Cancer -omics analysis on the ISB-CGC platform

2020-03-17

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

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

[Diagram showing the ecosystem of data contributors, consumers, and services related to cancer research data.]

Legend:
- Blue: Available to researchers
- Green: Development
- White: Future Nodes

Data Contributors and Consumers:
- Biomedical Researchers
- Tool Developers
- Data Scientists

Authentication & Authorization:
- NCI Cloud Resources
- APIs

DCF Digital ID / Metadata Services
- IndexD

Domain-specific Data Model
- Cloud-based Data Repository
- Node Portal

Institute for Systems Biology

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

NIH NATIONAL CANCER INSTITUTE

#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
Our Approach at ISB-CGC

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

raw data
- fastq, bam, image files, etc

tabular data
- clinical, molecular, reference, metadata

Google Cloud Storage

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

- Cloud-based web service from Google Cloud used for handling and analyzing big data
- In the world of “omics”, it can facilitate high-throughput data analysis on the Cloud inexpensively in the following ways:
  
  **Storage:**
  - Store the results from large-scale pipelines/workflows in centralized BigQuery tables
  - First **10 GB** of storage per month are free. **$0.02 per GB** thereafter (e.g. store VCFs, MAFs, tab-delimited files)

  **Analysis:**
  - Use standard SQL to query large -omics data, the first **TB** of query data is free a month. **$5.00 per TB** of queries thereafter.
  - Preview or interrogate data without worrying about downloading data file by file
  - Seamlessly integrate BigQuery tables with commonly used data analysis tools including R and Jupyter notebooks
Attributes of Google BigQuery that make it ideal for use in research

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

<table>
<thead>
<tr>
<th>DNASeq (MAF, VCF)</th>
<th>Clinical &amp; Biospecimen</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA methylation</td>
<td></td>
</tr>
<tr>
<td>Protein (RPPA)</td>
<td>miRNAseq</td>
</tr>
<tr>
<td>RNASeq (gene, isoform, exon, junction, etc)</td>
<td>SNP array (genotype calls, allele- and segment-copy-number values)</td>
</tr>
</tbody>
</table>
Over 300 open access BigQuery tables hosted by ISB-CGC

- Derived (analyzed) molecular datasets (TCGA, TARGET, CCLE)
  - Expression (RNA, protein), copy number, mutations, methylation, clinical, etc.

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

- Metadata tables
  - Indexes of files, Google Cloud file paths, case ID, etc.
Multiple easy avenues for computing on data on ISB-CGC

ISB-CGC enables full command line access to analyze cloud hosted data via a collection of powerful tools and technologies along with the ability to install your own tools.
Some example use-cases of the three entry points to ISB-CGC

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

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

**Interactive big data exploration and analysis**
- Interactive data exploration in BigQuery
- Use R or Python to perform custom multivariate analyses
- Develop and customize bioinformatics tools and pipelines
  - *etc…*
We provide data exploration tools through our web app

ISB-CGC web tools

Google BigQuery

Google VMs
Building cohorts using the ISB-CGC web app
ISB-CGC: Interactive Apps
Integrated visualization methods for Big Data

- Integrated genome viewer (view read pile-ups)
- caMicroscope (view histology)
- OHIF (view radiology)

---

**Integrated Visualization Tool:**
---

**Program** | **Case Barcode** | **Disease Code** | **Exp. Strategy** | **Platform** | **Data Category** | **Data Type** | **Data Format** | **File Size**
--- | --- | --- | --- | --- | --- | --- | --- | ---
TCGA | TCGA-OL-A680 | BRCA | Genotyping array | Affymetrix SNP Array 6.0 | Simple nucleotide variation | Genotypes | TXT | 20.9 MB
--- | --- | --- | --- | --- | --- | --- | --- | ---
TCGA | TCGA-OL-A680 | BRCA | Genotyping array | Affymetrix SNP Array 6.0 | Simple nucleotide variation | Genotypes | TXT | 20.9 MB
--- | --- | --- | --- | --- | --- | --- | --- | ---
TCGA | TCGA-OL-A680 | BRCA | RNA-Seq | Illumina HiSeq | Raw sequencing data | Aligned reads | BAM | 7.8 GB
--- | --- | --- | --- | --- | --- | --- | --- | ---
TCGA | TCGA-OL-A680 | BRCA | WX5 | Illumina HiSeq | Raw sequencing data | Aligned reads | BAM | 4.9 GB
--- | --- | --- | --- | --- | --- | --- | --- | ---
TCGA | TCGA-OL-A680 | BRCA | WX5 | Illumina HiSeq | Raw sequencing data | Aligned reads | BAM | 7.2 GB
--- | --- | --- | --- | --- | --- | --- | --- | ---
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/bg_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.
Three entry points for exploring cancer data on ISB-CGC

ISB-CGC WebApp

Google BigQuery

Google VMs
VMs enable advanced bioinformatic workflows
Some workflows we’ve enabled for ISB-CGC end-users

Multiple PanCancer Atlas projects, including:
- Germline-variant calling
- Fusion gene analysis
- T-cell and B-cell receptor analysis
- Viral DNA screening
- MYC pathway analysis (BQ)
- 8-oxoG filtering (MC3 project)

Other end-user projects include:
- SMC-RNA Dream challenge (supporting both the organizers and many participants)
- Tumor-specific alternative polyadenylation
- ML algorithm evaluation & benchmarking
- RNA seq alignment to novel transcriptome(s)
- mRNA expression quantitation
- Targeted de-novo assembly
- Structural variations (WGS + SNP6 data)
- Metagenomics / cancer analysis
- Statistical meta-analysis of miRNAs in cancer
- Code/tutorial development
- GDC hg38 TCGA miRNA QC (w/ BCGSC)
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.

- Pre-built VM images
- BigQuery and bigQueryR
- googleAuthR
- R
- Shiny by RStudio
- Bioconductor
- IPython
- Jupyter
- Cloud notebooks and workspaces.
- Cloud Datalab
Analyze correlation between TCGA samples & GTEx tissue types quickly and cheaply!
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
Use Google BigQuery to easily connect your research to public datasets

ISB-CGC and Other Public Datasets

SQL

Private User Data and Derived Results
A typical work setup across multiple browser tabs

Google web interface

Notebook (R or Python)

Searchable web docs

- Built in syntax checking
- Integrate with notebooks to generate your own publication quality visuals
- Google web interface
- Notebook (R or Python)
- Searchable web docs

A typical work setup across multiple browser tabs
What you need to know to interoperate with ISB-CGC

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

- Service Account is the Authentication and Authorization method that researchers’ 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
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
Some example typical ISB-CGC use-cases...

1) Fire up VMs to run pipelines using any workflow language of your choice

2) Build cohorts on the web-app and download file manifests with locations of files to use for analyses

3) The ISB-CGC Gold Standard Use-Case (featured in our demo)
   a) Use BigQuery to identify useful public data
   b) Transition to notebooks to perform multivariate analysis
   c) Leverage public data analysis tools (i.e., bioconductor)
   d) Combine your own data with public data seamlessly
   e) Generate beautiful figures
Questions?
ISB-CGC Team

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

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