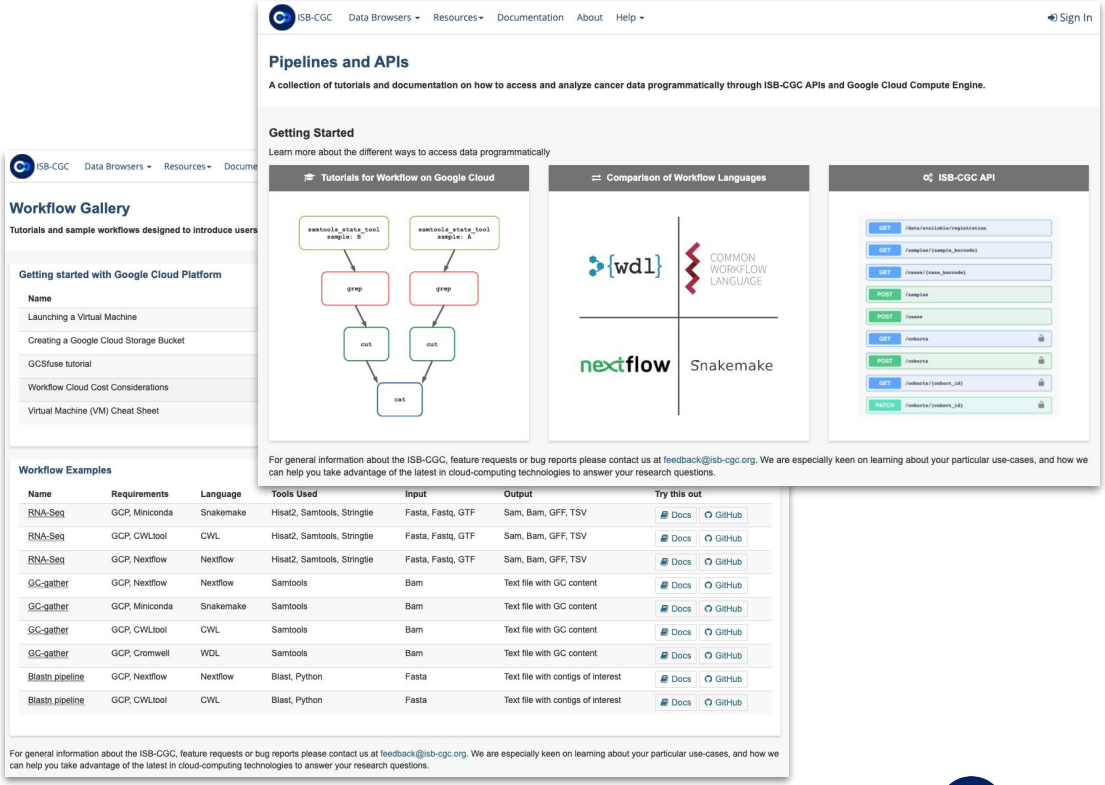
1 GB network egress from North America to all region destinations per month (excluding China and Australia)

ISB-CGC

We have several workflow and pipeline tutorials at ISB-CGC

Tutorials and sample workflows designed to introduce and enable users on how to run workflows in CWL, Nextflow, Snakemake and WDL on GCP

https://isb-cgc.appspot.com/programmatic_access/



ISB-CGC

Example RNASeq quantification workflow using a Docker image

STAR_alignment_run.sh — Desktop Add License

```
wget https://api.gdc.cancer.gov/data/25aa497c-e615-4cb7-8751-71f744f9691f \
    $path_to_reference/gencode.v22.annotation.gtf.gz
gunzip gencode.v22.annotation.gtf.gz

path_to_reference=/home/fseidl/genome
mkdir $path_to_reference/star_index_oh75

sudo docker run --rm -v $path_to_reference:/data -t broadinstitute/gtex_rnaseq:V8 \
    /bin/bash -c "STAR \
        --runMode genomeGenerate \
        --genomeDir /data/star_index_oh75 \
        --genomeFastaFiles /data/GRCh38.d1.vd1.fa \
        --sjdbGTFfile /data/gencode.v22.annotation.gtf \
        --sjdbOverhang 75 \
        --runThreadN 4"

sudo docker run --rm -v $path_to_reference:/data -t broadinstitute/gtex_rnaseq:V8 \
    /bin/bash -c "rsem-prepare-reference \
        /data/GRCh38.d1.vd1.fa \
        /data/rsem_reference \
        --gtf /data/gencode.v22.annotation.gtf \
        --num-threads 4"

path_to_data=/home/fseidl/bams
input_bam=HG00182.mapped.ILLUMINA.bwa.FIN.low_coverage.20120522.bam
sample_id=HG00182
mkdir $path_to_data

gsutil cp gs://genomics-public-data/1000-genomes/bam/$input_bam $path_to_data

sudo docker run --rm -v $path_to_data:/data -t broadinstitute/gtex_rnaseq \
    /bin/bash -c "/src/run_SamToFastq.py /data/$input_bam -p $sample_id -o /data"

# STAR alignment
sudo docker run --rm -v $path_to_data:/data -v $path_to_reference:/genome -t broadinstitute/gtex_rnaseq:V8 \
    /bin/bash -c "/src/run_STAR.py \
        /genome/star_index_oh75 \
        /data/${sample_id}_1.fastq.gz \
        /data/${sample_id}_2.fastq.gz \
        ${sample_id} \
        --threads 4 \
        --output_dir /tmp/star_out && mv /tmp/star_out /data/star_out"
```

Line: 31:30 Shell Script (Bash) Soft Tabs: 4

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

ISB-CGC

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
```

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.

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.

Built for batch processing

Preemptible VMs for affordable batch processing on fault-tolerant workloads to save you time and money.

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.

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.

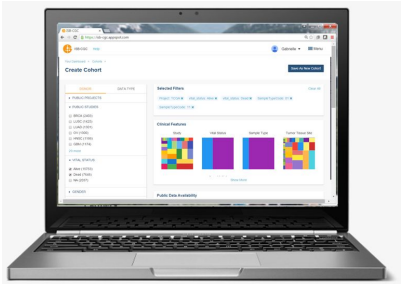
Ready for AI / ML

Bring your data closer to public datasets and advanced analytics that come along with GCP.

https://cloud.google.com/life-sciences

ISB-CGC

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



BigQuery

ISB-CGC hosted tables

User Tables

Cloud Life Sciences

Cloud Storage

Cloud-hosted cancer data

User Data

Compute Engine

VM

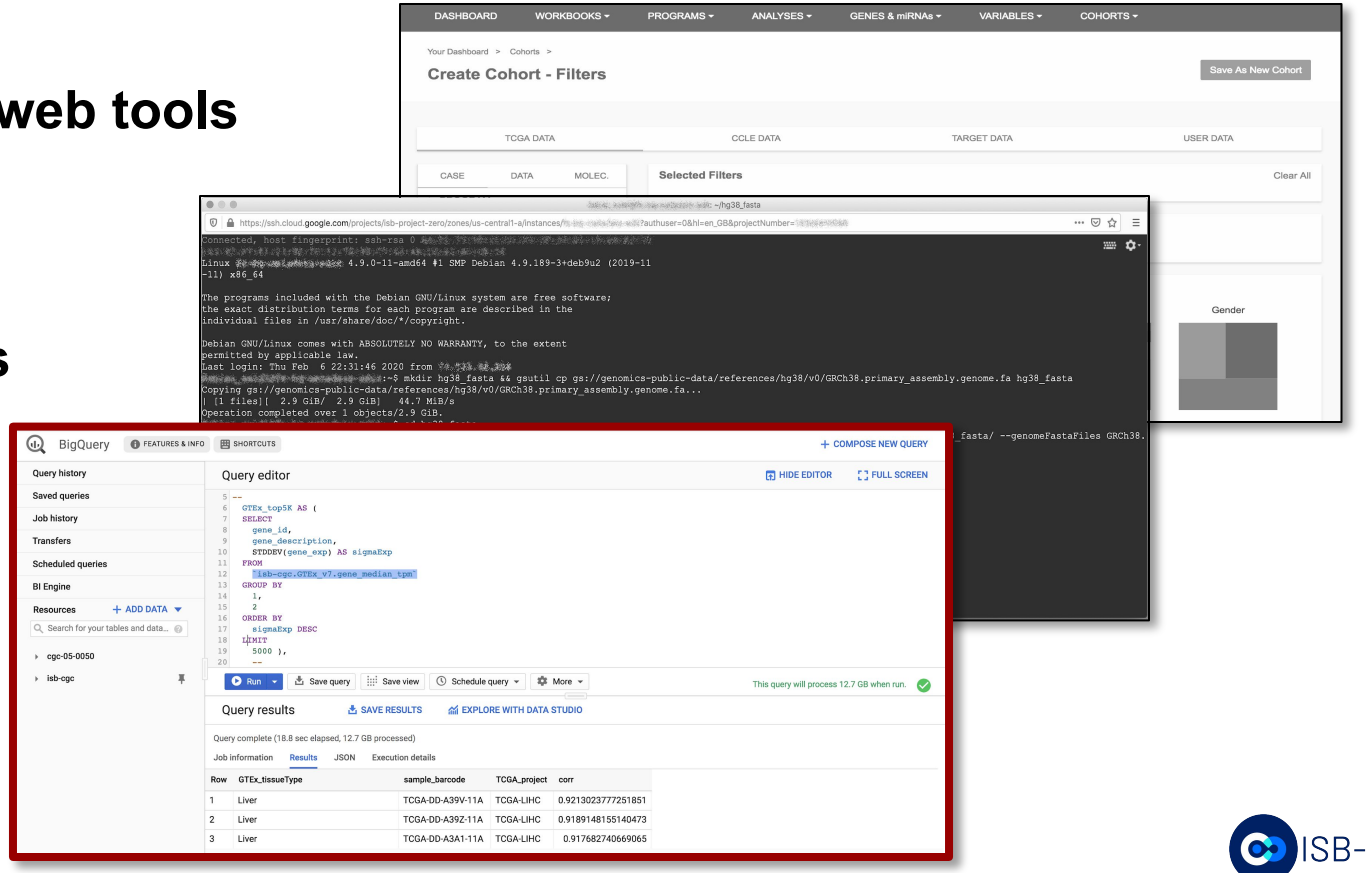
ISB-CGC

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC web tools

Google VMs

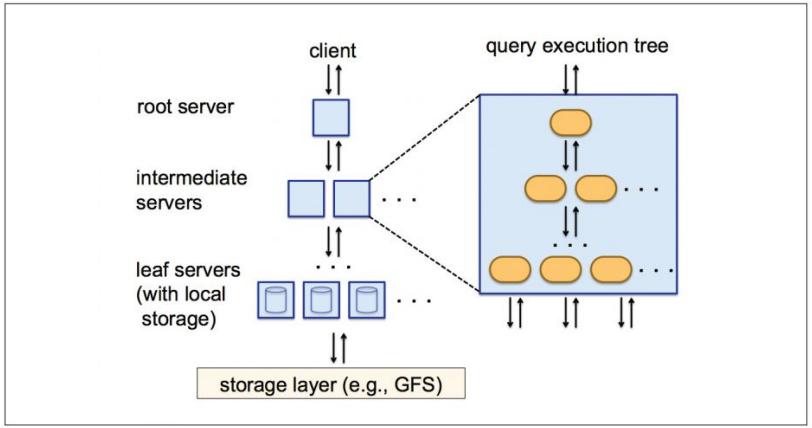
BigQuery



ISB-CGC

Attributes of Google BigQuery that make it ideal for cancer data analytics

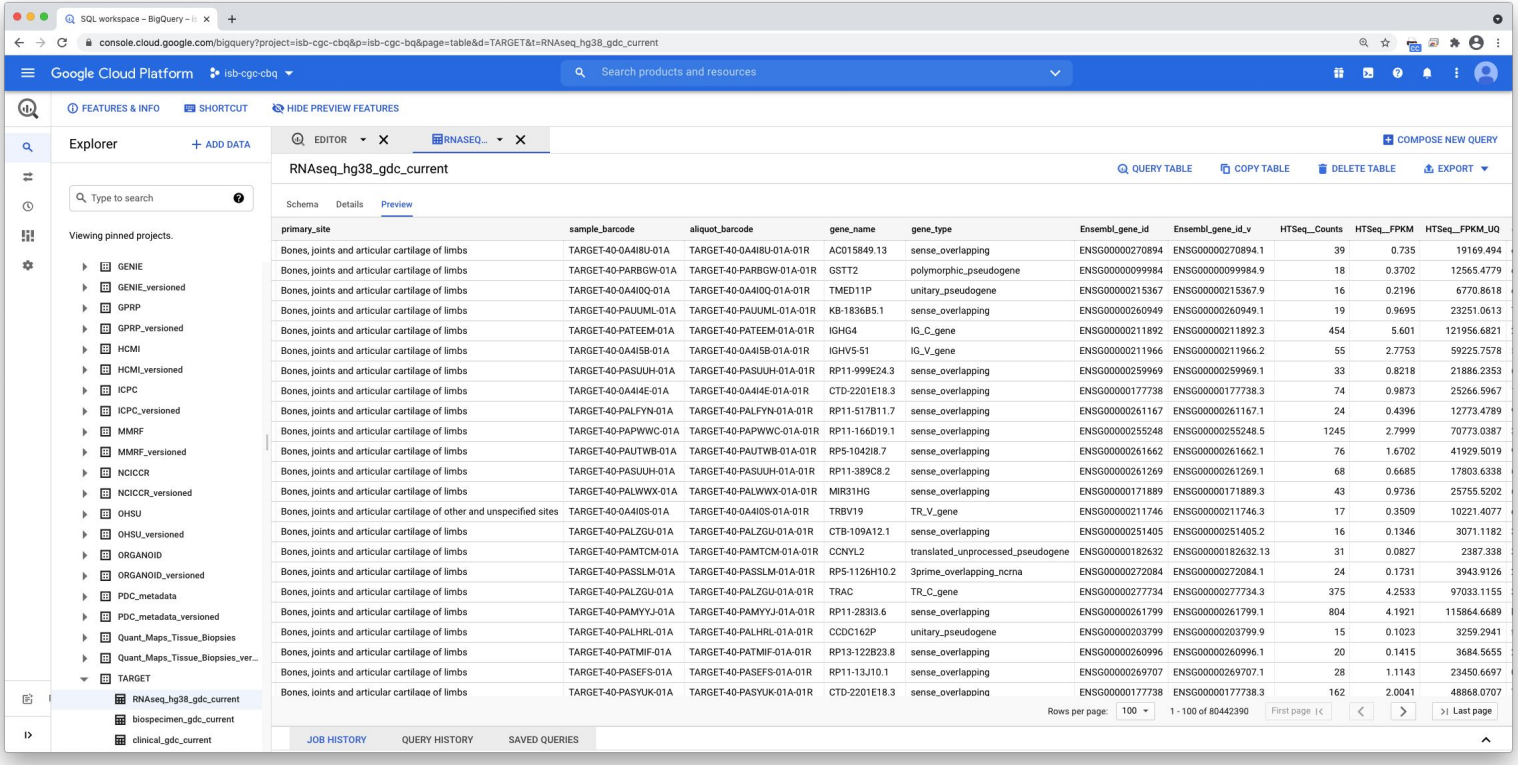
- Columnar database ideal for storing tabular data (RNA sequencing etc.)
- Query speed is automatically scaled by multiprocessing
- Powerful SQL language interface, including user defined functions
- Can join tables based on shared variables
- ISB-CGC combines data of a similar type into single BigQuery tables
  - For example: ~150 individual MAF files were combined to generate a single table



Tree architecture of Dremel

ISB-CGC

You can think of individual BigQuery tables very simply as a tabular database



primary_site	sample_barcode	aliquot_barcode	gene_name	gene_type	Ensembl_gene_id	Ensembl_gene_id_v	HTSeq__Counts	HTSeq__FPKM	HTSeq__FPKM_UQ
Bones, joints and articular cartilage of limbs	TARGET-40-0A4I8U-01A	TARGET-40-0A4I8U-01A-01R	AC015849.13	sense_overlapping	ENSG00000270894	ENSG00000270894.1	39	0.735	19169.494
Bones, joints and articular cartilage of limbs	TARGET-40-PARBGW-01A	TARGET-40-PARBGW-01A-01R	GSTT2	polymorphic_pseudogene	ENSG00000099984	ENSG00000099984.9	18	0.3702	12565.4779
Bones, joints and articular cartilage of limbs	TARGET-40-0A4I0Q-01A	TARGET-40-0A4I0Q-01A-01R	TMED11P	unitary_pseudogene	ENSG00000215367	ENSG00000215367.9	16	0.2196	6770.8618
Bones, joints and articular cartilage of limbs	TARGET-40-PAUUML-01A	TARGET-40-PAUUML-01A-01R	KB-1836B5.1	sense_overlapping	ENSG00000260949	ENSG00000260949.1	19	0.9695	23251.0613
Bones, joints and articular cartilage of limbs	TARGET-40-PATEEM-01A	TARGET-40-PATEEM-01A-01R	IGHG4	IG_C_gene	ENSG00000211892	ENSG00000211892.3	454	5.601	121956.6821
Bones, joints and articular cartilage of limbs	TARGET-40-0A4I5B-01A	TARGET-40-0A4I5B-01A-01R	IGHV5-51	IG_V_gene	ENSG00000211966	ENSG00000211966.2	55	2.7753	59225.7578
Bones, joints and articular cartilage of limbs	TARGET-40-PASUUH-01A	TARGET-40-PASUUH-01A-01R	RP11-999E24.3	sense_overlapping	ENSG00000259969	ENSG00000259969.1	33	0.8218	21886.2353
Bones, joints and articular cartilage of limbs	TARGET-40-0A4I4E-01A	TARGET-40-0A4I4E-01A-01R	CTD-2201E18.3	sense_overlapping	ENSG00000177738	ENSG00000177738.3	74	0.9873	25266.5967
Bones, joints and articular cartilage of limbs	TARGET-40-PALFYN-01A	TARGET-40-PALFYN-01A-01R	RP11-517B11.7	sense_overlapping	ENSG00000261167	ENSG00000261167.1	24	0.4396	12773.4789
Bones, joints and articular cartilage of limbs	TARGET-40-PAPWWC-01A	TARGET-40-PAPWWC-01A-01R	RP11-166D19.1	sense_overlapping	ENSG00000255248	ENSG00000255248.5	1245	2.7999	70773.0387
Bones, joints and articular cartilage of limbs	TARGET-40-PAUTWB-01A	TARGET-40-PAUTWB-01A-01R	RP5-1042I8.7	sense_overlapping	ENSG00000261662	ENSG00000261662.1	76	1.6702	41929.5019
Bones, joints and articular cartilage of limbs	TARGET-40-PASUUH-01A	TARGET-40-PASUUH-01A-01R	RP11-389C8.2	sense_overlapping	ENSG00000261269	ENSG00000261269.1	68	0.6685	17803.6338
Bones, joints and articular cartilage of limbs	TARGET-40-PALWWX-01A	TARGET-40-PALWWX-01A-01R	MIR31HG	sense_overlapping	ENSG00000171889	ENSG00000171889.3	43	0.9736	25755.5202
Bones, joints and articular cartilage of other and unspecified sites	TARGET-40-0A4I0S-01A	TARGET-40-0A4I0S-01A-01R	TRBV19	TR_V_gene	ENSG00000211746	ENSG00000211746.3	17	0.3509	10221.4077
Bones, joints and articular cartilage of limbs	TARGET-40-PALZGU-01A	TARGET-40-PALZGU-01A-01R	CTB-109A12.1	sense_overlapping	ENSG00000251405	ENSG00000251405.2	16	0.1346	3071.1182
Bones, joints and articular cartilage of limbs	TARGET-40-PAMTCM-01A	TARGET-40-PAMTCM-01A-01R	CCNYL2	translated_unprocessed_pseudogene	ENSG00000182632	ENSG00000182632.13	31	0.0827	2387.338
Bones, joints and articular cartilage of limbs	TARGET-40-PASSLM-01A	TARGET-40-PASSLM-01A-01R	RP5-1126H10.2	3prime_overlapping_ncrna	ENSG00000272084	ENSG00000272084.1	24	0.1731	3943.9126
Bones, joints and articular cartilage of limbs	TARGET-40-PALZGU-01A	TARGET-40-PALZGU-01A-01R	TRAC	TR_C_gene	ENSG00000277734	ENSG00000277734.3	375	4.2533	97033.1155
Bones, joints and articular cartilage of limbs	TARGET-40-PAMYYJ-01A	TARGET-40-PAMYYJ-01A-01R	RP11-283I3.6	sense_overlapping	ENSG00000261799	ENSG00000261799.1	804	4.1921	115864.6689
Bones, joints and articular cartilage of limbs	TARGET-40-PALHRL-01A	TARGET-40-PALHRL-01A-01R	CCDC162P	unitary_pseudogene	ENSG00000203799	ENSG00000203799.9	15	0.1023	3259.2941
Bones, joints and articular cartilage of limbs	TARGET-40-PATMIF-01A	TARGET-40-PATMIF-01A-01R	RP13-122B23.8	sense_overlapping	ENSG00000260996	ENSG00000260996.1	20	0.1415	3684.5655
Bones, joints and articular cartilage of limbs	TARGET-40-PASEFS-01A	TARGET-40-PASEFS-01A-01R	RP11-13JJ10.1	sense_overlapping	ENSG00000269707	ENSG00000269707.1	28	1.1143	23450.6697
Bones, joints and articular cartilage of limbs	TARGET-40-PASYUK-01A	TARGET-40-PASYUK-01A-01R	CTD-2201E18.3	sense_overlapping	ENSG00000177738	ENSG00000177738.3	162	2.0041	48868.0707

ISB-CGC

The unique benefit is that you can mine these data quickly and cheaply

RUN MORE SAVE

This query will process 33 GiB when run.

```
1  select avg(HTSeq__FPKM_UQ) as average_fpkm
2  from `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current`
3  where aliquot_barcode = 'TCGA-3X-AAV9-01A-72R-A41I-07'
4  and gene_type = 'protein_coding'
5
```

Query results SAVE RESULTS EXPLORE DATA

Query complete (1.9 sec elapsed, 33 GB processed)

Job information Results JSON Execution details

Row	average_fpkm
1	342956.0106327599

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

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

Query history
Saved queries
Job history
Transfers
Scheduled queries
BI Engine
Resources + ADD DATA
Search for your tables and data...
cgc-05-0050
isb-cgc

Query editor | HIDE EDITOR | FULL SCREEN

```
--
  GTEx_top5K AS (
  SELECT
    gene_id,
    gene_description,
    STDDEV(gene_exp) AS sigmaExp
  FROM
    `isb-cgc.GTEx_v7.gene_median_tpm`
  GROUP BY
    1,
    2
  ORDER BY
    sigmaExp DESC
  LIMIT
    5000 ),
    --
```

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

This query will process 12.7 GB when run.

Query results | SAVE RESULTS | EXPLORE WITH DATA STUDIO

Query complete (18.8 sec elapsed, 12.7 GB processed)

Job information | Results | JSON | Execution details

Row	GTEx_tissueType
1	Liver
2	Liver
3	Liver

Query complete (18.8 sec elapsed, 12.7 GB processed)

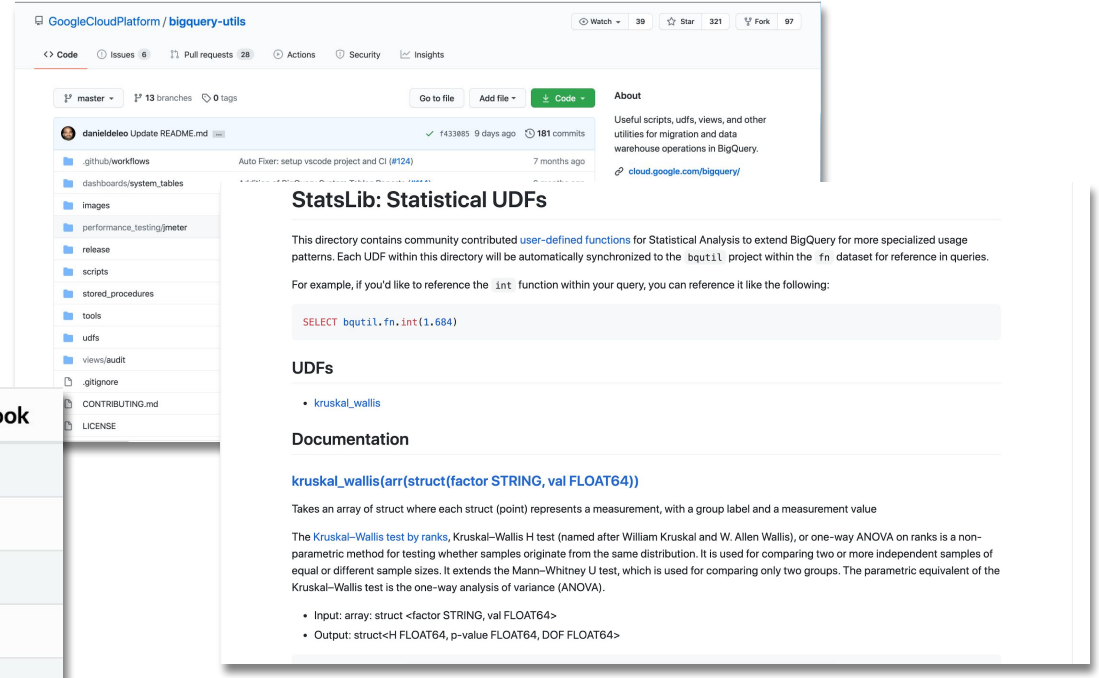
Job information | Results | JSON | Execution details

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

ISB-CGC

Custom User Defined Functions allow users to expand beyond the default functionality

- We are contributing commonly used statistical methods as UDFs to the Google BQ StatsLib
- Users will be able to conduct the following statistical analyses directly within BigQuery



Data type	Data type	Statistical test/notebook
Gene expression	Clinical	Kruskal-Wallis score
Gene expression	Somatic mutation	T-test score
Gene expression	Gene expression	Spearman Correlation
Somatic mutation	Clinical	Chi Square test
Somatic mutation	Somatic Mutation	Fisher's exact test

ISB-CGC

Google Cloud Platform Free Tier lets you perform introductory queries at no cost

DATA ANALYTICS

BigQuery

1 TB

Queries per month

Fully managed, petabyte scale, analytics data warehouse.

1 TB of querying per month

10 GB of storage

COMPUTE
ud Run
million
uests per month
y managed environment to run stateless
ainers.

DATABASE
Firestore
1 GB
Storage
Scalable NoSQL document database.

COMPUTE
Compute Engine
1
F1-micro instance per month
Scalable, high-performance virtual machines.

AGE
ud Storage
GB
ths regional storage
n-class performance, reliability, and
g 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.

UTE
ogle Kubernetes Engine
lusters
ize clusters
click container orchestration via
netes 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.

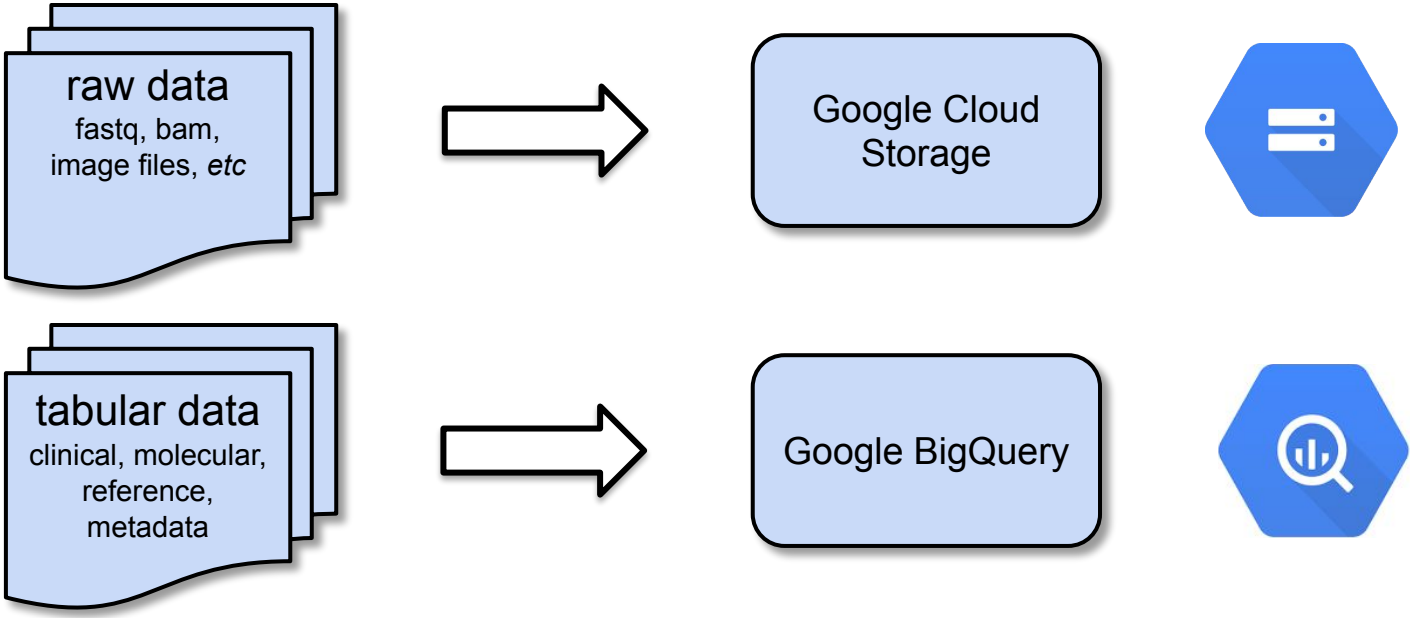
ANALYTICS
Query
TB
Queries per month
Fully managed, petabyte scale, analytics data warehouse.

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.

ISB-CGC

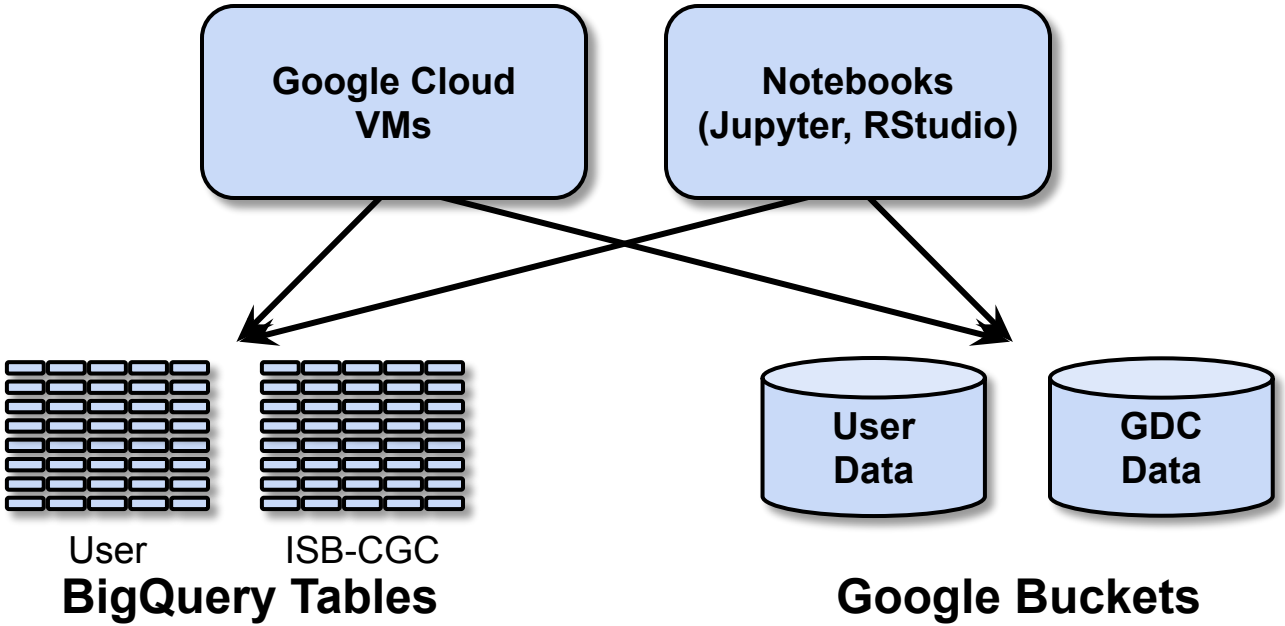
How do users access data on ISB-CGC?



ISB-CGC

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



ISB-CGC

In total, ISB-CGC hosts multiple TBs of cancer research data across 440 open-access BigQuery tables

- TCGA, TARGET, CCLE:
  - Clinical & Biospecimen information
  - Molecular data
    - mRNA, miRNA, protein expression
    - DNA copy-number
    - DNA methylation
    - Somatic mutations
- Reference information:
  - Gencode, Ensembl, miRBase, …
  - hg19 to hg38 'liftover'
  - 450K DNA methylation annotations
  - Kaviar, COSMIC, …
  - GDC metadata
- Other publicly available resources:
  - Google-hosted public datasets
  - Tute Genomics (ANNOVAR) annotations
- ***Your*** data!

Google BigQuery

ISB-CGC

This is potentially overwhelming for data discovery, we have a table search UI to help

BigQuery Table Search

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

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.

Status
CURRENT

Name

Program
Choose Programs...

Category
- CLINICAL BIOSPECIMEN DATA
- FILE METADATA
- GENOMIC REFERENCE DATABASE
- PROCESSED -OMICS DATA

Reference Genome
ALL

Source
Choose Sources...

Data Type
Choose Data Types...

Experimental Strategy
Choose 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		
CCLE 2016 - COPY NUMBER SEGMENTS	CCLE	PROCESSED -OMICS DATA	BROAD	COPY NUMBER SEGMENT	CURRENT	760,192	2/27/2016		
CCLE 2016 - FASTQC METRICS	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,249	3/28/2016		
CCLE 2016 - FILE METADATA	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,915	3/29/2016		
CCLE 2016 - SAMPLE INFORMATION	CCLE	PROCESSED -OMICS DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	929	2/26/2016		
CCLE 2016 - SOMATIC MUTATION	CCLE	PROCESSED -OMICS 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.

https://isb-cgc.appspot.com/bq_meta_search/

ISB-CGC

Table schemas are easily accessible through this UI

BigQuery Table Search

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

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.

Status
CURRENT

Name

Program
Choose Programs...

Category
- CLINICAL BIOSPECIMEN DATA
- FILE METADATA
- GENOMIC REFERENCE DATABASE
- PROCESSED -OMICS DATA

Reference Genome
ALL

Source
Choose Sources...

Data Type
Choose Data Types...

Experimental Strategy
Choose 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		
CCLE 2016 - COPY NUMBER SEGMENTS	CCLE	PROCESSED -OMICS DATA	BROAD	COPY NUMBER SEGMENT	CURRENT	760,192	2/27/2016		
CCLE 2016 - FASTQC METRICS	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,249	3/28/2016		
CCLE 2016 - FILE METADATA	CCLE	PROCESSED -OMICS DATA	BROAD	FILE METADATA	CURRENT	1,915	3/29/2016		
CCLE 2016 - SAMPLE INFORMATION	CCLE	PROCESSED -OMICS DATA	BROAD	BIOSPECIMEN SUPPLEMENT	CURRENT	929	2/26/2016		
CCLE 2016 - SOMATIC MUTATION	CCLE	PROCESSED -OMICS 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		

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. TC-71, NIH:OVCAR-3, 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(CN/2) mean value estimate

Labels access : open | data_type : copy_number_segment | program : ccle | reference_genome_0 : hg19 | source : broad | category : processed_-omics_data | status : current

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.

https://isb-cgc.appspot.com/bq_meta_search/

ISB-CGC

Benefits of the ISB-CGC BigQuery Table Search

- 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 of interested? An “open” button takes users directly to the GCP BigQuery Console.

ISB-CGC

Use Google BigQuery to easily connect your research to public datasets

ISB-CGC and Other
Public Datasets

↓

SQL

↑

Private User Data
and Derived Results

ISB-CGC

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

BigQuery interface for building and experimenting with queries

Jupyter notebook to visualize and share analysis

BigQuery BETA | Go to Classic UI

Query history
Saved queries
Job history
Transfers
Resources + ADD DATA
Search for your tables and datasets
- isb-project-zero
- isb-cgc
- isb-cgc-02-0001

Query editor

```
select t.case_barcode, t.project_short_name,t.age_at_diagnosis,s.dx_age
from `isb-project-zero.SEER.SEER_view02`  as s
JOIN `isb-project-zero.SEER.SEER_view01` as x ON s.patient_id_number = x.patie
JOIN `isb-cgc.TCGA_bioclin_v0.Clinical` as t ON x.icd_10_code = t.icd_10
```

Processing location: US

Run | Save query | Save view | More

Query results | SAVE RESULTS | EXPLORE IN DATA STUDIO

Query complete (0.015 sec elapsed, cached)

Job information | Results | JSON | Execution details

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-A4I1	TCGA-THCA	37	50
23004	TCGA-FK-A3S3	TCGA-THCA	43	84
23005	TCGA-DE-A4MC	TCGA-THCA	43	50
23006	TCGA-EL-A3GO	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

CODE | TEXT | CELL | CELL

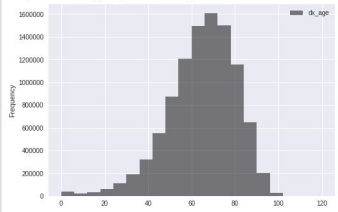
```
%%bigquery --project isb-project-zero df
select * from `isb-project-zero.SEER.SEER_view02` where dx_year >=2000 and dx_age <=40 LIMIT 10
```

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

```
df = pd.io.gbq.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)
```

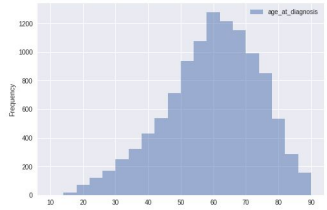
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0871bdc9e8>
```




```
df2 = pd.io.gbq.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)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f086f1d6f28>
```



ISB-CGC

BigQuery integrates with a variety of commonly used analysis tools

bigrquery and
bigQueryR

googleAuthR

Pre-built VM images

IP[y]:
IPython

Cloud notebooks
and workspaces.

Cloud Datalab

ISB-CGC

Interactive step-by-step guide on BigQuery data analysis

Cancer Data Discovery in the Cloud through ISB-CGC

1. Introduction

ISB-CGC

ISB-CGC, one of the National Cancer Institute's Cloud Resources, uniquely hosts cancer data including somatic mutations, copy number variations, gene and, protein expressions, etc. from widely used cancer datasets including TCGA, TARGET and many more in Google BigQuery.

Google BigQuery is a massively-parallel analytics engine ideal for tabular data. ISB-CGC has combined data scattered over tens of thousands of files into easily accessible BigQuery tables. This novel approach allows our users to quickly analyze data from thousands of patients in ISB-CGC curated BigQuery tables.

- Users can explore and learn more about the ISB-CGC hosted BigQuery tables via an interactive BigQuery Table Search User Interface. Users can find tables of interest based on category, reference genome build, data type and free-form text search.
- Users with Google Cloud Platform (GCP) projects can analyze patient, biospecimen, and molecular data from many NCI funded programs such as TCGA, TARGET, CCLE, GTEx all in ISB-CGC's BigQuery tables. SQL queries can be executed in many ways including through the Google Cloud BigQuery web console, Jupyter notebooks, and R scripts.

In this tutorial, you will:

- Analyze gene expression and protein abundance differences between two types of TCGA kidney cancers, **Kidney Renal Papillary Carcinoma (TCGA-KIRP)** and **Kidney Chromophobe (TCGA-KICH)**.
- Build a cohort of patients with these cancer types and extract their respective gene expression and protein abundance data from ISB-CGC TCGA Google BigQuery tables.
- Connect to Google BigQuery tables from within R for data analysis and visualization

What you'll learn:

- **How to explore the ISB-CGC BigQuery Table Search User Interface**
- **How to build and run queries in the Google BigQuery Console**
- **How to use R notebooks for data analysis and visualization**
- **How to use Bioconductor packages on data in ISB-CGC BigQuery tables**

Next

Report a mistake

https://isb-cgc.appspot.com/how_to_discover/#0

ISB-CGC

Some examples of analyses run with our tools

BigQuery can be used to join tables of different data types

```
with gexp as
    (SELECT
        project_short_name,
        case_barcode,
        gene_name,
        avg(HTSeq__FPKM) as mean_gexp
    FROM `isb-cgc.TCGA_hg38_data_v0.RNAseq_Gene_Expression`
    WHERE
        project_short_name in ('TCGA-KIRP', 'TCGA-KICH')
        AND gene_type = 'protein_coding'
    GROUP BY
        project_short_name, case_barcode, gene_name
),
pexp as
    (SELECT
        project_short_name,
        case_barcode, gene_name,
        avg(protein_expression) as mean_pexp
    FROM `isb-cgc.TCGA_hg38_data_v0.Protein_Expression`
    WHERE
        project_short_name in ('TCGA-KIRP', 'TCGA-KICH')
    GROUP BY
        project_short_name, case_barcode, gene_name
)
SELECT
    gexp.project_short_name,
    gexp.case_barcode,
    gexp.gene_name,
    gexp.mean_gexp,
    pexp.mean_pexp
FROM gexp
    inner join pexp
    on
    gexp.project_short_name = pexp.project_short_name
    AND gexp.case_barcode = pexp.case_barcode
    AND gexp.gene_name = pexp.gene_name
```

Query results SAVE RESULTS EXPLORE DATA

Query complete (3.6 sec elapsed, 36.5 GB processed)

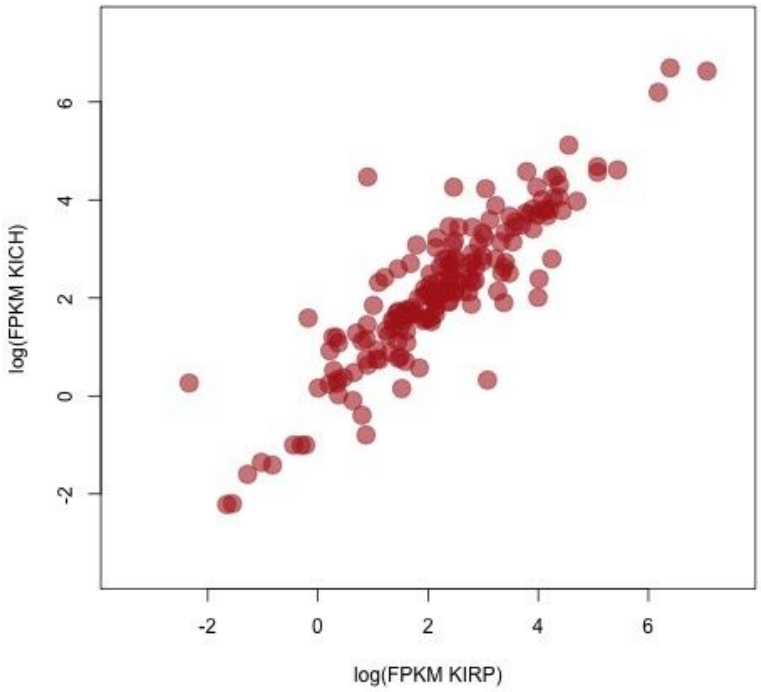
Job information Results JSON Execution details

Row	project_short_name	case_barcode	gene_name	mean_gexp	mean_pexp
1	TCGA-KIRP	TCGA-BQ-5882	CASP9	4.36721774268	-0.77899630325
2	TCGA-KIRP	TCGA-BQ-5893	CASP9	3.51383227626	-0.00487536525000024
3	TCGA-KIRP	TCGA-UZ-A9PV	CASP9	5.18471681617	0.608095901
4	TCGA-KIRP	TCGA-HE-7130	CASP9	9.66588010427	1.26419392825
5	TCGA-KIRP	TCGA-BQ-7051	CASP9	4.960828359680001	0.16363261725
6	TCGA-KIRP	TCGA-B9-A5W7	CASP9	3.48913273215	0.4692597255
7	TCGA-KIRP	TCGA-B9-A69E	CASP9	2.31606330472	-0.79851830675
8	TCGA-KIRP	TCGA-IA-A83W	CASP9	7.0151441055	0.0390634209999998
9	TCGA-KIRP	TCGA-KV-A74V	CASP9	3.6511469304	0.58510363975
10	TCGA-KIRP	TCGA-DW-7963	CASP9	2.72821430732	-0.0676392402500001
11	TCGA-KIRP	TCGA-B1-A47O	CASP9	4.38451034062	-0.30207994575
12	TCGA-KIRP	TCGA-SX-A7SM	CASP9	5.19366866041	-0.12840562825
13	TCGA-KIRP	TCGA-BQ-7061	CASP9	5.52320872773	-0.4549548795
14	TCGA-KIRP	TCGA-AL-A5DJ	CASP9	3.94040060142	-0.22713179975
15	TCGA-KIRP	TCGA-B9-A8YH	CASP9	4.40211877405	0.307956009
16	TCGA-KIRP	TCGA-J7-A8I2	CASP9	5.40615807819	0.9955787775
17	TCGA-KIRP	TCGA-IZ-8196	CASP9	4.76919558442	0.0767987447499998
18	TCGA-KIRP	TCGA-UZ-A9PR	CASP9	5.16376898893	0.41371152025
19	TCGA-KIRP	TCGA-BQ-5875	CASP9	3.445547761625	-0.35423371525
20	TCGA-KIRP	TCGA-SX-A7SS	CASP9	3.32930470316	-0.12148286825
21	TCGA-KIRP	TCGA-DW-7840	CASP9	4.47098482015	-0.1197760605

ISB-CGC

This query can be executed and the results downloaded and plotted from R

```r

# Load data into a data frame
expression_data <- bq_table_download(bq_project_query (project, query = sql_expression))

expression_data$id <- paste(expression_data$project_short_name, expression_data$case_barcode,
sep='.')
cases <- unique(expression_data$id)

# Transform the expression_data data frame, so that columns are samples, rows are genes.
list_exp <- lapply(cases, function(case){
    temp <- expression_data[expression_data$id == case, c('gene_name', 'mean_gexp')]
    names(temp) <- c('gene_name', case)
    return(temp)
})

gene_exps <- Reduce(function(x, y) merge(x, y, all=T, by="gene_name"), list_exp)

exp_p <- gene_exps[,grep('KIRP', names(gene_exps))]
exp_c <- gene_exps[,grep('KICH', names(gene_exps))]
plot(log(rowMeans(exp_p)), log(rowMeans(exp_c)),
    xlab='log(FPKM KIRP)', ylab='log(FPKM KICH)',
    xlim=c(-3.5,7.5), ylim=c(-3.5,7.5), pch=19, cex=2,
    col=rgb(178,34,34,max=255,alpha=150))
```

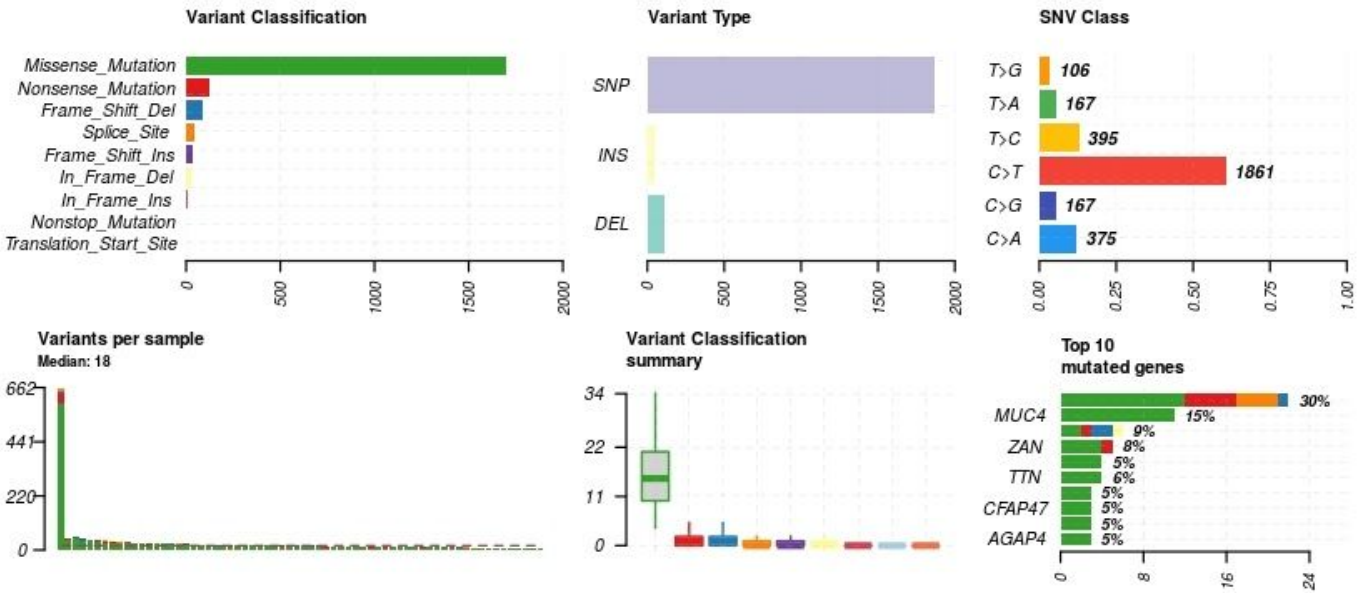


Lots of data types can directly be plugged into Bioconductor pipelines

```r
# Use BigQuery to load TCGA somatic mutation data for our cancers of interest.
sql_kich <- "
SELECT
    Hugo_Symbol,
    Chromosome,
    Start_Position,
    End_Position,
    Reference_Allele,
    Tumor_Seq_Allele2,
    Variant_Classification,
    Variant_Type,
    sample_barcode_tumor
FROM
    `isb-cgc.TCGA_hg38_data_v0.Somatic_Mutation`
WHERE
    project_short_name = 'TCGA-KICH'"

# Put data into a dataframe
maf_kich <- bq_table_download(bq_project_query (project, query = sql_kich))

# Rename column 9 to the field name required by maftools.
colnames(maf_kich)[9] <- "Tumor_Sample_Barcode"

# Convert data frames to maftools objects.
kich <- read.maf(maf_kich)

# Maftools plots
plotmafSummary(maf = kich, rmOutlier = TRUE,
               addStat = 'median', dashboard = TRUE,
               titvRaw = FALSE)
```

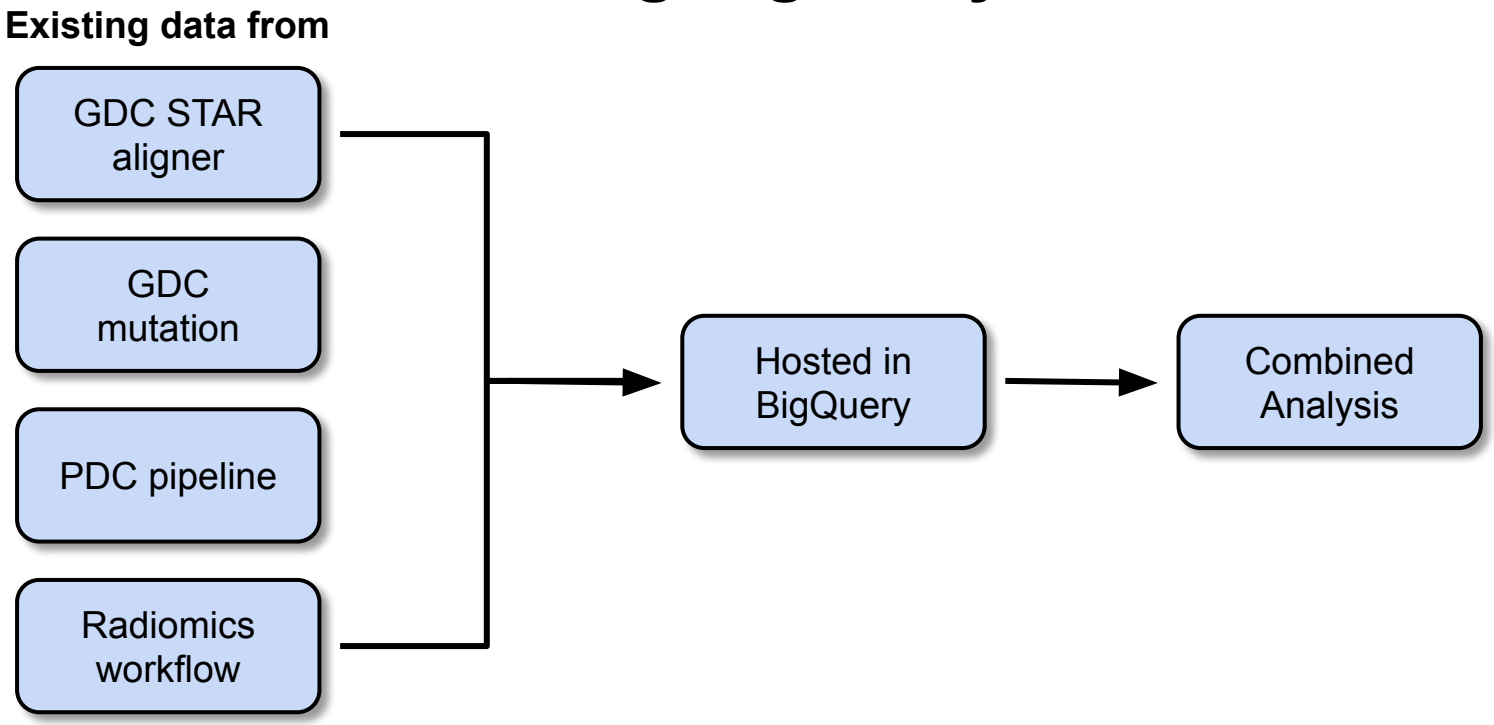
ISB-CGC

A plot of MAF characteristics using the Bioconductor maftools library



ISB-CGC

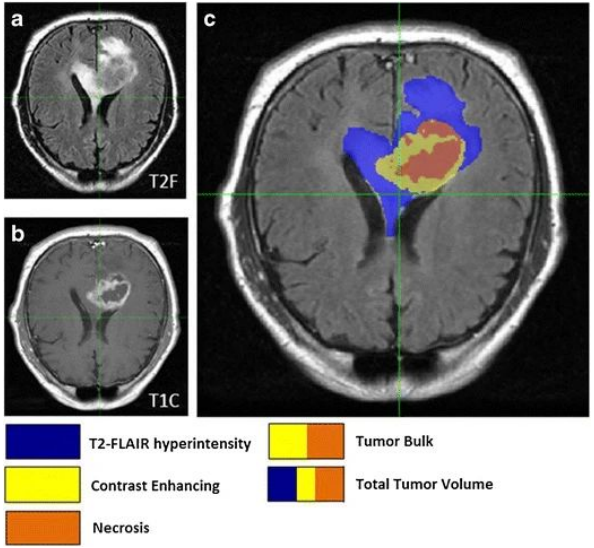
Connecting existing genomics, proteomics, and radiomics data from the respective data commons using BigQuery

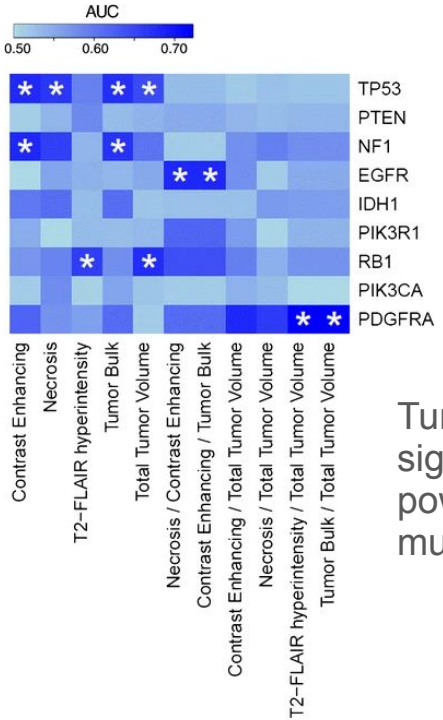


ISB-CGC

Collaboration between ISB-CGC Cloud Resource and IDC

Tumor identification and characterization by machine learning





Tumor features have significant predictive power about the mutations present

Gutman, et al.
Diagnostic Neuroradiology 2015

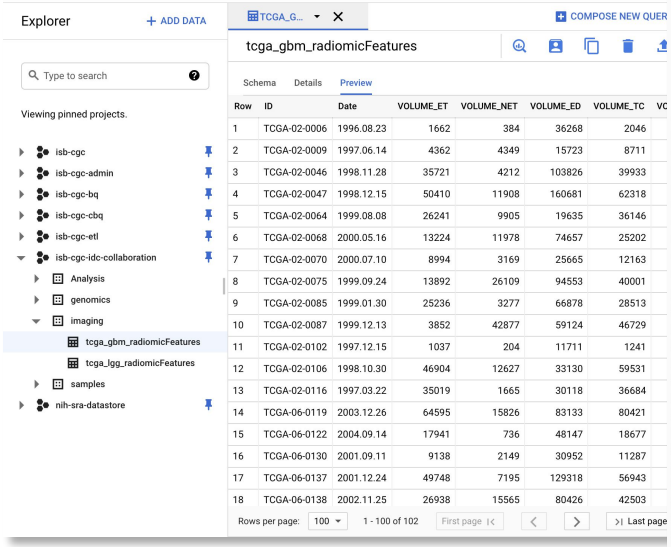
ISB-CGC

We have created workflows to join data and run statistical tests

Using an initial feature set from Bakas, et al. *Nature Scientific Data*, 2017 to pilot analyses

Generated significance values from the correlation between gene expression and tumor features

Subset of somatic mutations for each case id to be used for association analyses



Row	ID	Date	VOLUME_ET	VOLUME_NET	VOLUME_ED	VOLUME_TC	VC
1	TCGA-02-0006	1996.08.23	1662	384	36268	2046	
2	TCGA-02-0009	1997.06.14	4362	4349	15723	8711	
3	TCGA-02-0046	1998.11.28	35721	4212	103826	39933	
4	TCGA-02-0047	1998.12.15	50410	11908	160681	62318	
5	TCGA-02-0064	1999.08.08	26241	9905	19635	36146	
6	TCGA-02-0068	2000.05.16	13224	11978	74657	25202	
7	TCGA-02-0070	2000.07.10	8994	3169	25665	12163	
8	TCGA-02-0075	1999.09.24	13892	26109	94553	40001	
9	TCGA-02-0085	1999.01.30	25236	3277	66878	28513	
10	TCGA-02-0087	1999.12.13	3852	42877	59124	46729	
11	TCGA-02-0102	1997.12.15	1037	204	11711	1241	
12	TCGA-02-0106	1998.10.30	46904	12627	33130	59531	
13	TCGA-02-0116	1997.03.22	35019	1665	30118	36684	
14	TCGA-06-0119	2003.12.26	64595	15826	83133	80421	
15	TCGA-06-0122	2004.09.14	17941	736	48147	18677	
16	TCGA-06-0130	2001.09.11	9138	2149	30952	11287	
17	TCGA-06-0137	2001.12.24	49748	7195	129318	56943	
18	TCGA-06-0138	2002.11.25	26938	15565	80426	42503	

RUN SAVE SCHEDULE MORE

```
#get gene expression data
rnaseq AS (
  select case_barcode, sample_barcode, gene_name, MAX(HTSeq__Counts) gexp
  from
    isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current as rna
  WHERE
    gene_type = 'protein_coding'
    AND project_short_name = 'TCGA-GBM'
  GROUP BY gene_name, case_barcode, sample_barcode
),
# join gene exp with image features
combined AS (
  SELECT gene_name, feature, gexp, value ,
    (RANK() OVER( PARTITION BY gene_name,feature ORDER BY gexp )) + (COUNT(*) OVER
    (RANK() OVER( PARTITION BY gene_name,feature ORDER BY value )) AS ranky
  FROM rnaseq
  JOIN isb-cgc-idc-collaboration.Analysis.unpivoted_tcga_gbm_radiomicFeatures img
  ON img.ID = rnaseq.case_barcode
    AND  STARTS_WITH(img.feature, "VOLUME")
),
# spearman correlation
correlation AS (
  SELECT gene_name, feature, COUNT(gexp) as n, CORR(rankx,ranky)  as corr
  FROM combined
  GROUP BY gene_name, feature
)
#p-value computation
SELECT  feature, gene_name, n, corr,
  `cgc-05-0042.functions.jstat_ttest`(t, n-2, 2) as p
FROM (
  SELECT *,
      ABS(corr)*SQRT( (n-2)/((1-corr*corr))) AS t
  FROM correlation
  WHERE  ABS(corr) < 1.0
    )
ORDER BY p
```

Row	gene	TCGA_DB_5275	TCGA_DH_A66D	TCGA_DH_A66F	TCGA_DU_6392	TCGA_DU_6402
2401	ENSG00000113621	0	0	0	0	0
2402	ENSG00000172508	0	0	0	0	0
2403	ENSG00000165028	0	0	0	0	0
2404	ENSG00000166159	0	0	0	0	0
2405	ENSG00000159212	0	0	0	0	0
2406	ENSG00000061794	0	0	0	0	0
2407	ENSG00000173812	0	1	0	0	0
2408	ENSG00000183011	0	0	0	0	0
2409	ENSG00000165175	0	0	0	0	0
2410	ENSG00000162595	0	0	0	0	0
2411	ENSG00000188846	0	0	0	0	0
2412	ENSG00000143353	0	0	0	0	0
2413	ENSG00000100350	0	0	0	0	0
2414	ENSG00000138382	0	0	0	0	0
2415	ENSG00000141295	0	0	0	0	0
2416	ENSG00000101152	0	0	0	0	0
2417	ENSG00000180483	0	0	0	0	0
2418	ENSG00000143545	0	0	0	0	0
2419	ENSG00000166974	0	0	0	0	0

ISB-CGC

With more effort these cross-omics analyses could lead to novel insights

Using a simple t-test run via SQL in BigQuery we were able to replicate a significant association between a mutated MUC16 gene and an increased tumor volume as measured by radiology.

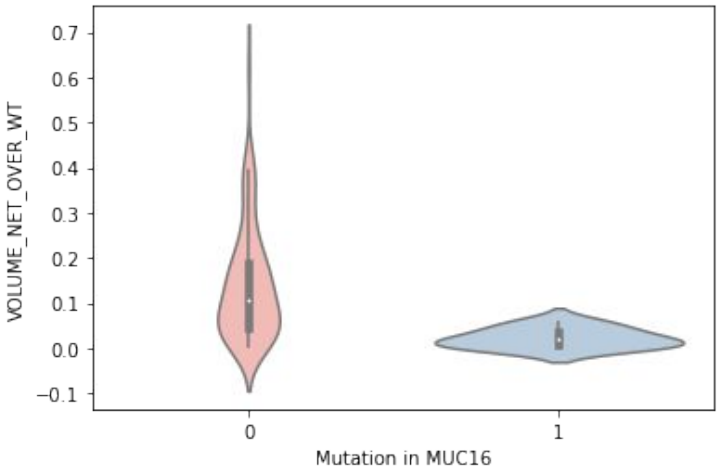
Query results SAVE RESULTS EXPLORE DATA

Query complete (2.5 sec elapsed, 68.2 MB processed)

Job information Results JSON Execution details

Row	Study	radiomic_feature	gene	n1	n0	avg1	avg0	pvalue
1	GBM	VOLUME_NET_OVER_WT	MUC16	6	96	0.02313391	0.139958265	1.5318989464578824E-11
2	GBM	VOLUME_NET_OVER_ED	MUC16	6	96	0.036119709	0.333068168	6.005824166041314E-10
3	GBM	VOLUME_NET_OVER_BRAIN	MUC16	6	96	0.00180942	0.011286297	6.260805815986468E-8
4	GBM	VOLUME_NET	MUC16	6	96	2660	16385.6875	1.2425312760676515E-7
5	GBM	VOLUME_ET_over_TC	MUC16	6	96	0.90486	0.691931177	2.036526115878079E-5
6	GBM	VOLUME_NET_over_TC	MUC16	6	96	0.095138	0.308068906	2.0366175438960117E-5

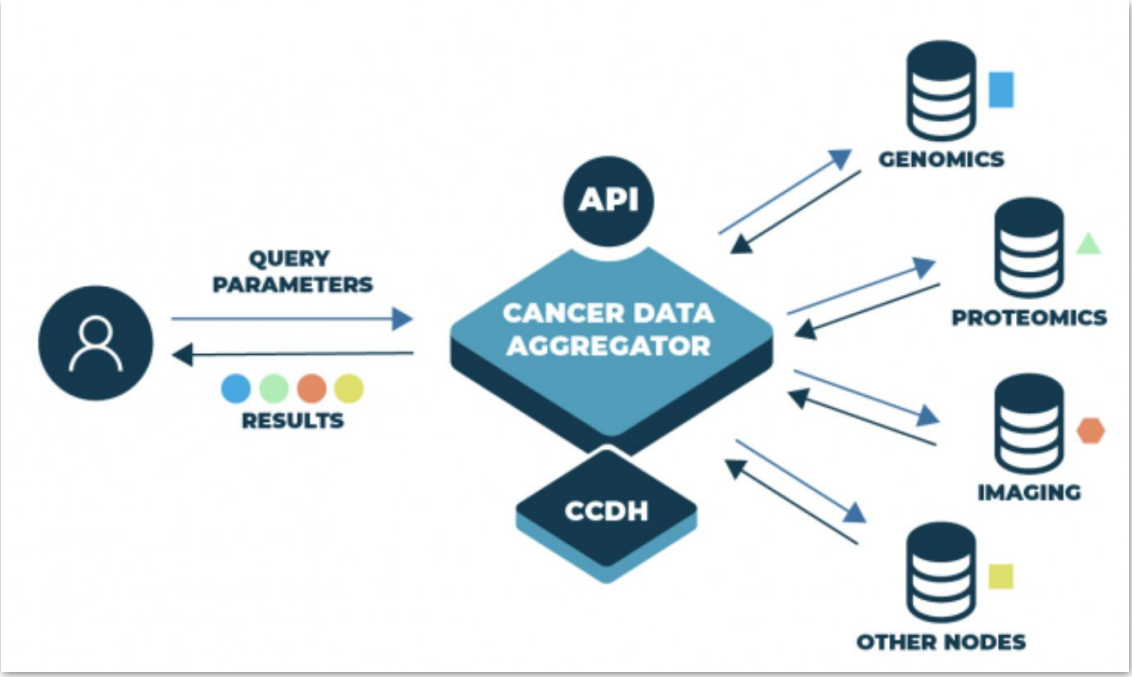
Rows per page: 100 1 - 100 of 210 First page



ISB-CGC

The Cancer Data Aggregator (CDA) is aimed to bridge data discovery between the Data Commons

Federated realtime search across nodes



ISB-CGC

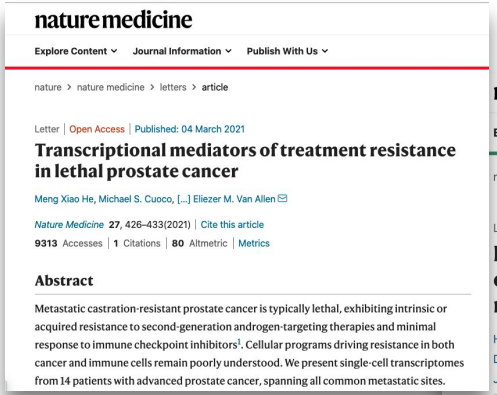
Some more workflows done by ISB-CGC end-users

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)


nature medicine

Explore Content | Journal Information | Publish With Us

nature > nature medicine > letters > article

Letter | Open Access | Published: 04 March 2021

Transcriptional mediators of treatment resistance in lethal prostate cancer

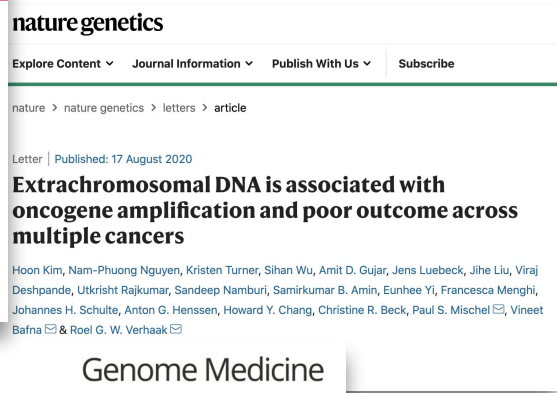
Meng Xiao He, Michael S. Cuoco, [...] Eliezer M. Van Allen

Nature Medicine **27**, 426–433(2021) | Cite this article

9313 Accesses | 1 Citations | 80 Altmetric | Metrics

Abstract

Metastatic castration-resistant prostate cancer is typically lethal, exhibiting intrinsic or acquired resistance to second-generation androgen-targeting therapies and minimal response to immune checkpoint inhibitors[1]. Cellular programs driving resistance in both cancer and immune cells remain poorly understood. We present single-cell transcriptomes from 14 patients with advanced prostate cancer, spanning all common metastatic sites.


nature genetics

Explore Content | Journal Information | Publish With Us | Subscribe

nature > nature genetics > letters > article

Letter | Published: 17 August 2020

Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers

Hoon Kim, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, Viraj Deshpande, Utkrisht Rajkumar, Sandeep Namburi, Samirkumar B. Amin, Eunhee Yi, Francesca Menghi, Johannes H. Schulte, Anton G. Henssen, Howard Y. Chang, Christine R. Beck, Paul S. Mischel, Vineet Bafna & Roel G. W. Verhaak


Oak *et al. Genome Medicine* (2020) 12:51
https://doi.org/10.1186/s13073-020-00744-3

Genome Medicine

RESEARCH | Open Access

Ancestry-specific predisposing germline variants in cancer

Check for updates

Ninad Oak[1,2], Andrew D. Cherniack[3,4], R. Jay Mashl[5,6], TCGA Analysis Network, Fred R. Hirsch[7], Li Ding[5,6,8,9], Rameen Beroukhim[3,4,10], Zeynep H. Gümüş[11,12], Sharon E. Plon[2,13,14] and Kuan-lin Huang[11,12,15*]

with many other manuscripts and grants currently in progress or submitted

ISB-CGC

Resources to help you get started on ISB-CGC

- Our documentation can be found at: https://isb-cancer-genomics-cloud.readthedocs.io/

- Our workflow tutorials can be found at: https://isb-cgc.appspot.com/programmatic_access/

- Finally, a quick guide to BigQuery analysis is at: https://isb-cgc.appspot.com/how_to_discover/#0

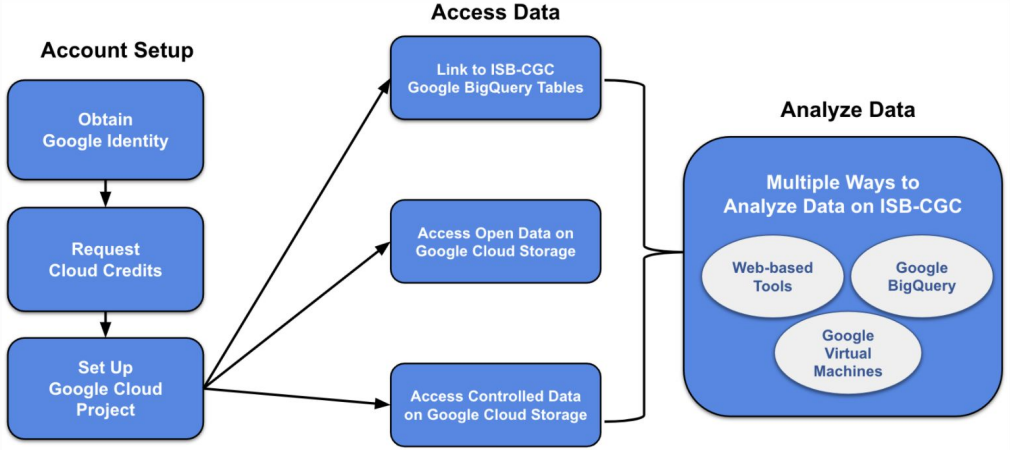
For feedback/suggestions or need help getting starting, contact us at feedback@isb-cgc.org

ISB-CGC

Contact us about setting up your own Google Cloud Platform Project!
feedback@isb-cgc.org

Quick-Start Guide

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. To get started, you'll need a Google Cloud Project. Additionally, to access controlled data, you'll also need dbGaP authorization.



https://isb-cancer-genomics-cloud.readthedocs.io

ISB-CGC

ISB-CGC Team

ISB

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

GENERAL DYNAMICS
Information Technology

David Pot
Madelyn Reyes
Kawther Abdilleh
Fabian Seidl
Deena Bleich
Owais Shahzada

Questions?

ISB-CGC