Cancer -omics analysis on the ISB-CGC platform

2020-03-17

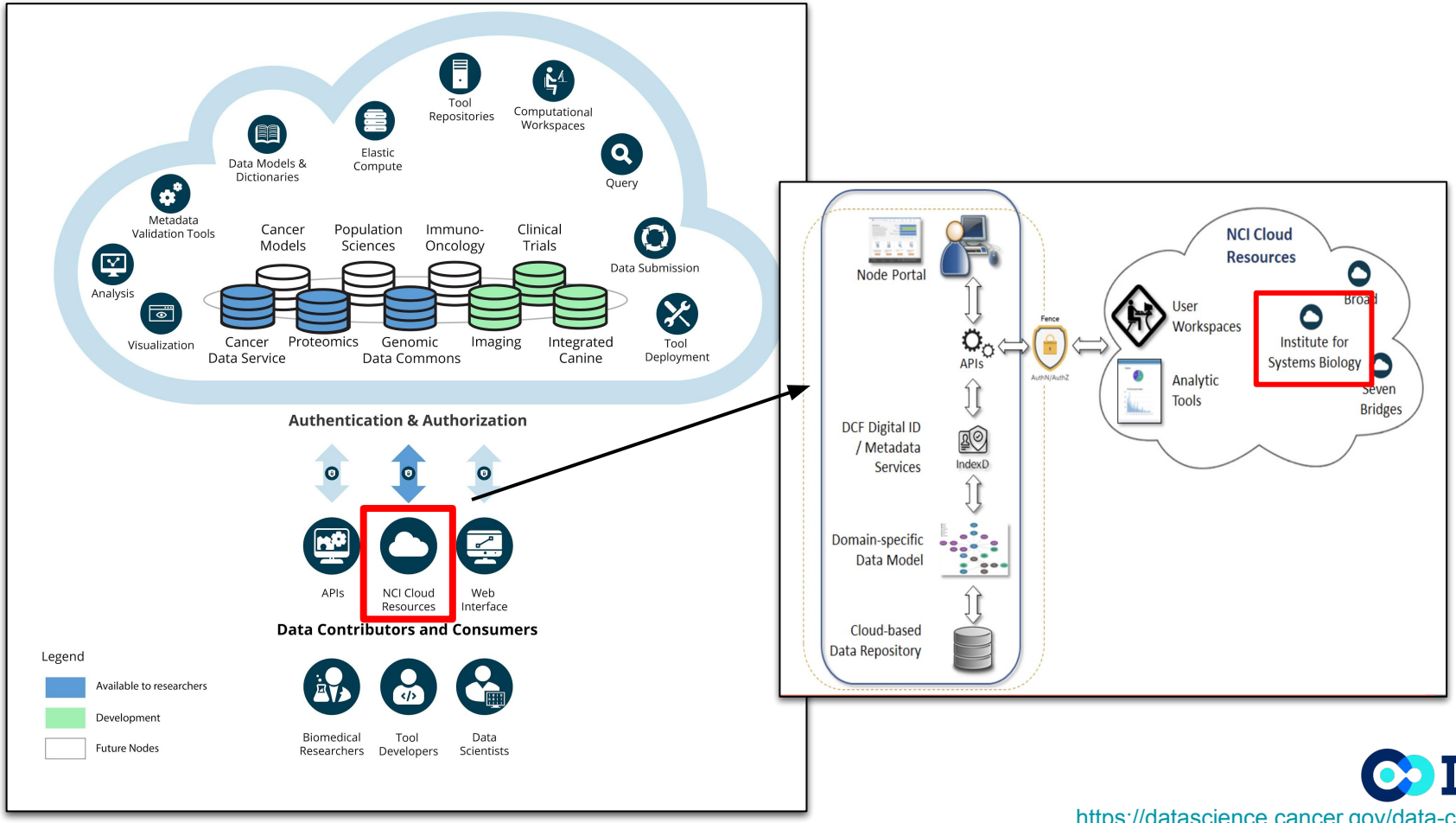
Kawther Abdilleh - Bioinformatics Scientist
Fabian Seidl - Bioinformatics Scientist
Bill Longabaugh - Co-PI ISB-CGC

Outline

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

ISB

NCI Cancer Research Data Commons Ecosystem



https://datascience.cancer.gov/data-commons

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

Data

- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Compute

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Security

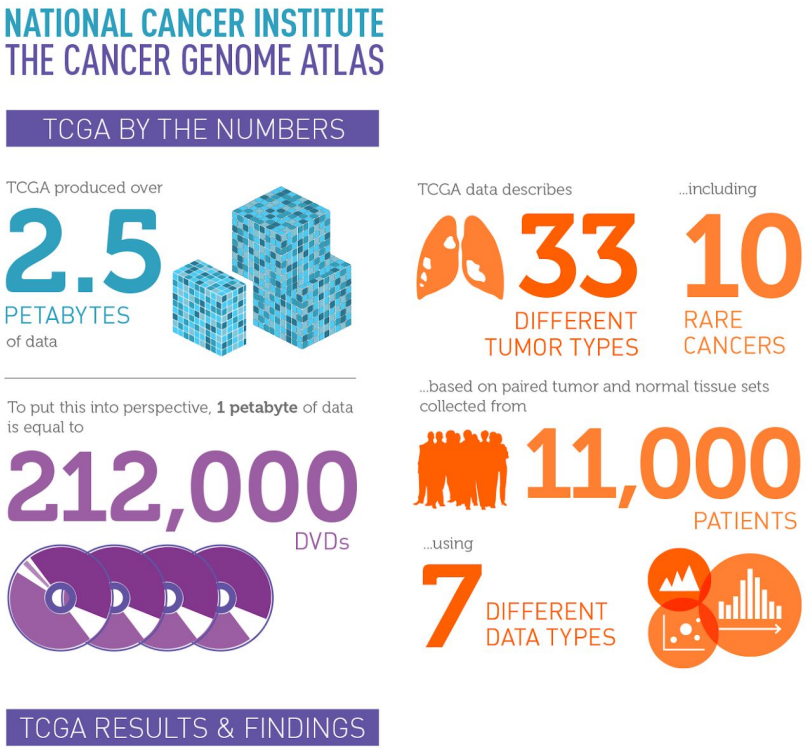
- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

#NCICloud

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

Common problems of big data:

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



Our mission at ISB-CGC

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

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

https://isb-cgc.org

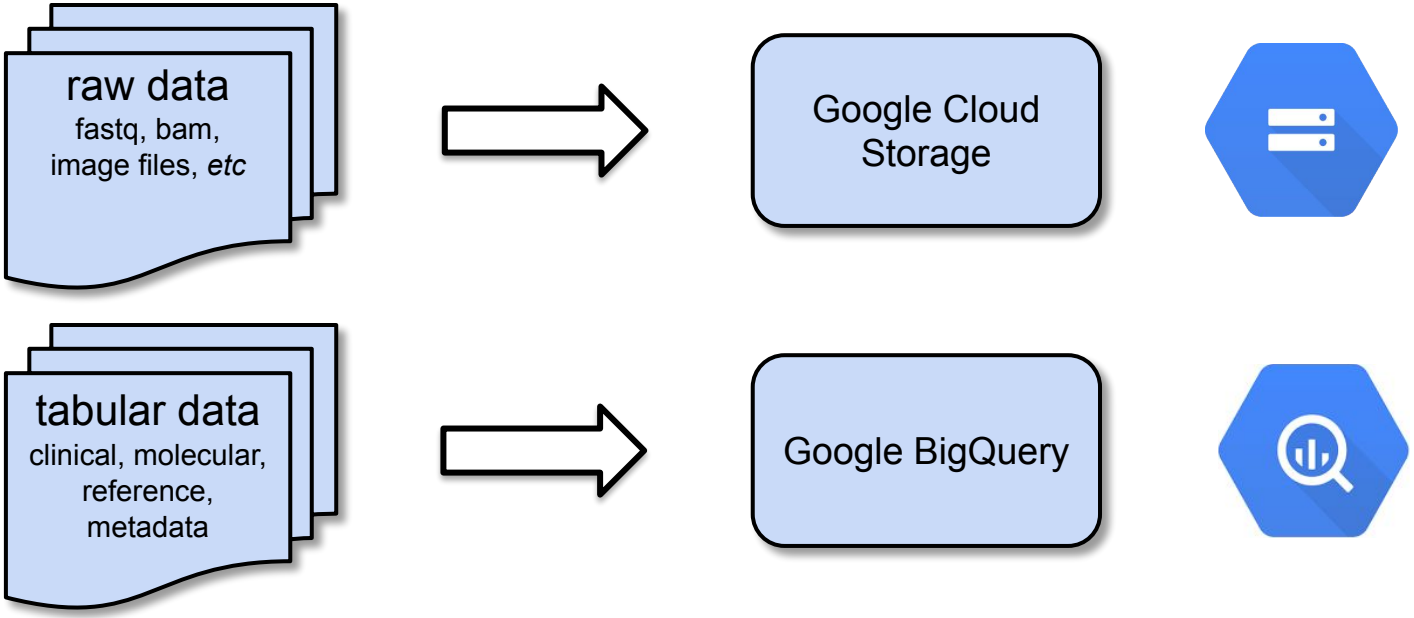
ISB

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

How do users access data on ISB-CGC?



ISB

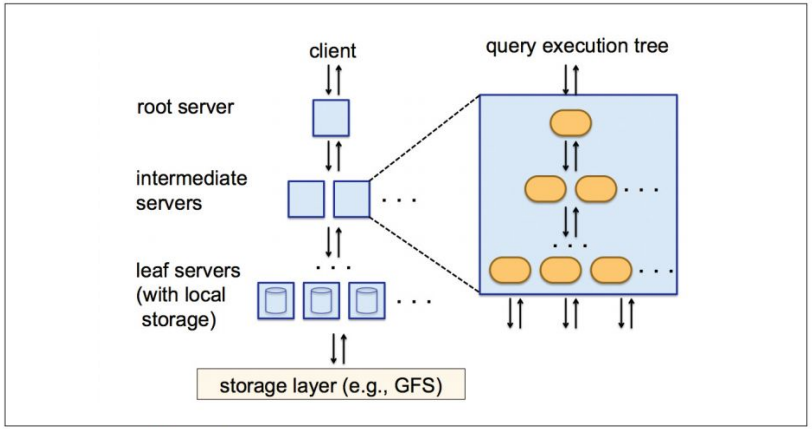
What is Google BigQuery and how does it enable –omics analyses?

- 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

ISB

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

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

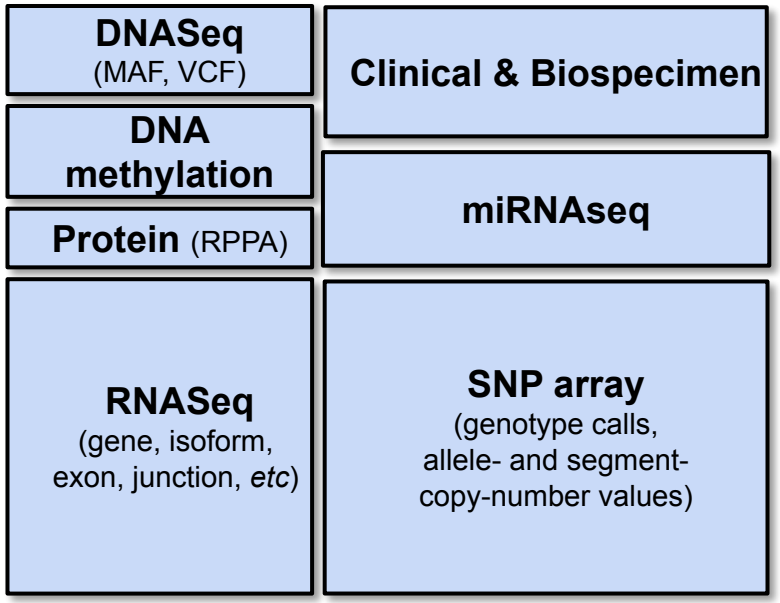


Tree architecture of Dremel

ISB

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



ISB

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.

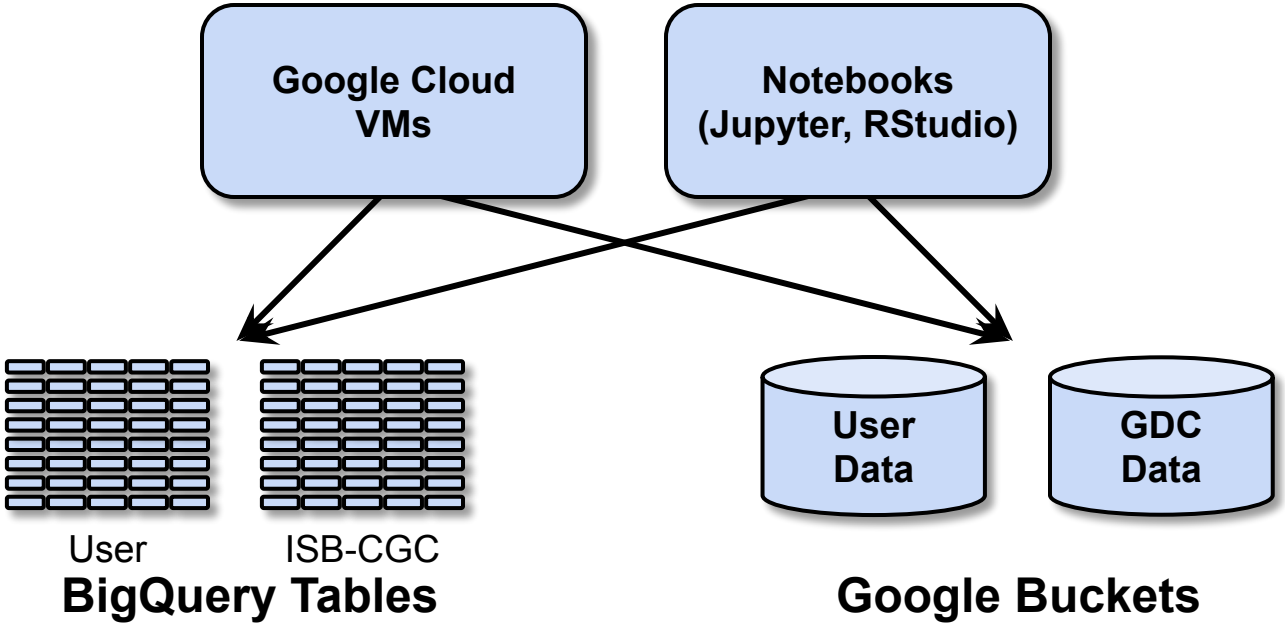
- Genomic reference tables
  - **PanCancer Atlas, COSMIC, ClinVar, cytoBand, dbSNP, Kaviar, Ensembl, Reactome, Gene Ontology, etc.**

- Metadata tables
  - Indexes of files, Google Cloud file paths, case ID, etc.

ISB

Multiple easy avenues for computing on data on ISB-CGC

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



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

Interactive web-based exploration

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

Direct Command line Access to VMs

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

Interactive big data exploration and analysis

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

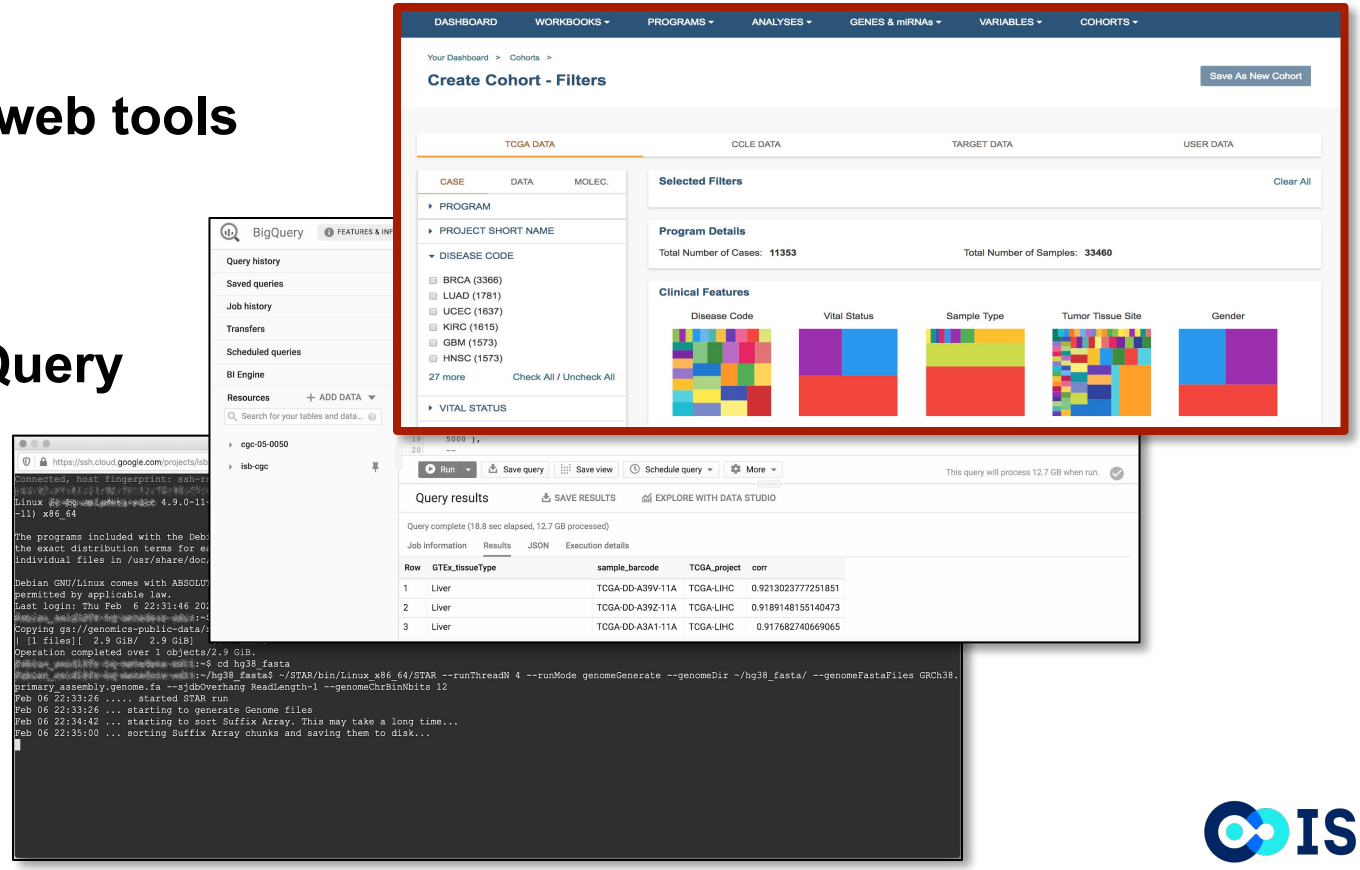
ISB

We provide data exploration tools through our web app

ISB-CGC web tools

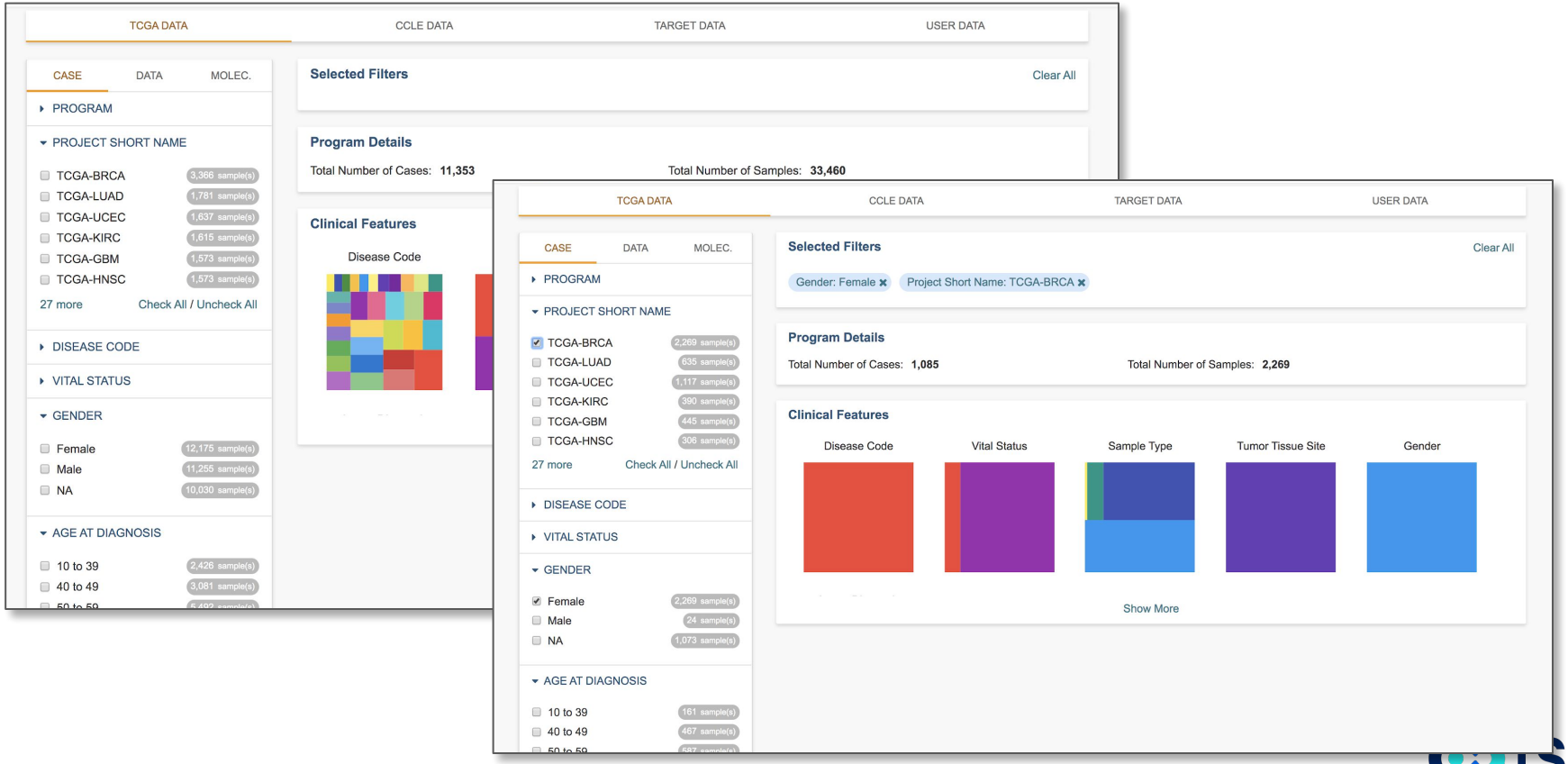
Google BigQuery

Google VMs

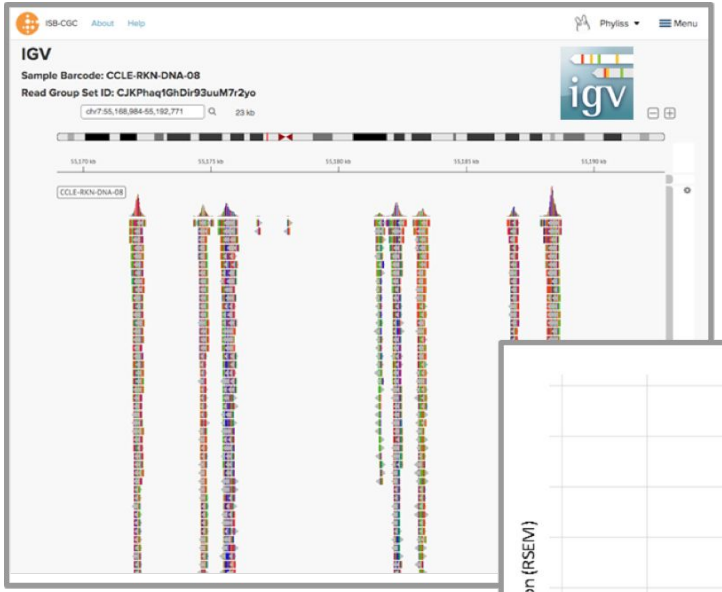


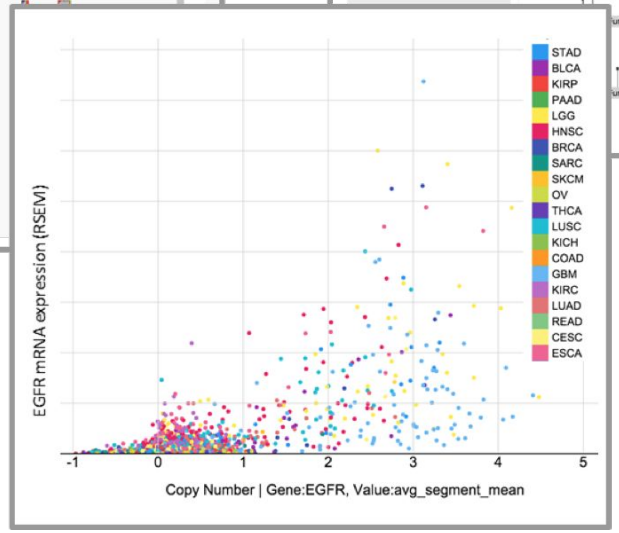
ISB

Building cohorts using the ISB-CGC web app



ISB-CGC: Interactive Apps





ISB

Integrated visualization methods for Big Data

Integrated genome viewer (view read pile-ups)

caMicroscope (view histology)

OHIF (view radiology)

Your Dash

File E

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

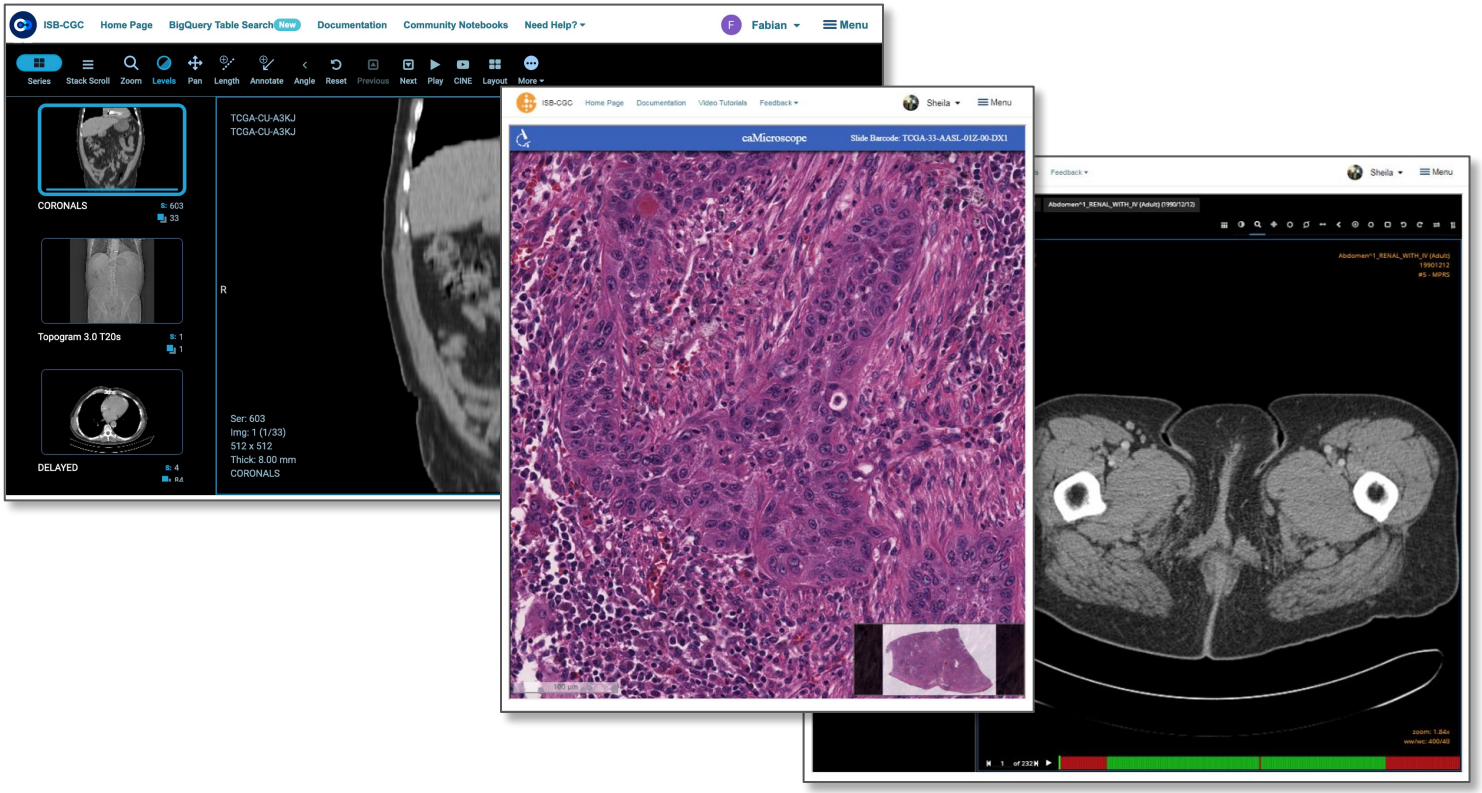
Show 25 **entries** **Page** Go Previous 1 2 3 ... 1588 Next

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

ISB-CGC: Interactive image viewers



ISB

The ISB-CGC BigQuery Table Search 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		

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

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

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

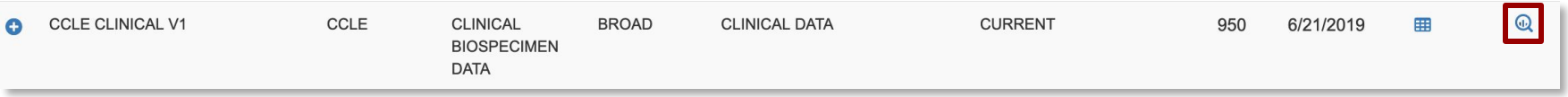
Have feedback or corrections? Please email us at feedback@isb-cgc.org.

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

ISB

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.



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.

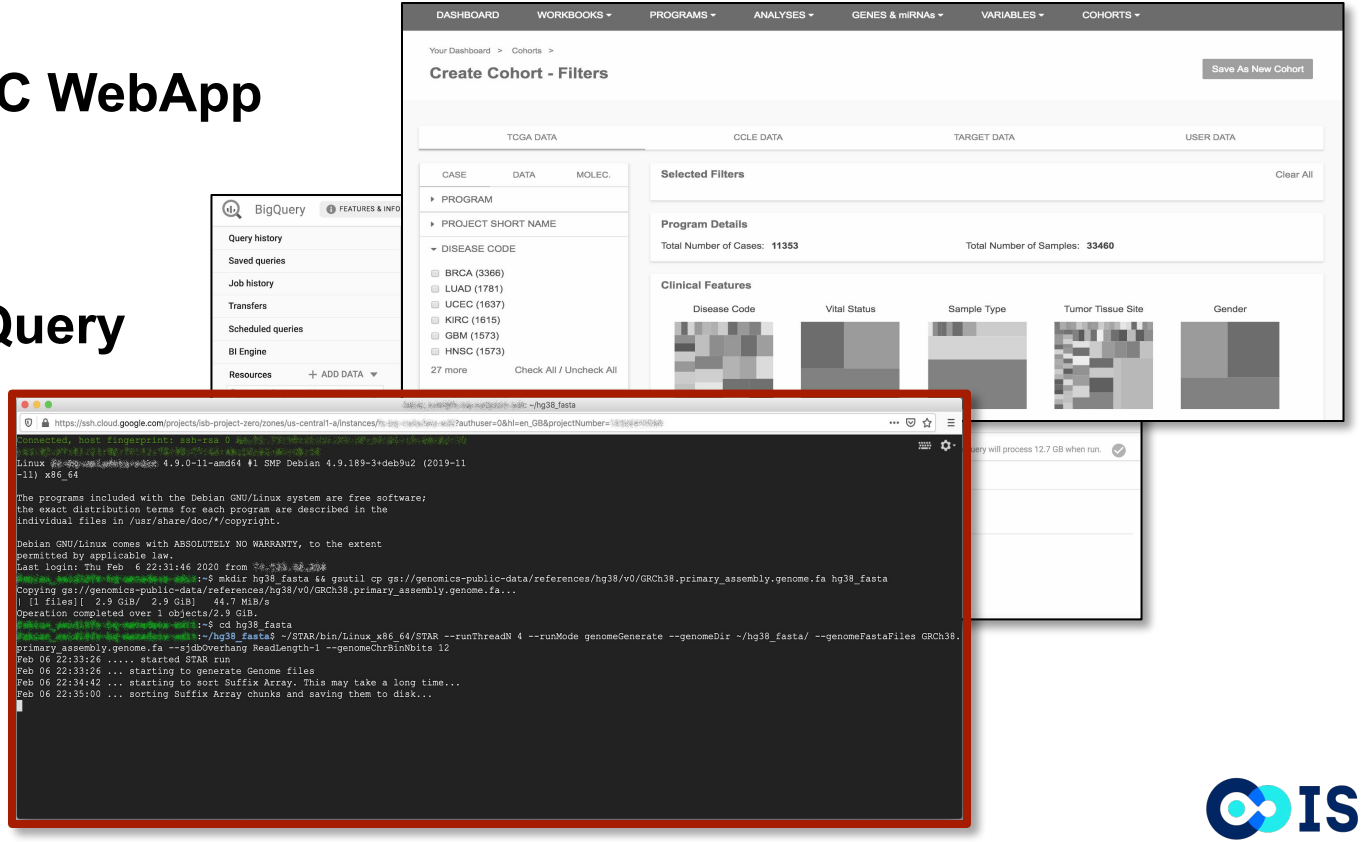
ISB

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC WebApp

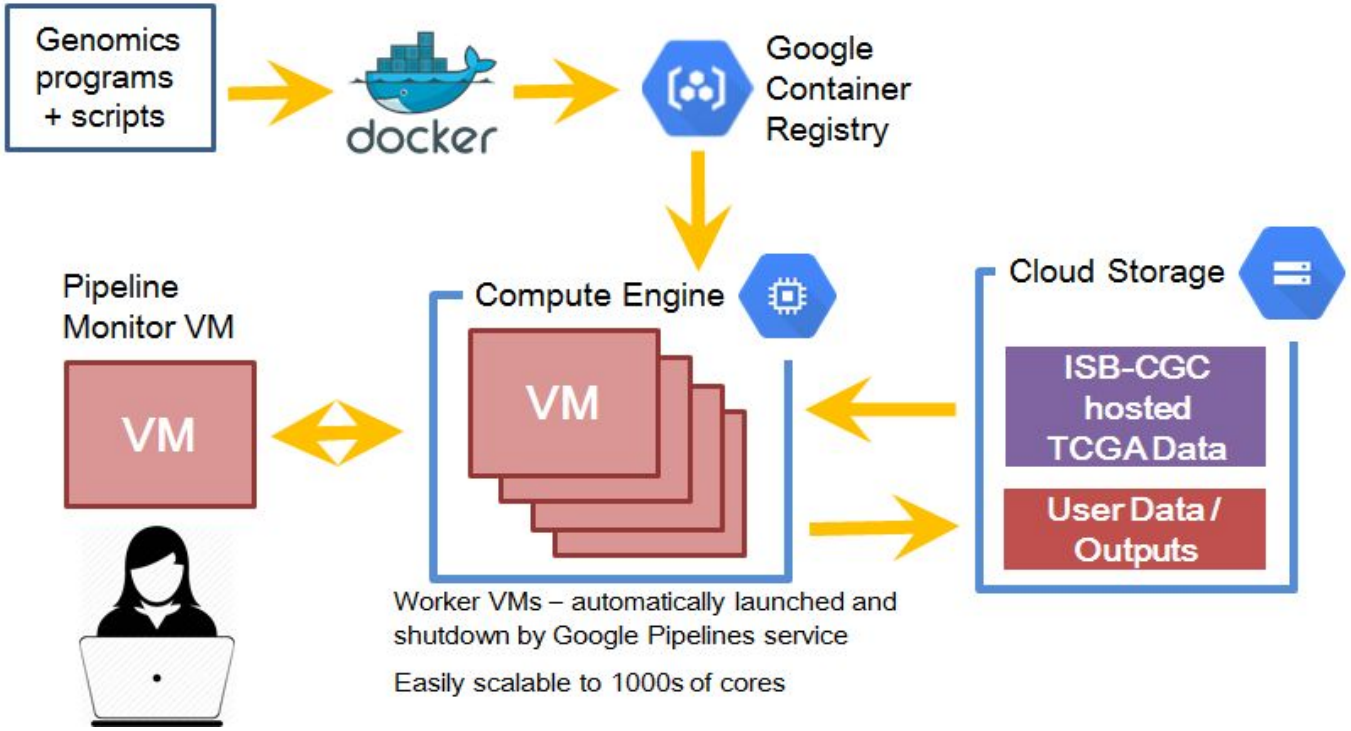
Google BigQuery

Google VMs



ISB

VMs enable advanced bioinformatic workflows



ISB

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

Multiple PanCancer Atlas projects, including:

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

Other end-user projects include:

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

SCIENTIFIC REPORTS /ABOUT /EDITORIAL BOARD /FOR AUTHORS

Sci Rep. 2016; 6: 39259. PMCID: PMC5159871
Published online 2016 Dec 16. doi: 10.1038/srep39259

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

PJ Tatlow[1] and Stephen R. Piccolo[a,1,2]

CSH Cold Spring Harbor Laboratory

bioRχiv beta THE PREPRINT SERVER FOR BIOLOGY

Pan-cancer analysis reveals complex tumor-specific alternative polyadenylation

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

101/160960

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

Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY
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

with many other manuscripts and grants currently in progress or submitted

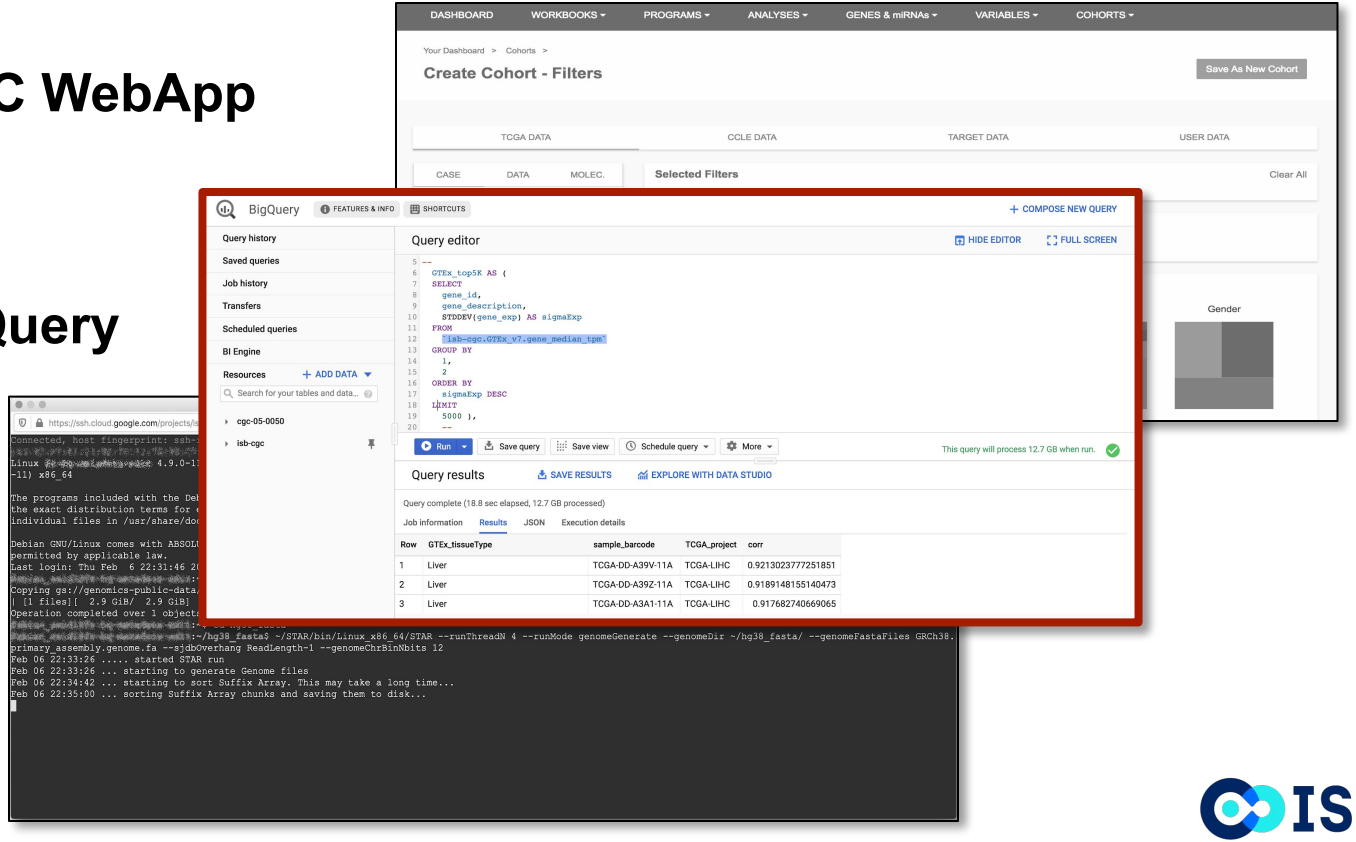
ISB

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC WebApp

Google BigQuery

Google VMs



ISB

BigQuery integrates with a variety of commonly used analysis tools

bigrquery and
bigQueryR

googleAuthR

Pre-built VM images

IP[y]:
IPython

Cloud Datalab

Cloud notebooks
and workspaces.

ISB

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

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.pati
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
```

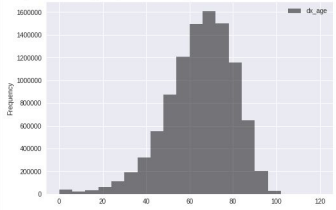
```
import pandas as pd
project_id = 'isb-project-zero'
```

```
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>
```

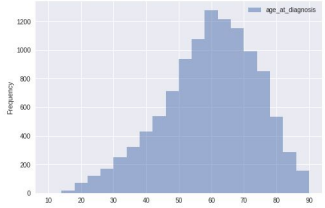



```
df2 = pd.io.gbq.read_gbq('''
select age_at_diagnosis
from
  `isb-cgc.TCGA_bioclin_v0.Clinical`
''',project_id='isb-project-zero', verbose=False, dialect='standard')

df2.head()
```

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

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



ISB

Use Google BigQuery to easily connect your research to public datasets

ISB-CGC and Other
Public Datasets

↓

SQL

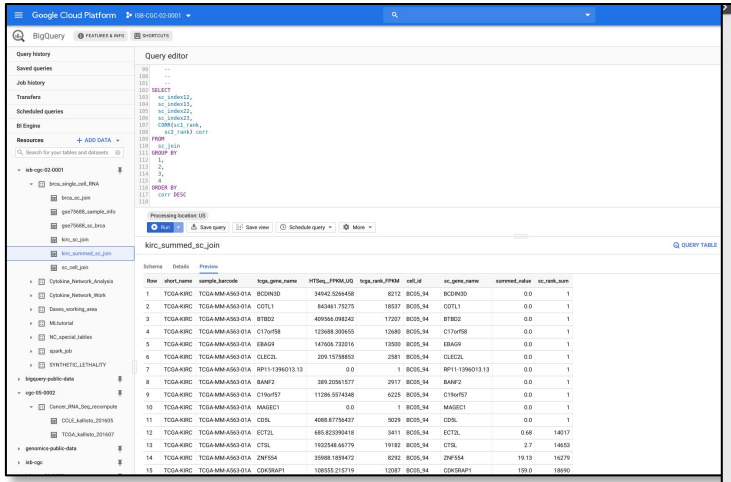
↑

Private User Data
and Derived Results

ISB

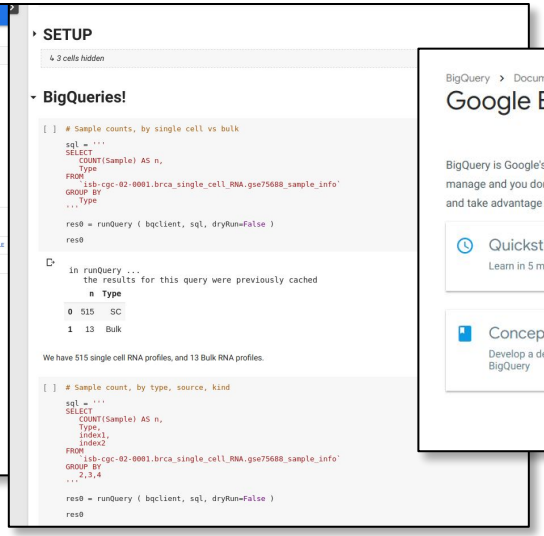
A typical work setup across multiple browser tabs

Google web interface



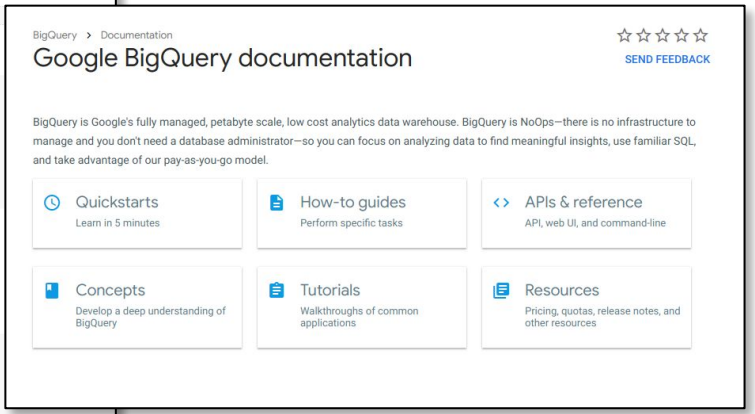
Built in syntax checking

Notebook (R or Python)



Integrate with notebooks to generate *your own* publication quality visuals

Searchable web docs



ISB

What you need to know to interoperate with ISB-CGC

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

ISB

Managing security and permissions via service accounts

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

ISB

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

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
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 machines.

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 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.

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

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

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

ISB

Questions?

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
Ron Taylor
Fabian Seidl
Deena Bleich
Mark Backus
Derrick Moore
Owais Shahzada