Cancer Data Analytics on the ISB-CGC platform

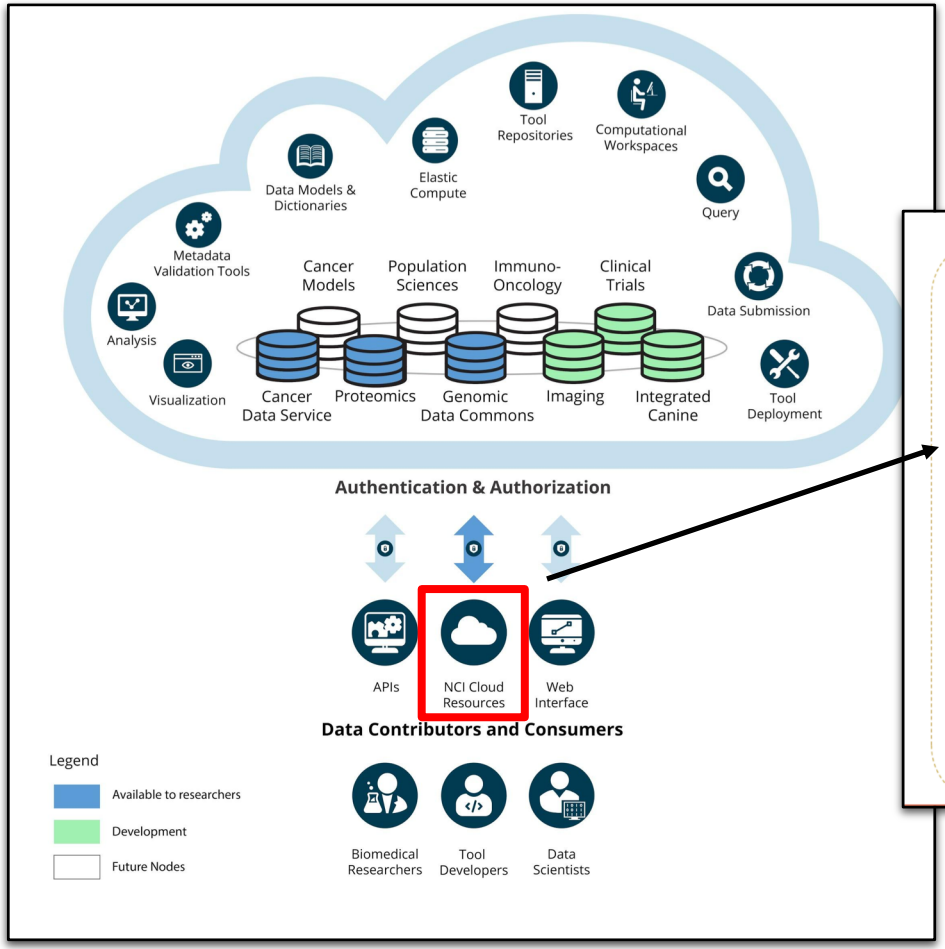
2020-04-10

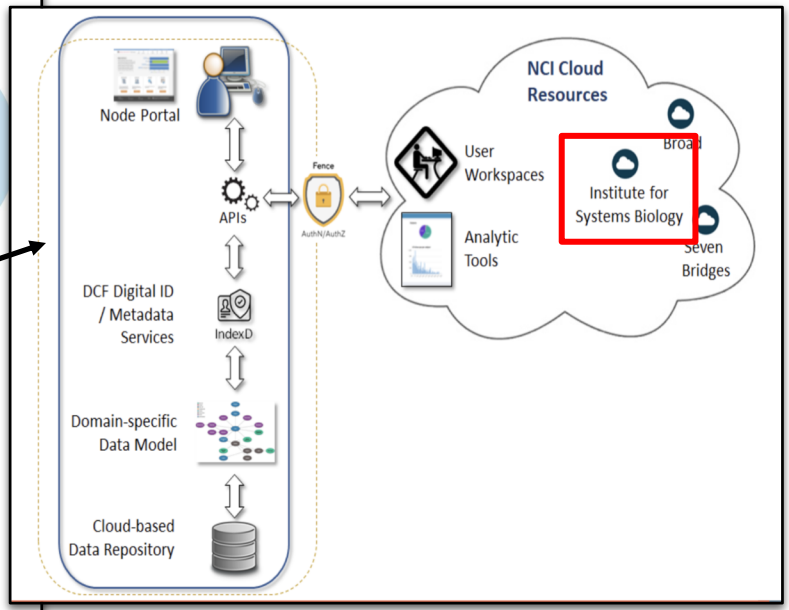
Kawther Abdilleh, PhD
Fabian Seidl, PhD

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





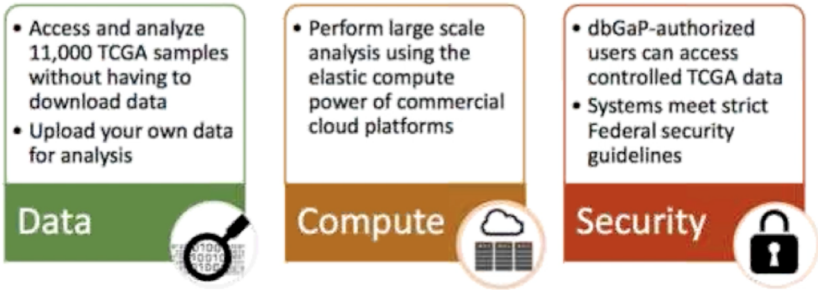
https://datascience.cancer.gov/data-commons

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



#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

ISB

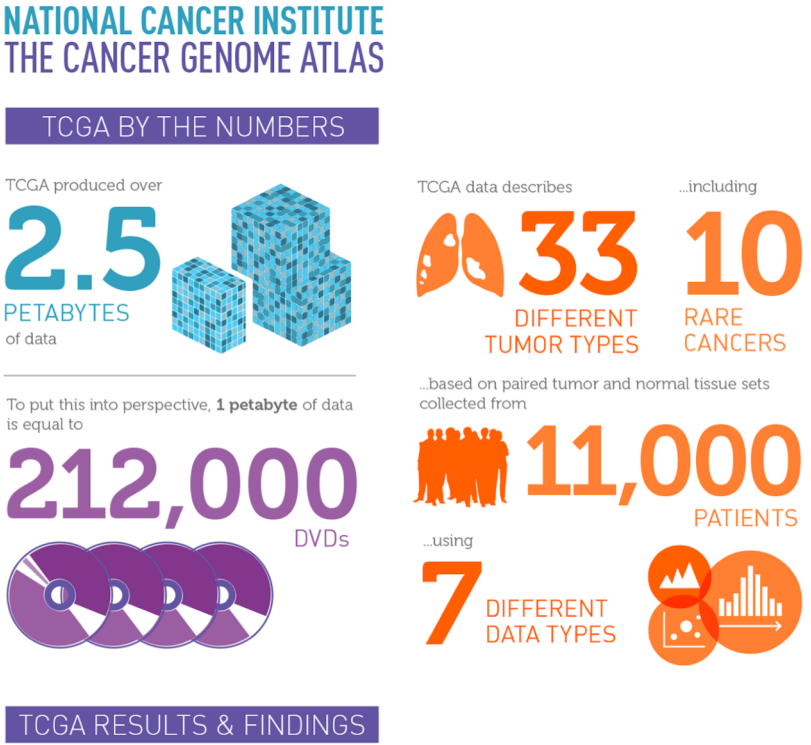
Our Approach at ISB-CGC

- 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

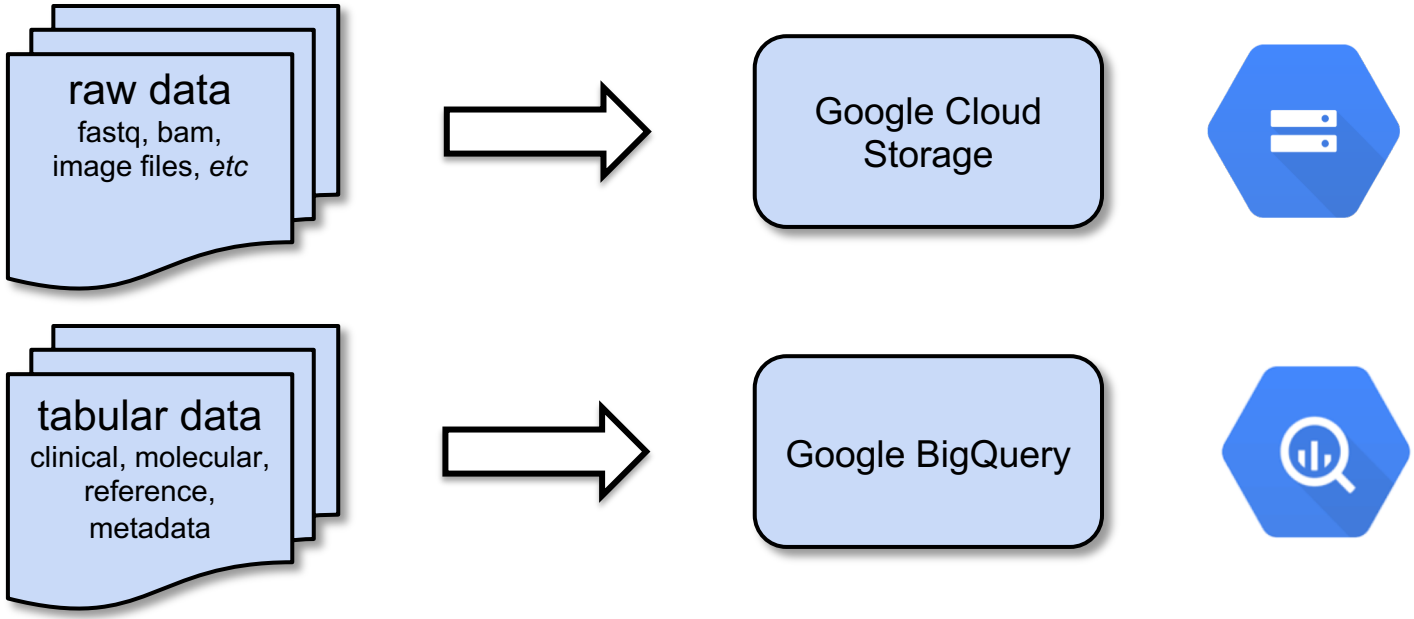
ISB-CGC provides Data as a Service (DaaS) solutions to the rapid growth of cancer data

Common problems of big data:

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



How do users access data on ISB-CGC?



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

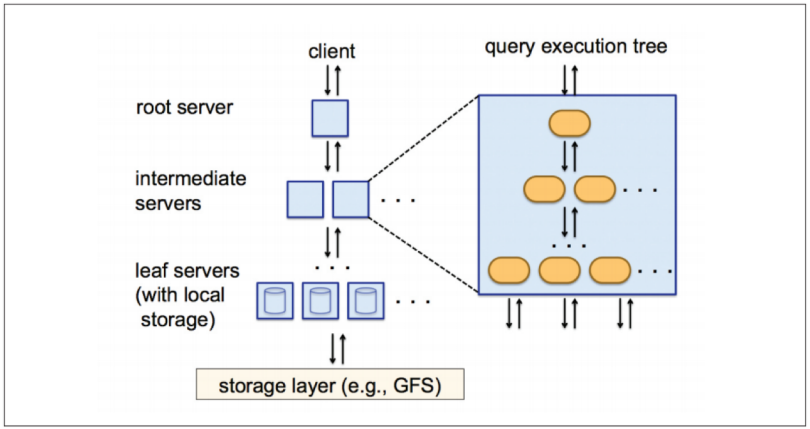
- Cloud-based web service from Google Cloud used for handling and analyzing big data
- In the world of “omics”, it can facilitate high-throughput data analysis on the Cloud inexpensively in the following ways:
- **Storage**:
  - Store the results from large-scale pipelines/workflows in centralized BigQuery 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

ISB

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

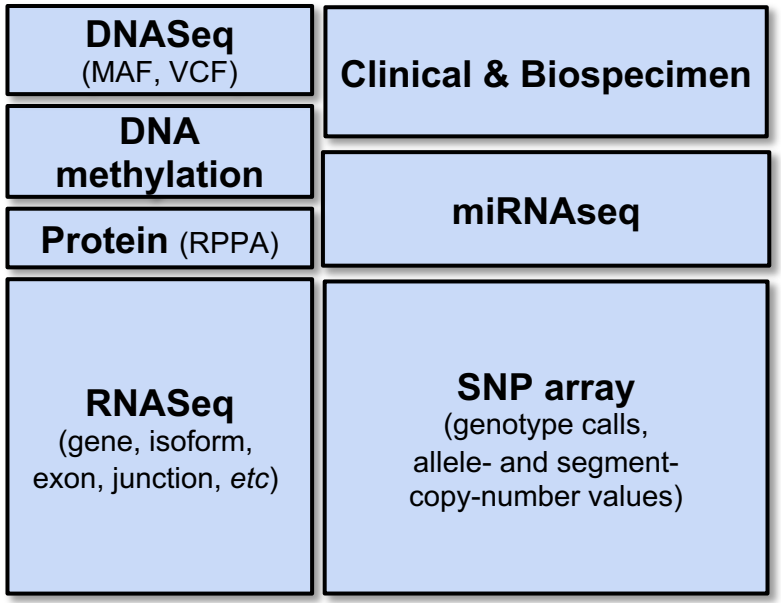


Tree architecture of Dremel

ISB

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.

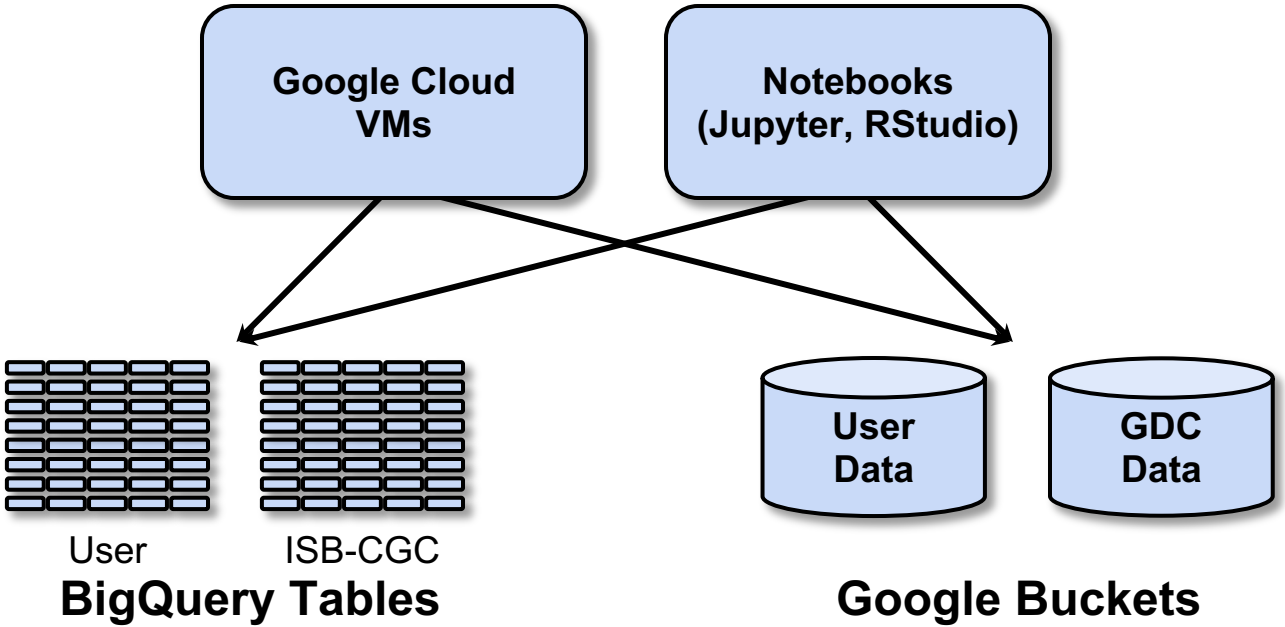
- Genomic reference tables
  - **PanCancer Atlas, COSMIC, ClinVar, cytoBand, dbSNP, Kaviar, Ensembl, Reactome, Gene Ontology, etc.**

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

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

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*…

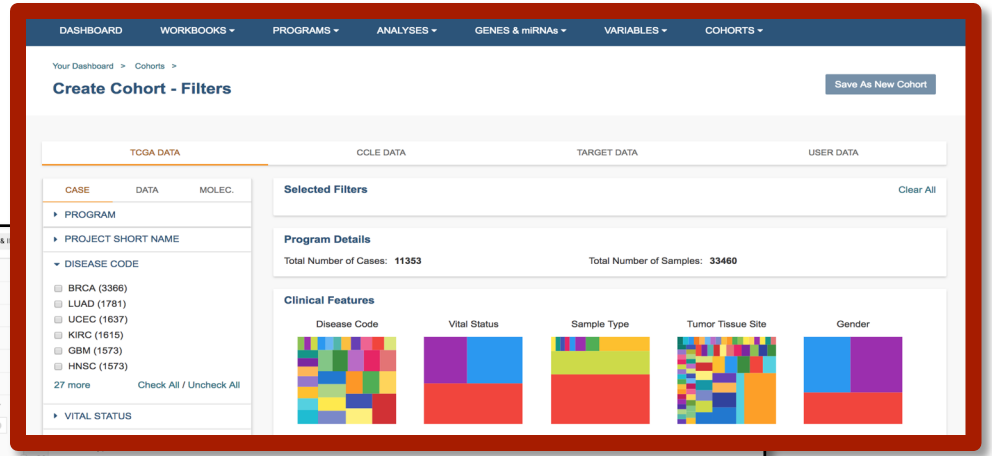
ISB

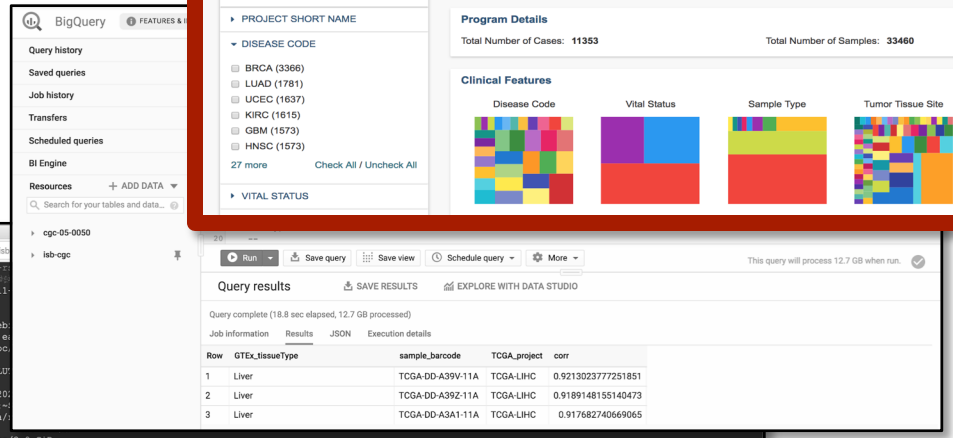
Three entry points for exploring cancer data on ISB-CGC

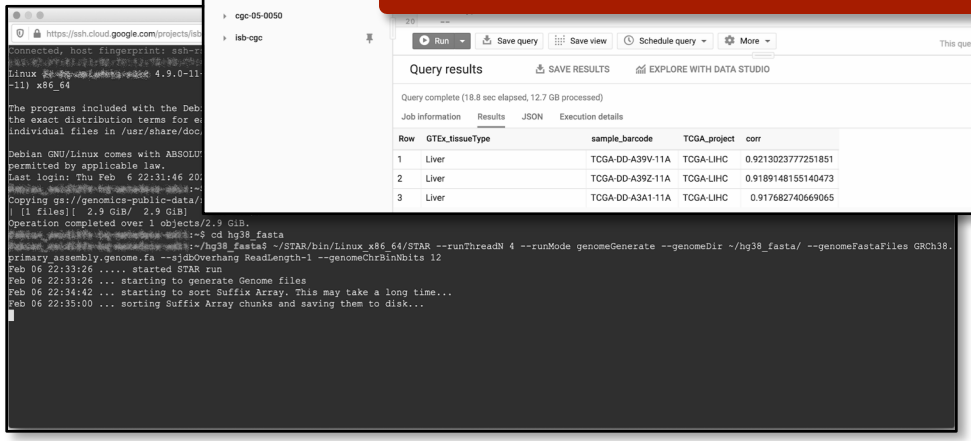
ISB-CGC web tools

Google BigQuery

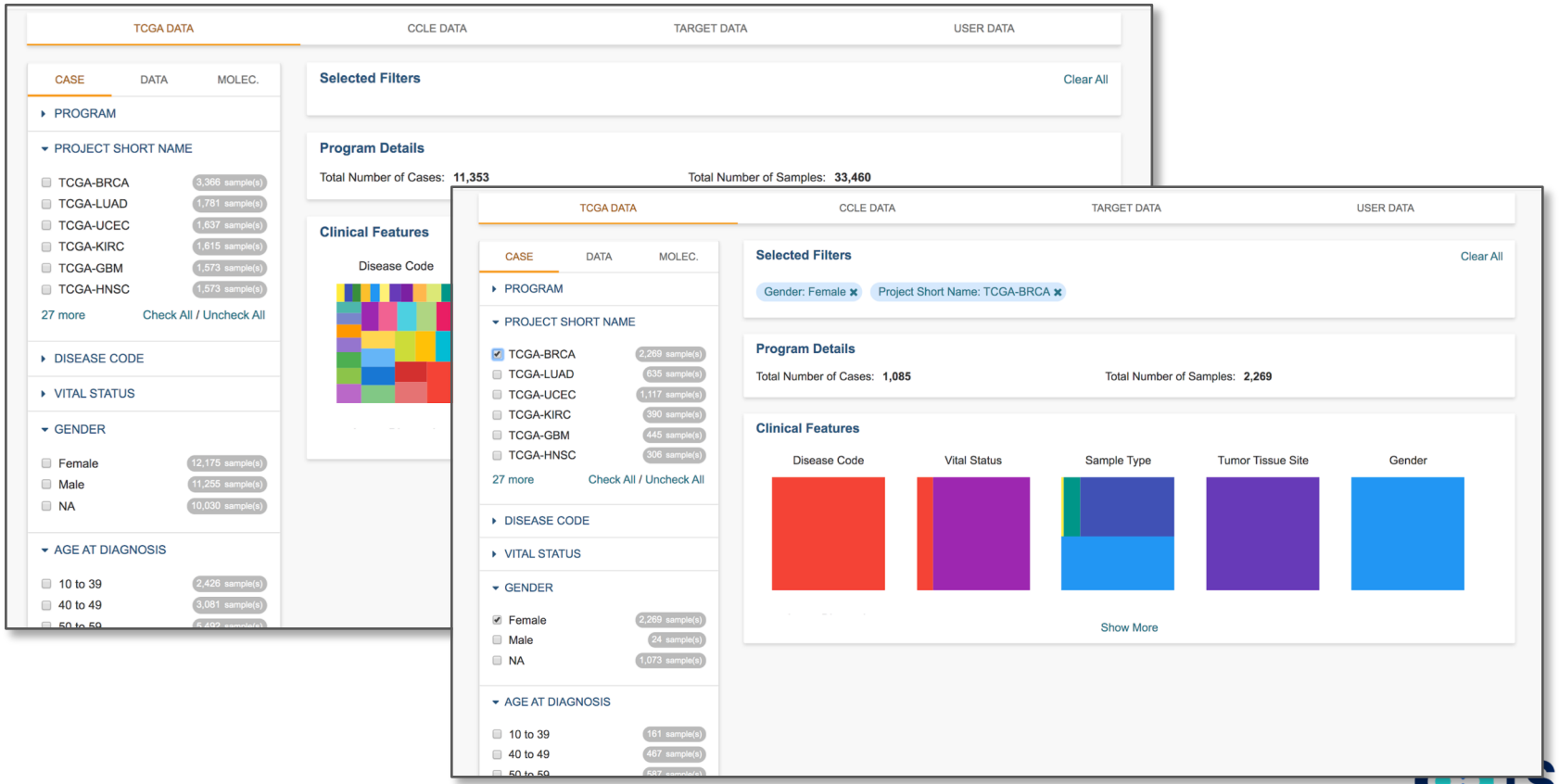
Google VMs



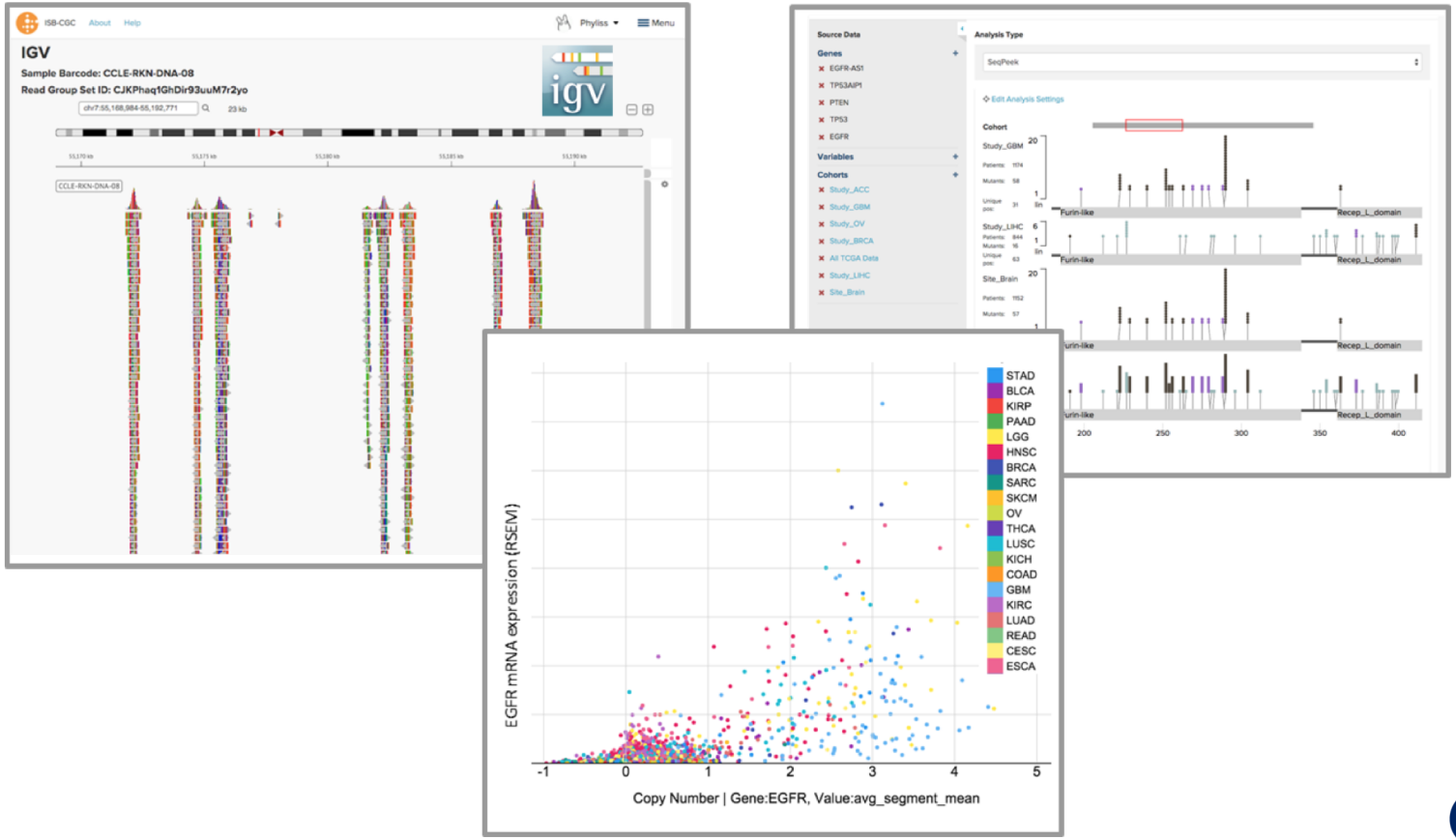




Interactive cohort-building using the ISB-CGC

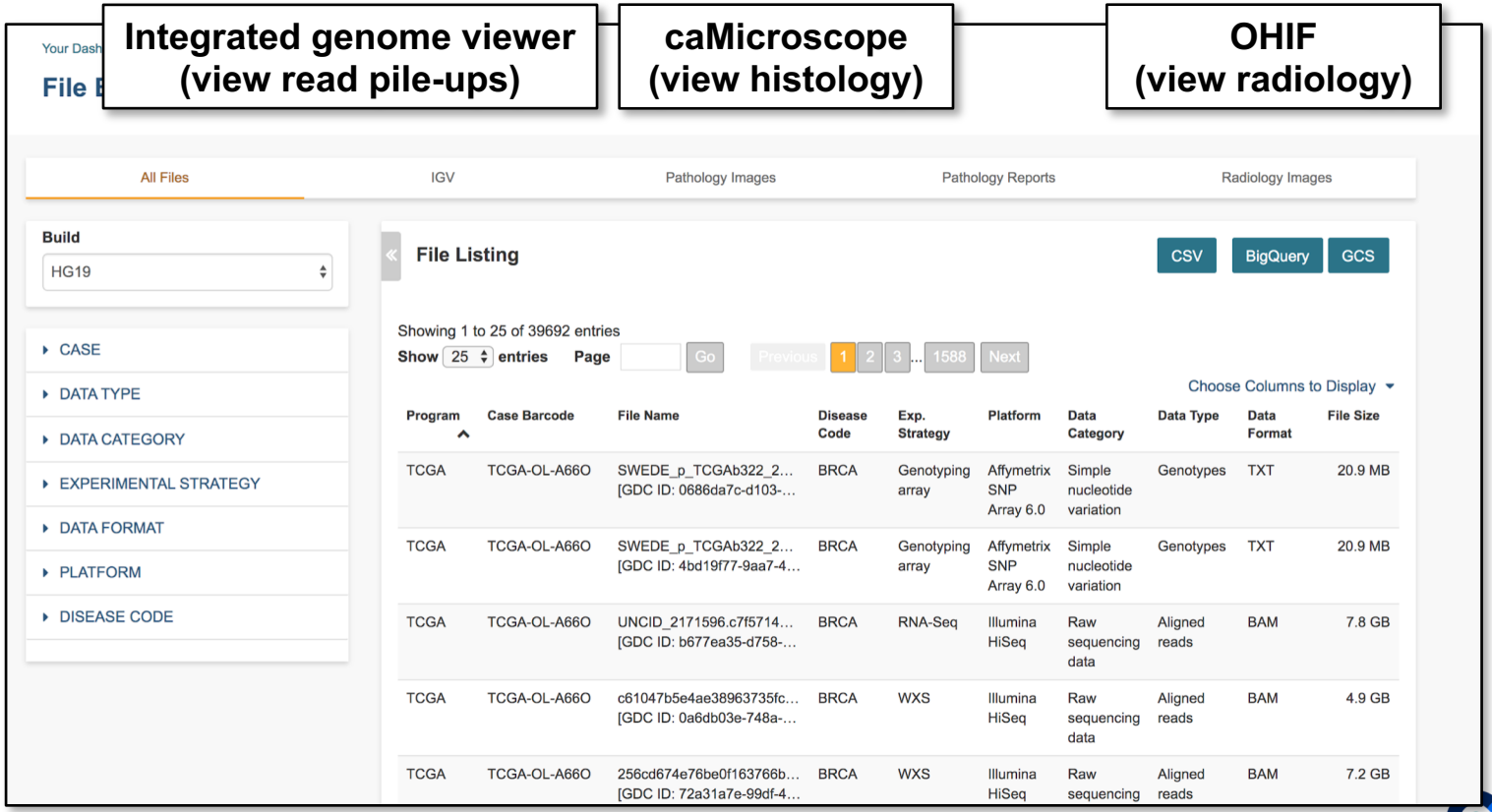


ISB-CGC: Interactive Apps



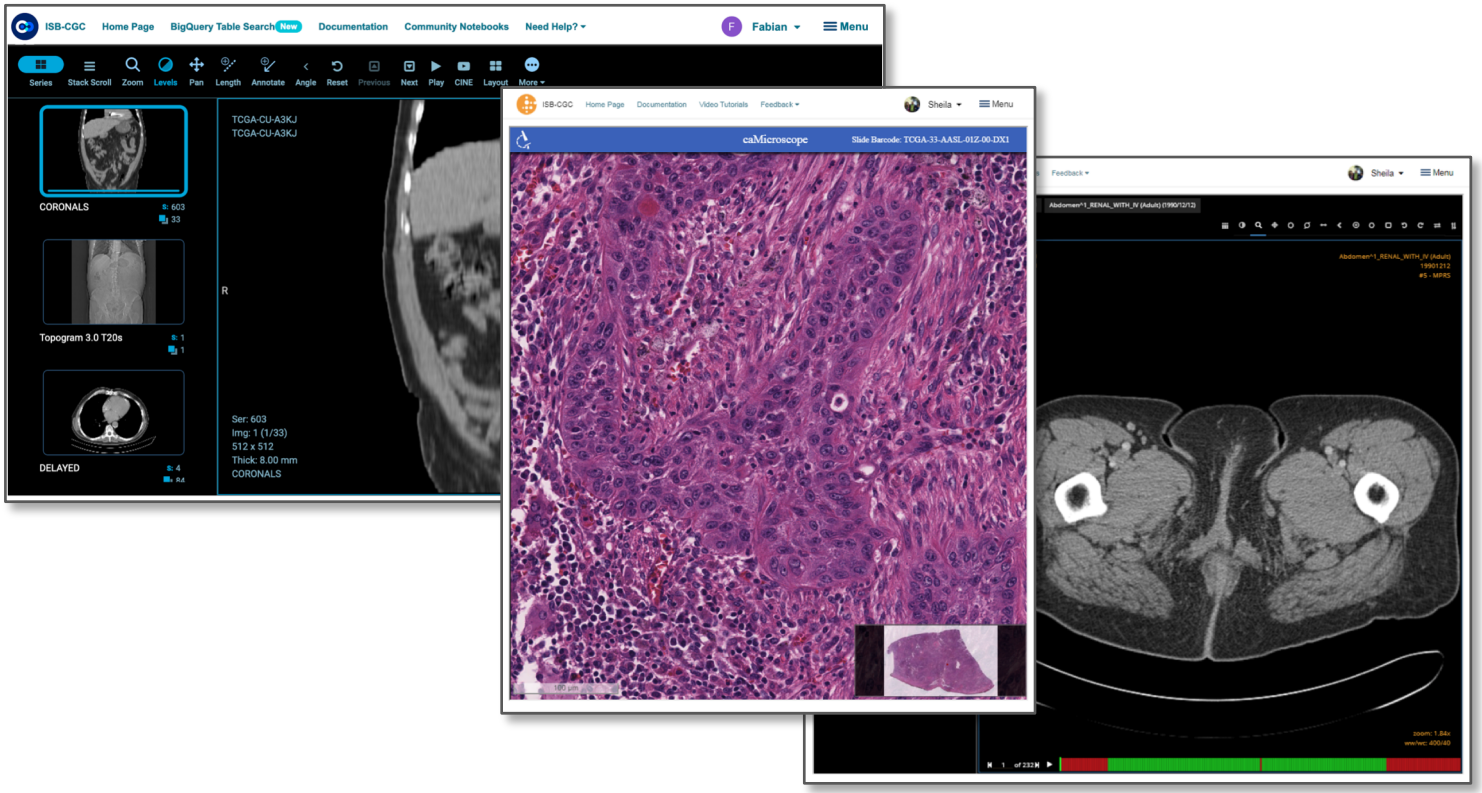
ISB

Integrated visualization methods for Big Data

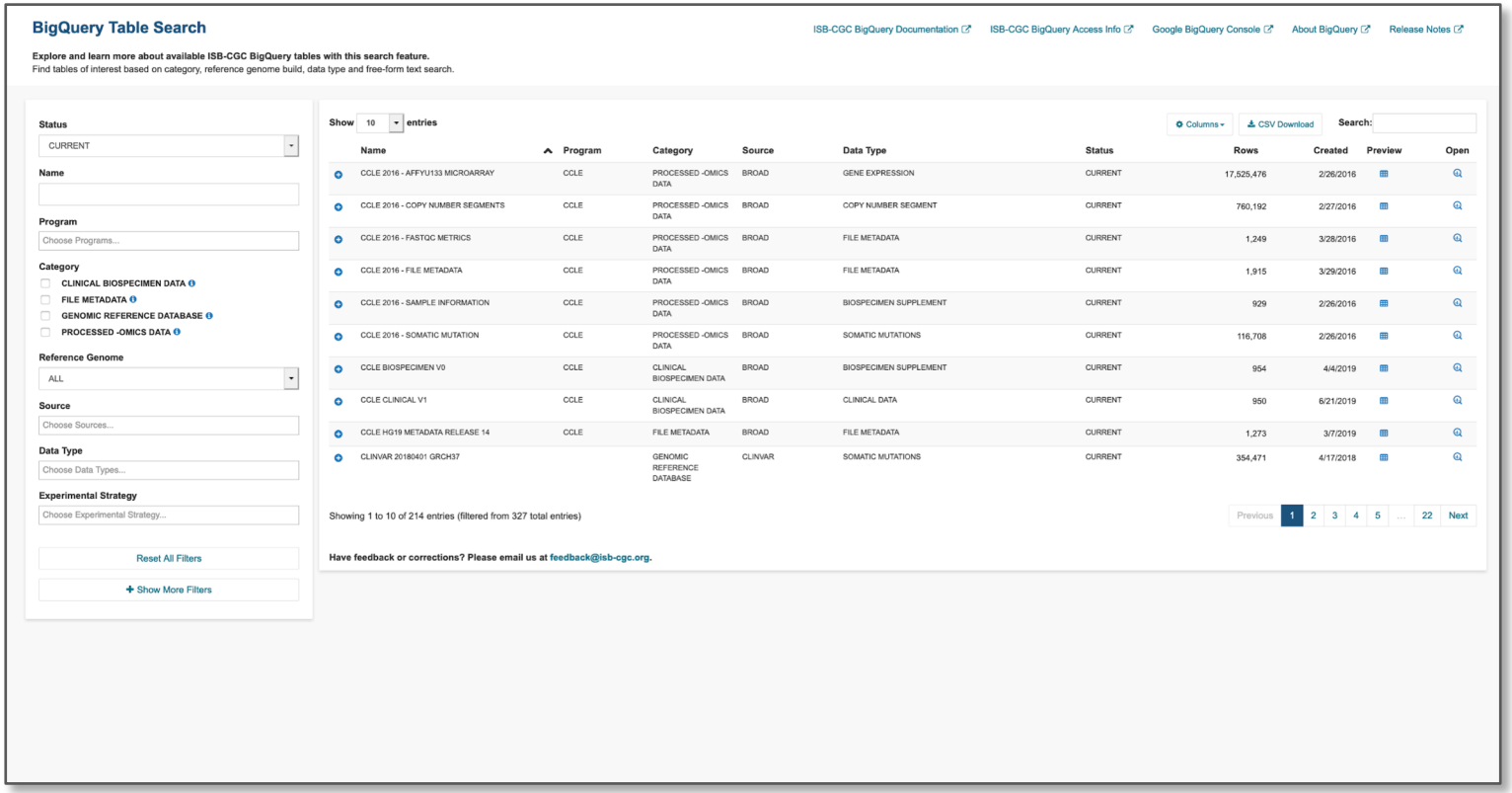


ISB

ISB-CGC: Interactive image viewers

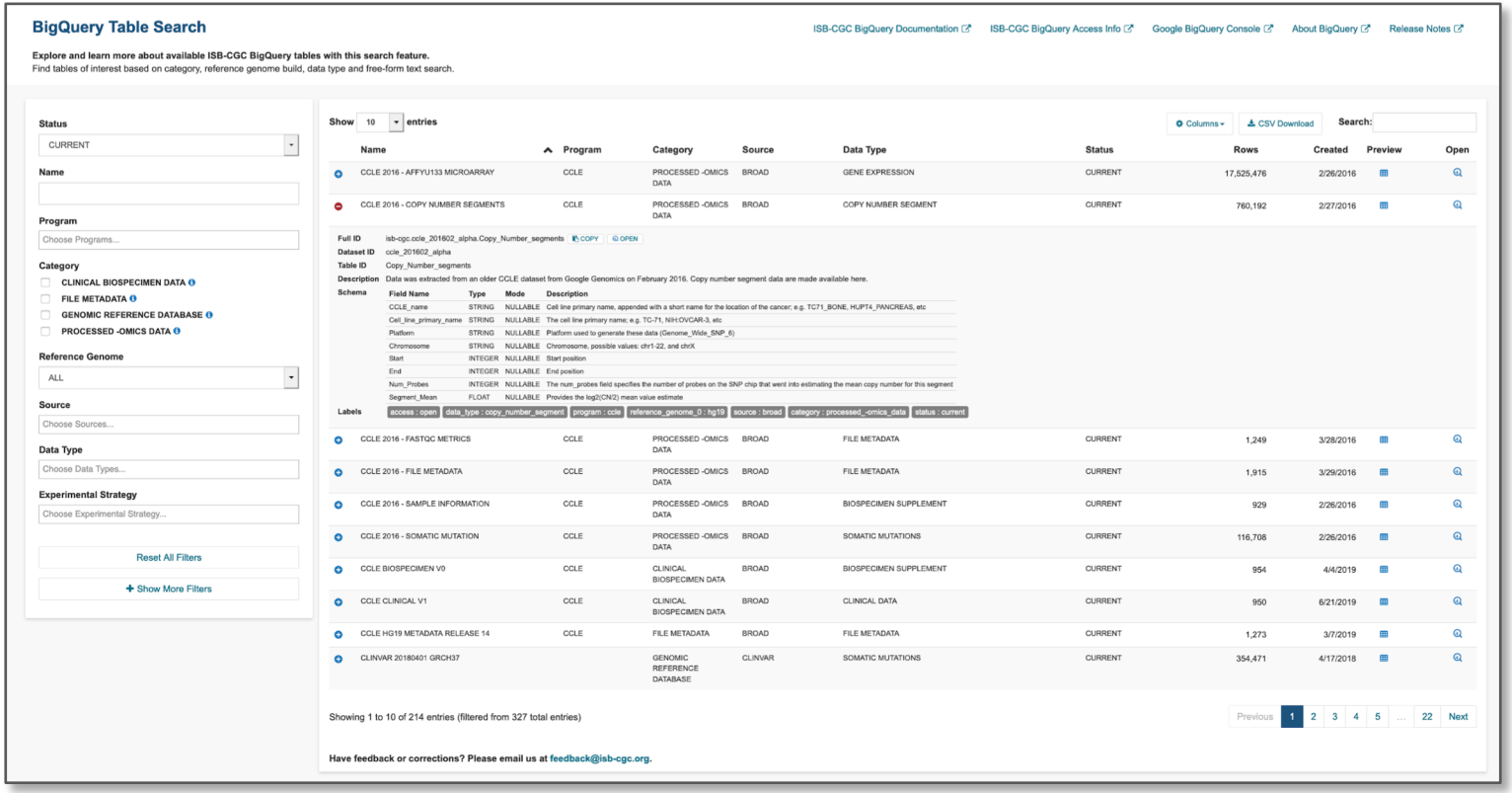


The ISB-CGC BigQuery Table Search UI



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

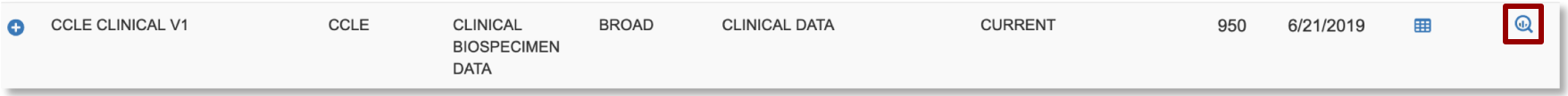
More information on a table at the click of a button!



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

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.

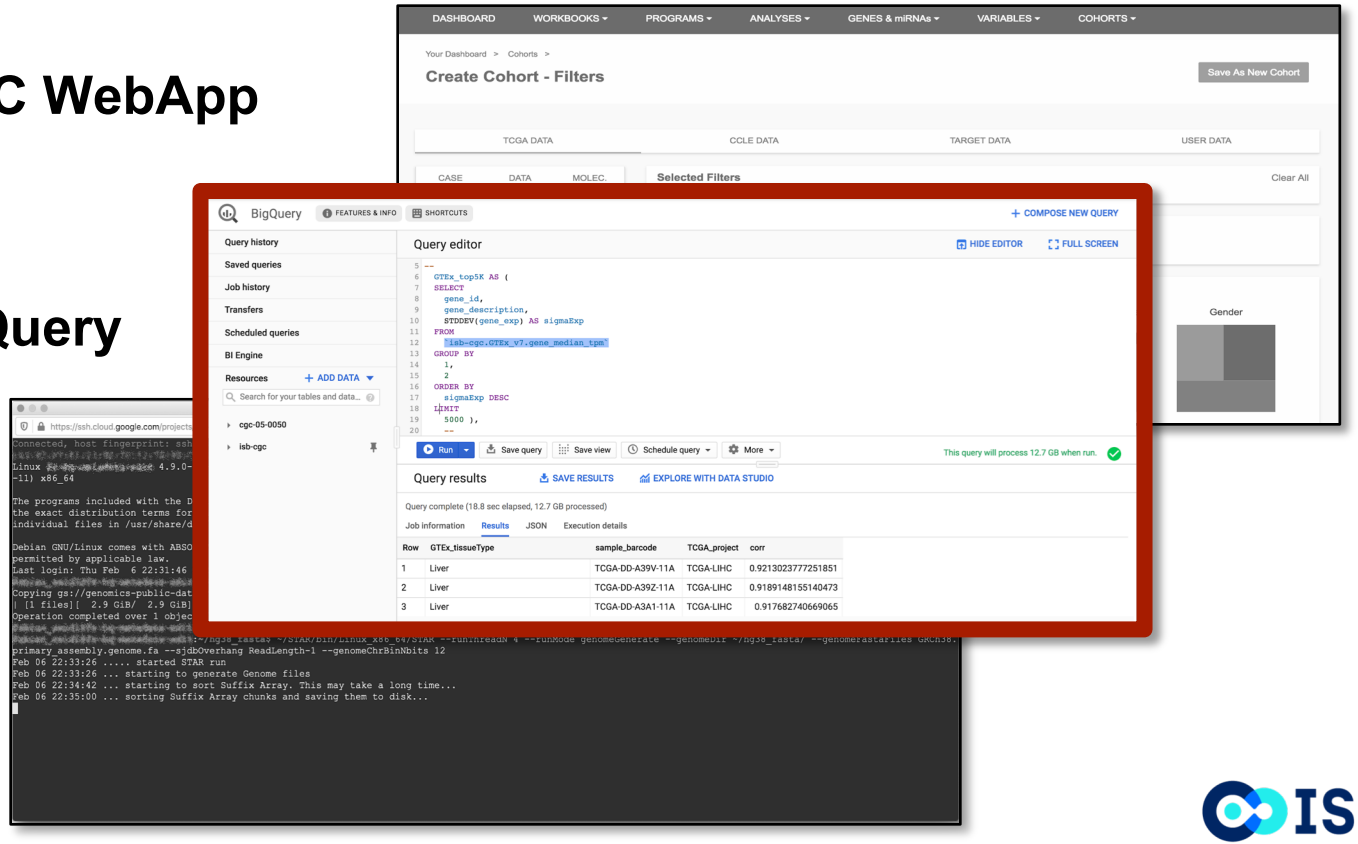
https://mitelmandatabase.isb-cgc.org/

Three entry points for exploring cancer data on ISB-CGC

ISB-CGC WebApp

Google BigQuery

Google VMs



BigQuery integrates with a variety of commonly used analysis tools

bigrquery and
bigQueryR

googleAuthR

Pre-built VM images

Cloud notebooks
and workspaces.

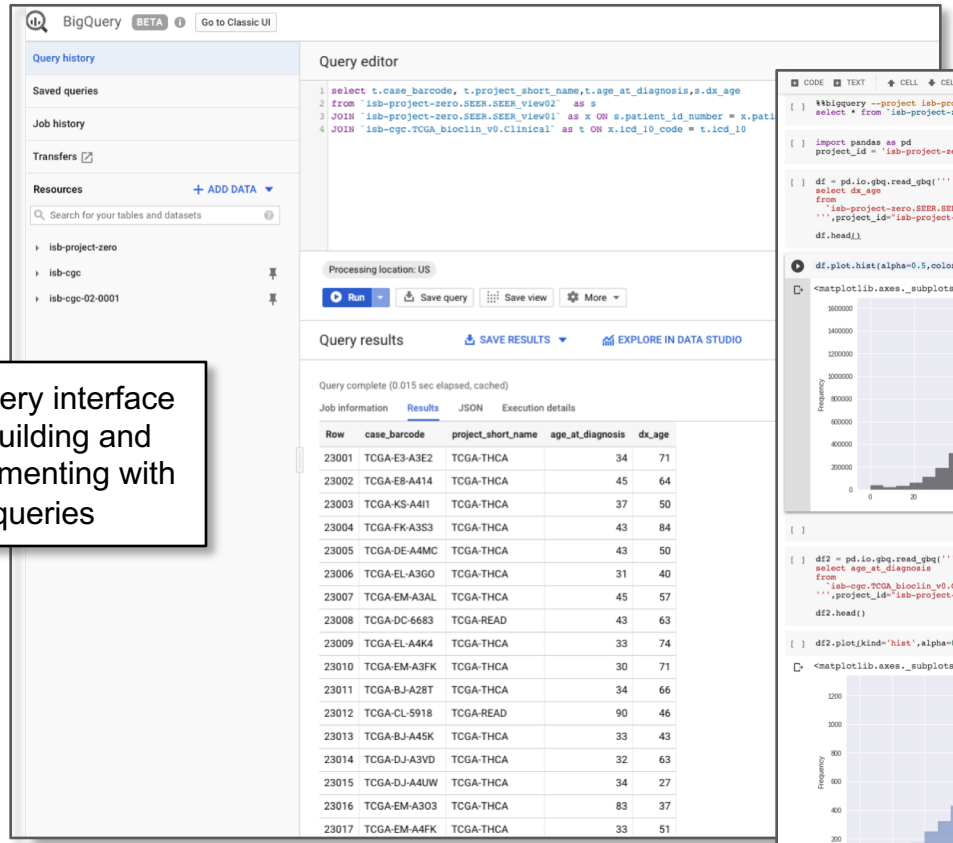
Analyze correlation between TCGA samples & GTEx tissue types quickly and cheaply!

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



BigQuery interface for building and experimenting with queries

Jupyter notebook to visualize and share analysis

ISB

Use Google BigQuery to easily connect your research to public datasets

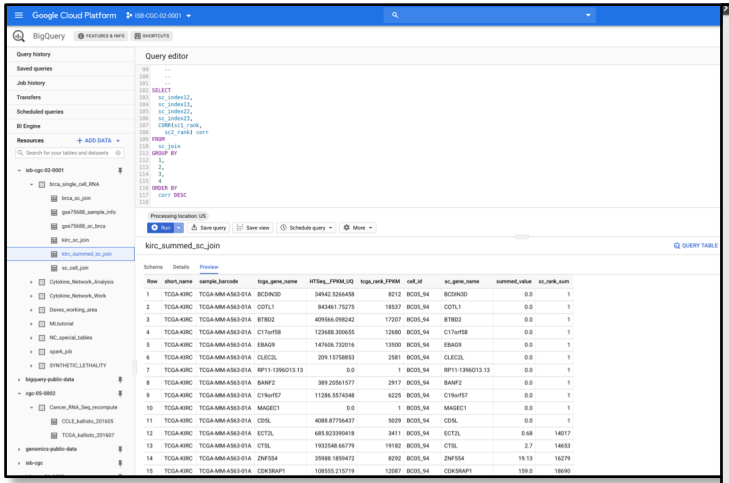
ISB-CGC and Other
Public Datasets

Private User Data
and Derived Results

ISB

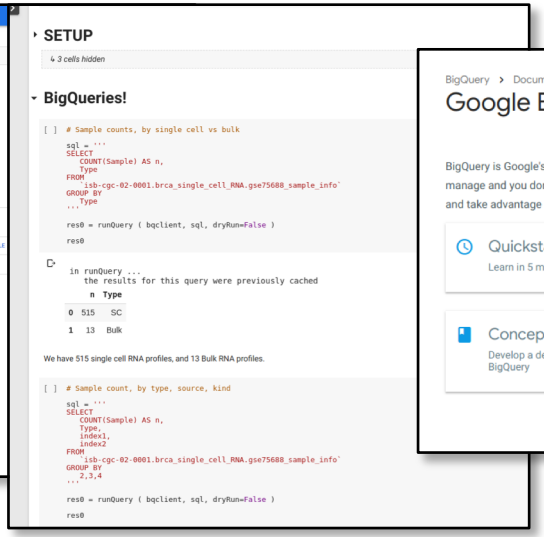
A typical work setup across multiple browser tabs

Google web interface



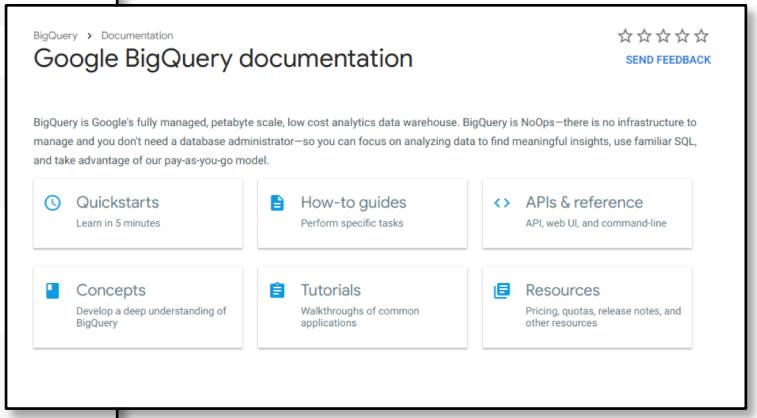
Built in syntax checking

Notebook (R or Python)



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

Searchable web docs



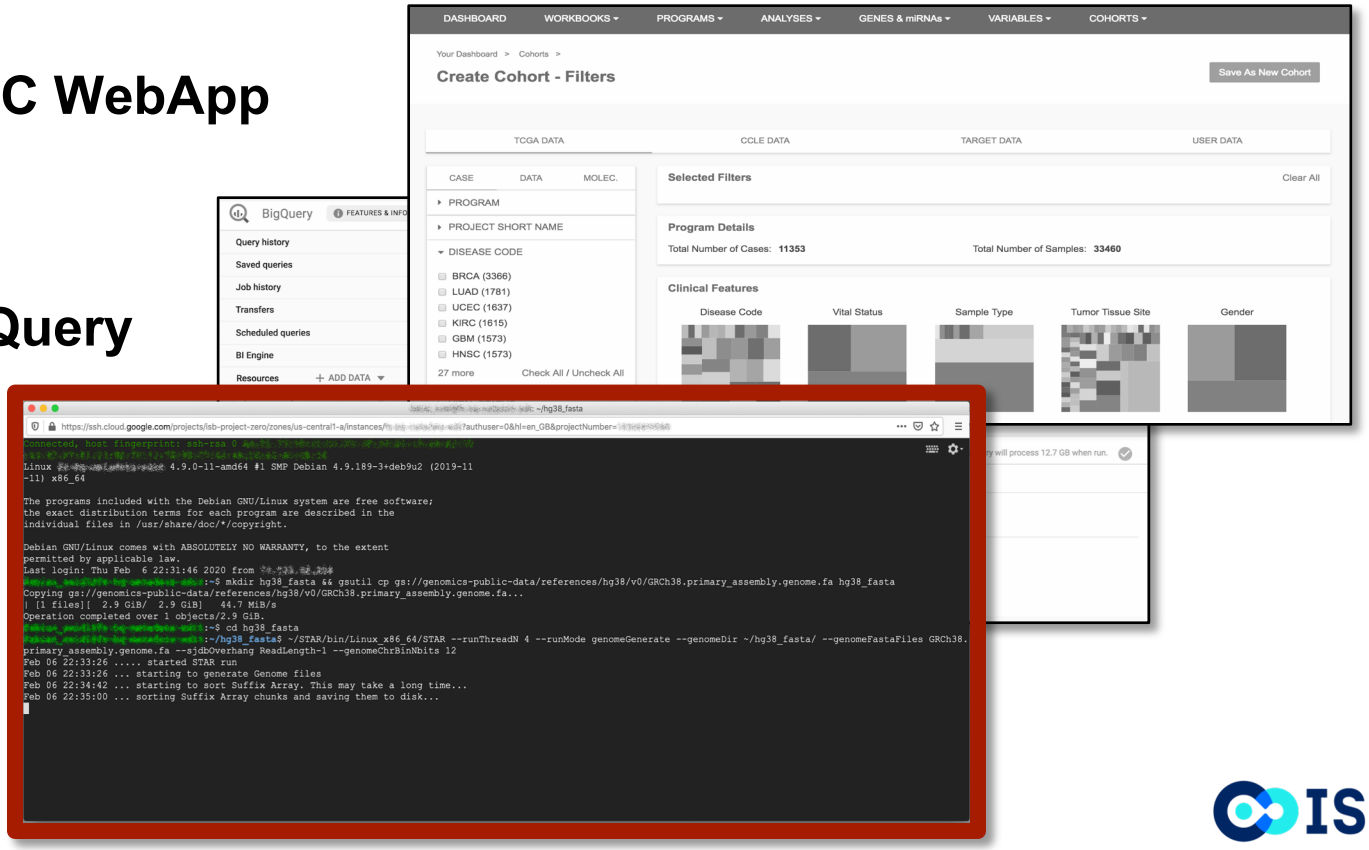
ISB

Three entry points for exploring cancer data on ISB-CGC

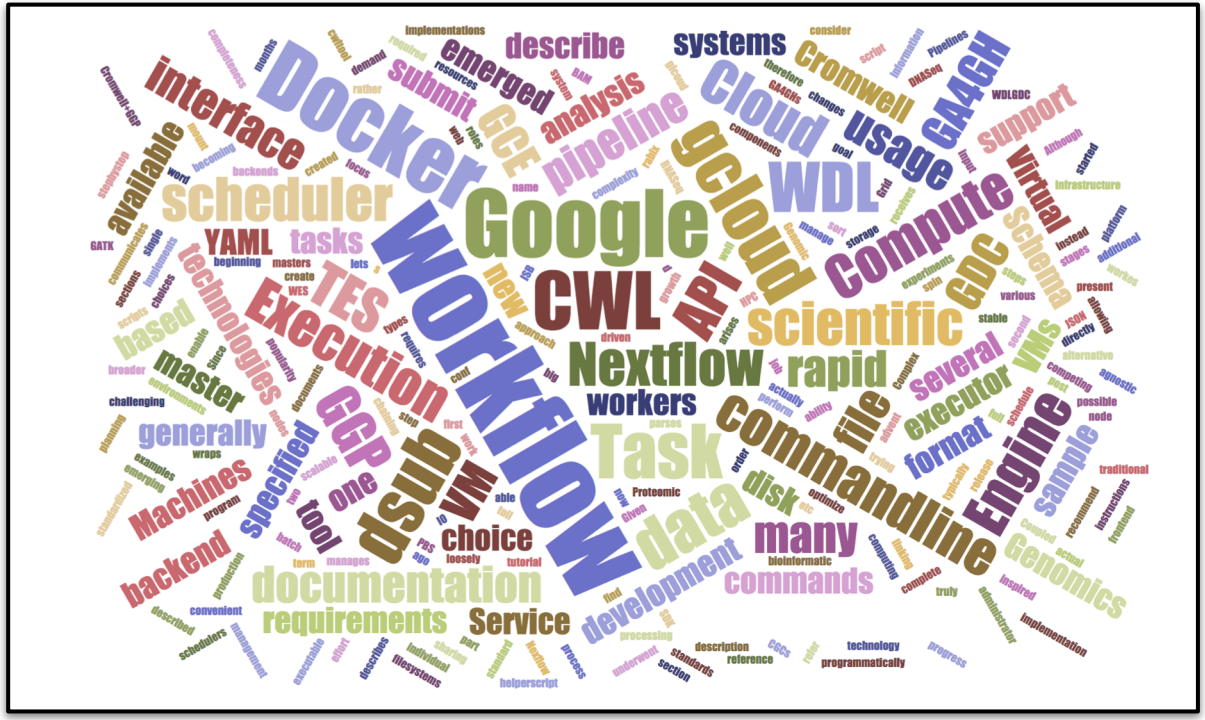
ISB-CGC WebApp

Google BigQuery

Google VMs



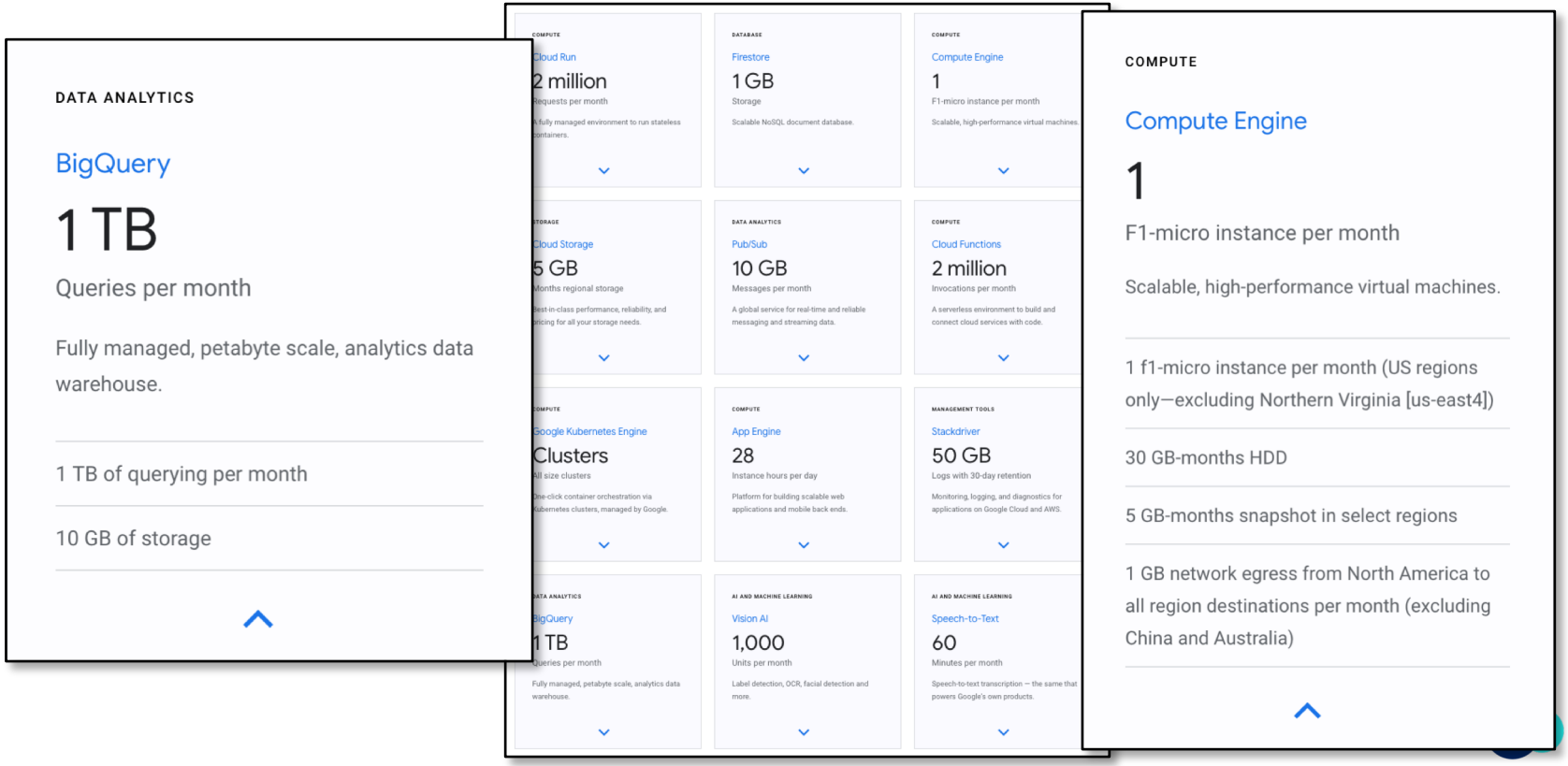
Running workflows on ISB-CGC



ISB-CGC does not make the choice for users but instead enables as many workflow technologies as possible through *documentation*, *support*, and where necessary *infrastructure*.

ISB

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



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

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

ISB

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.

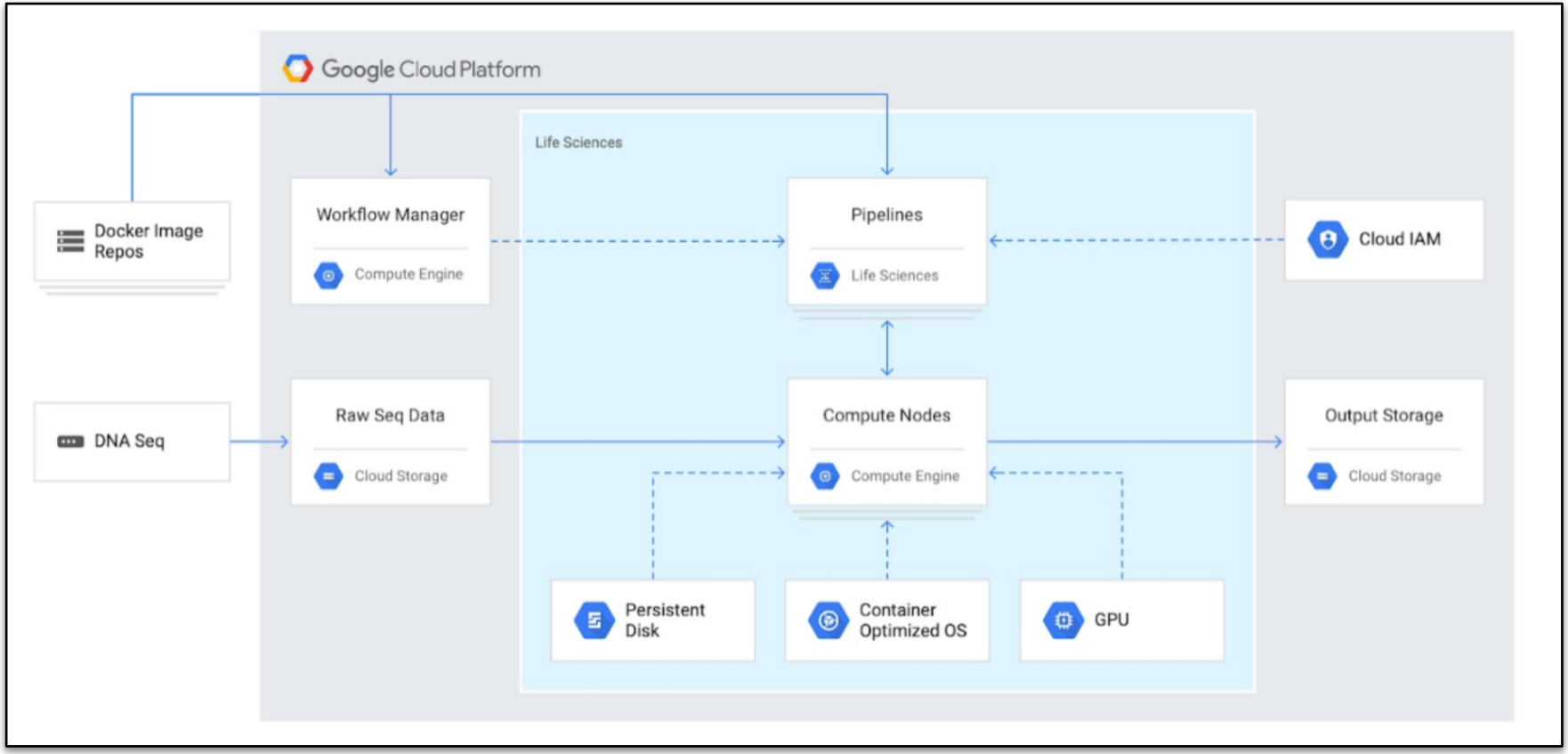
Ready for AI / ML

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

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

ISB

Google Cloud Life Sciences



https://cloud.google.com/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 Learnir
are two availble model types, linear and logistic. The first, linear regression, models a
variable given a selection of variables, both categorical (*eg* US postal code) and num
or weight). The second, logistic regression, models a binary label given some variable
for classification between two groups, which we have used extensively in this blog. A
groups we created had features like 'does or does not have a mutation in GATA3'. Ev
logistic regression is regularized! We have both L1 and L2 regularization available, w
similar to an implementation of elasticnet.

In the following examples, I'm going to be working in the BigQuery web interface, bu
possible to train and apply these models using the command line tool (bq), the REST
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
docs, at this time (July 2018) pricing is still under development. But essentially it's si
other query. You're charged according to how much data is processed in training the
storage fees for the model (first 10GB free). However, the actual data read for traini
more than just the size of the table. It's not entirely clear at this point, but when I lea
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, incase you missed it a few months ago, is that the R package used for interacting with BigQuery, bigrquery, 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.

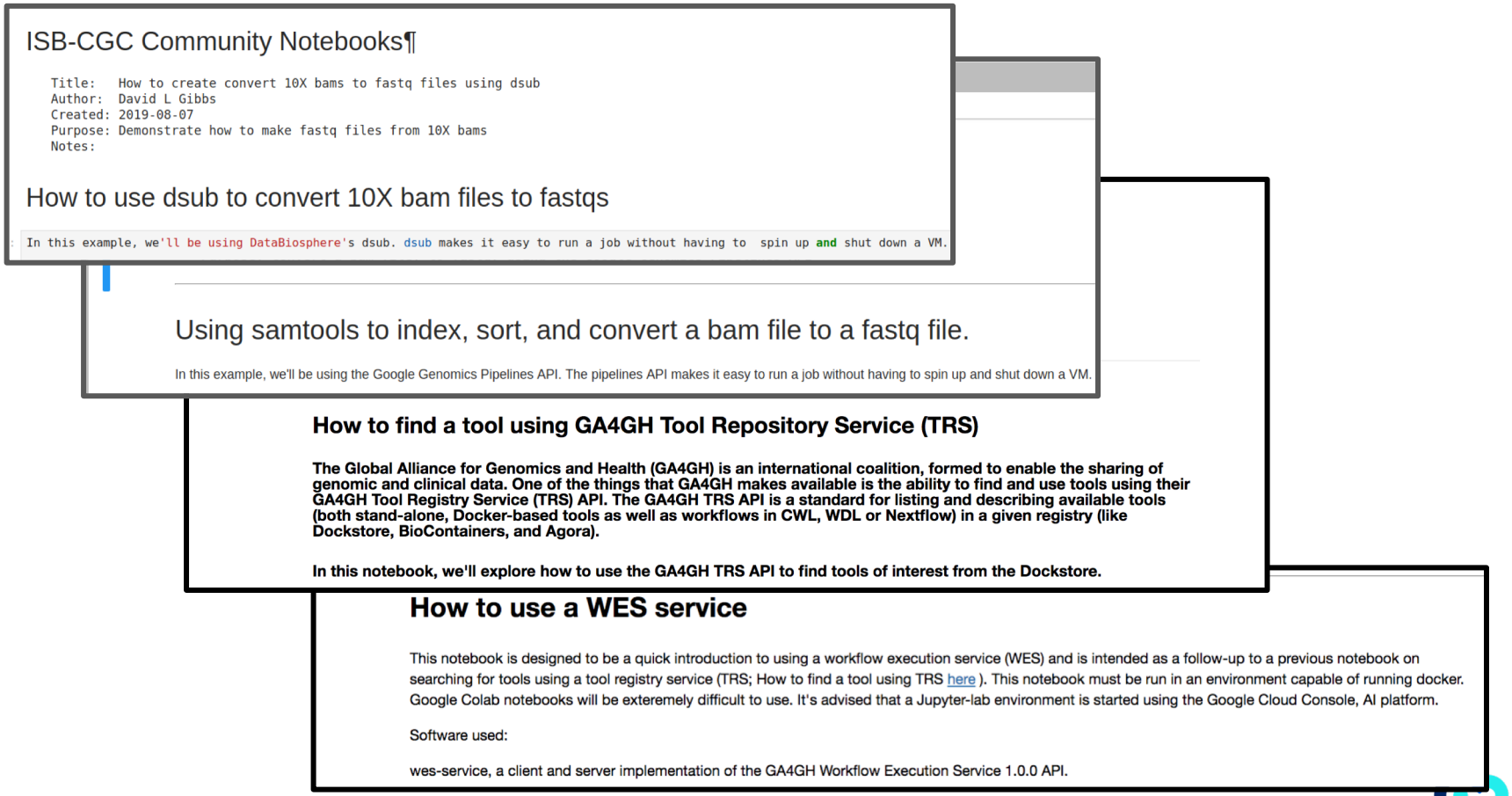
Working with BigQuery ML is quite a bit different than what we've done before. In the past, when working with BigQuery, we've computed different statistics, and we've even used those statistics for classification, but that work was all done in the SQL – including, for example, formulating a Z-score in 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).

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

ISB

Example workflows launched from cloud-based notebooks



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.

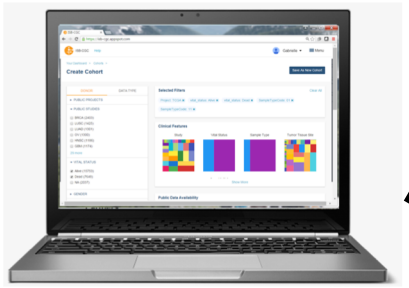
Today `dsub` supports Google Cloud as the backend batch job runner, along with a local provider for development and testing. With help from the community, we'd like to add other backends, such as a Grid Engine, Slurm, Amazon Batch, and Azure Batch.

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



BigQuery

ISB-CGC hosted tables

User Tables

Cloud Life Sciences

Cloud Storage

Cloud-hosted cancer data

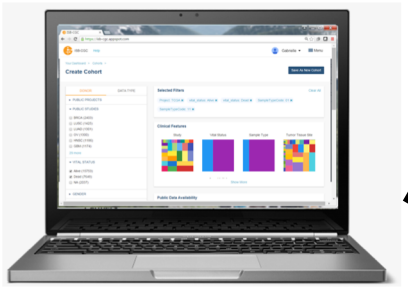
User Data

Compute Engine

VM

ISB

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

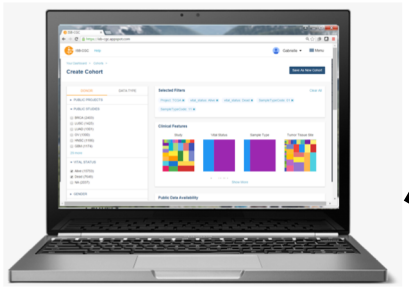
User Data

Compute Engine

VM

ISB

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

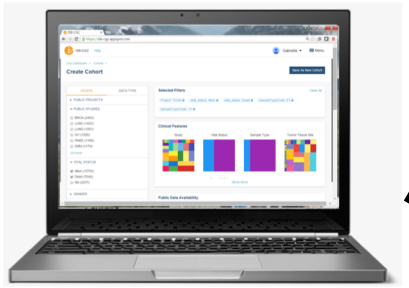
User Data

Compute Engine

VM

ISB

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

Example RNASeq quantification workflow using a Docker image

STAR_alignment_run.sh — Desktop

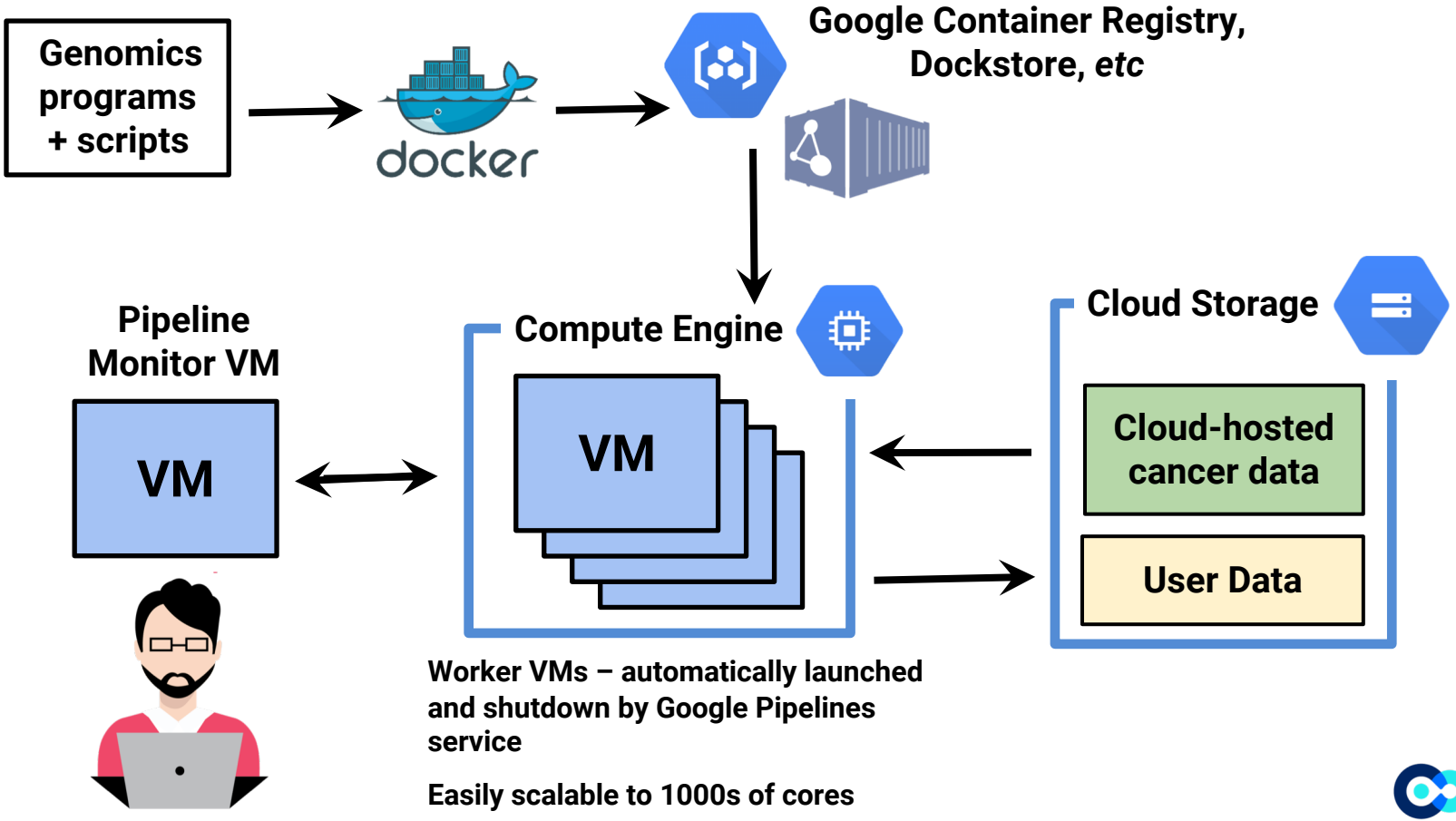
Add License

```
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"
```

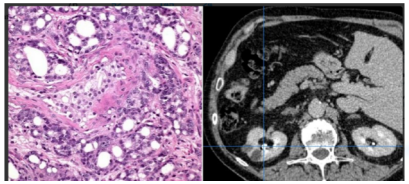
Line: 31:30 Shell Script (Bash) Soft Tabs: 4

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

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

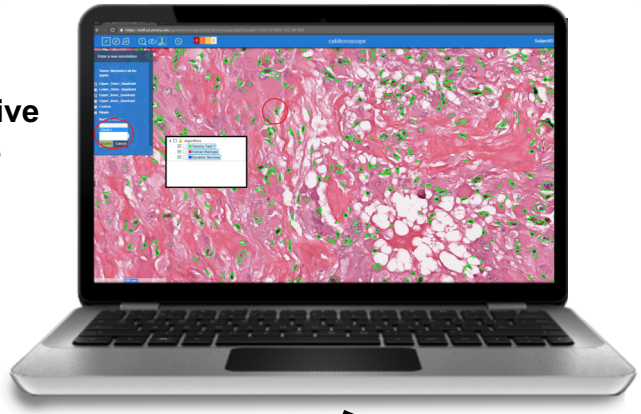


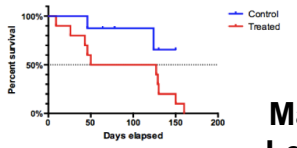
Performing deep learning and integrative analysis on pathology images

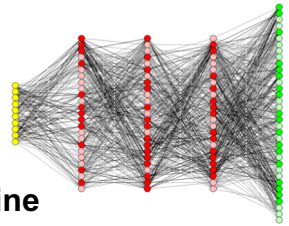


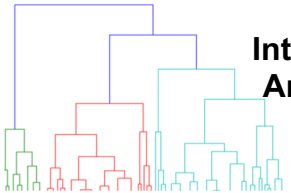
Pathology and Radiology images in Cloud Storage

Interactive Tools









Machine Learning & Integrative Analyses

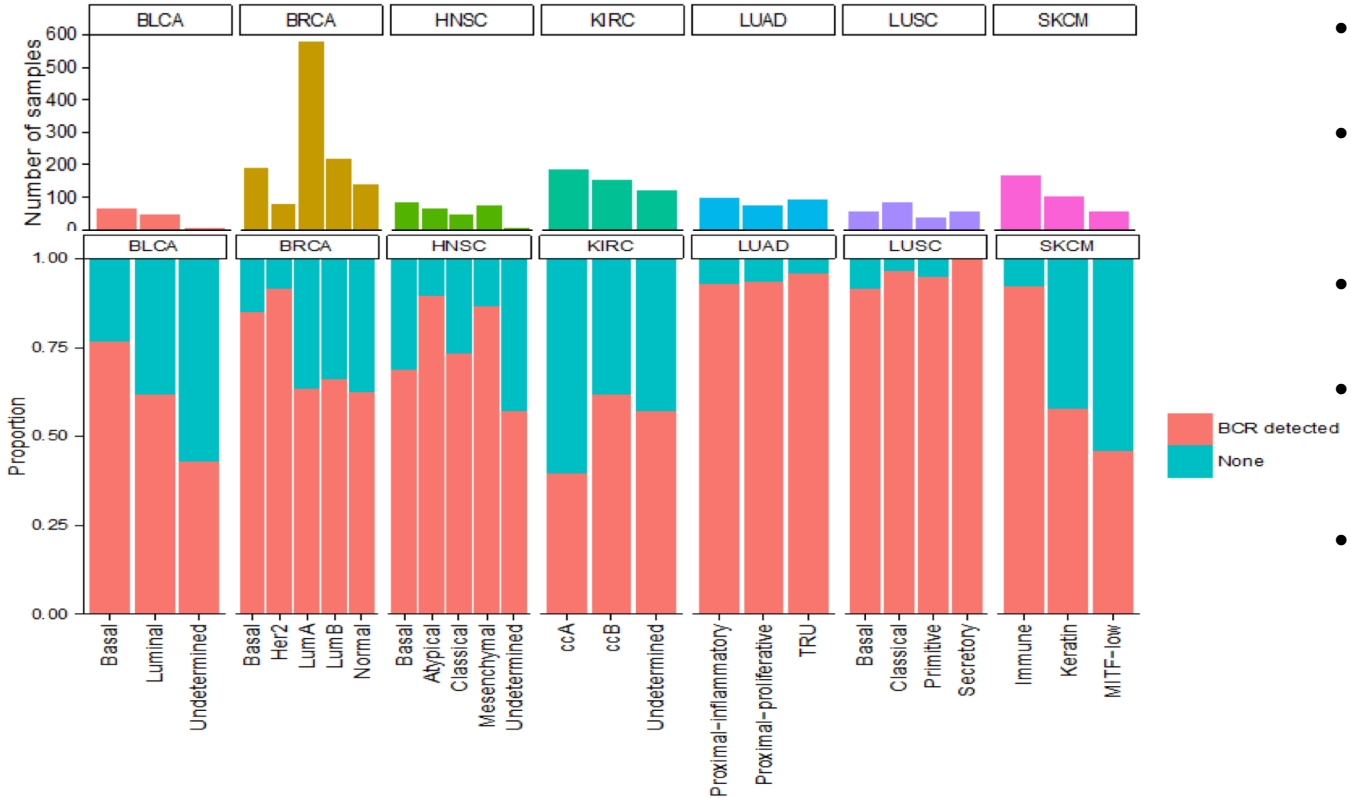
BIG QUERY

Automated Segmentation & Feature Extraction

Saltz et al. Cell Rep. 2018 Apr 3;23(1):181-193.e7

ISB

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

Mose, LE et al Bioinformatics. 2016 Dec 15;32(24):3729-3734

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

Tatlow PJ, Piccolo SR. Sci Rep. 2016 Dec 16;6:39259.

ISB

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

https://cloud.google.com/preemptible-vms

ISB

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. 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

Human Mutation
Variation, Informatics, and Disease
HGVS HUMAN GENOME VARIATION SOCIETY
Explore this journal >

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

101/160960

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

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

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