Getting started on the ISB-Cancer Genomics Cloud (ISB-CGC) powered by the Google Cloud Platform

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.
More about ISB-CGC, ISB-CGC Main Landing Page, Full documentation of the ISB-CGC platform, and FAQs.

Benefits of using the cloud: You don't have to download the data! Bring your compute and know-how to the data. Use cloud-native, compute scale as big as you can imagine, tools to analyze TBs and PBs of data!!

Goals of this tutorial:

1. How to get started on the ISB-CGC powered by the Google Cloud Platform (GCP)
2. Learn about cloud credits offered by Google and ISB-CGC
3. Learn about data hosted on ISB-CGC
4. Learn about how to get authorization to access controlled cancer genomics data
5. Setting up and registering a GCP project to use controlled access data
6. Enabling required GCP APIs
7. Learn about GCP's BigQuery tool
8. Learn how to run analyses on data stored in BigQuery or stored in Google Cloud Storage (GCS)

Take home message: Don't be intimidated by the cloud! Bring your computation to the data on ISB-CGC. If you've conducted bioinformatics analyses before using the command line or SQL, this will be just as easy (if not easier).

I. Getting Started:

1) ISB-CGC hosts both open-access and controlled-access cancer genomics data from the NCI.
   About ISB-CGC Cloud-Hosted Datasets

2) To access controlled-access data, dbGaP authorization is required.
   Accessing Controlled-Access Data and acquiring dbGaP authorization.

3) To work in GCP, you must first set up a GCP Project:
   - 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 ISB-CGC Cloud Credits Available for Researchers.
- How to request ISB-CGC Cloud Credits.

4) Registering the GCP project
5) Enable Required Google Cloud APIs

II. Accessing and analyzing data via BigQuery

- BigQuery is Google's native big data analysis tool. It is a serverless, highly scalable data warehouse tool that allows researchers to find meaningful insights from data using standard SQL queries CHEAPLY, and FAST!
- ISB-CGC has leveraged this powerful tool and uploaded multiple cancer genomics datasets into BigQuery tables that are open to the public. ISB-CGC Datasets in BigQuery and the always freshly updated Data Release Notes and Future Plans.
- To obtain access to the ISB-CGC project tables in BigQuery, users can link these tables to your GCP project as described here.
- ISB-CGC provides tutorials and walkthroughs on how to access BigQuery from the web-UI, programmatically in R, or through Google's native Jupyter notebook Cloud Datalab, and python examples.
- Every month, ISB-CGC provides an example analysis of cancer genomics data using BigQuery in our Query of the Month blog.

III. Accessing and analyzing data stored in GCS

- All open-access data on ISB-CGC are stored in a publically available GCS bucket (gs://isb-cgc-open).
- All controlled-access data are stored in Google Cloud Storage (GCS) in their original form as obtained from the GDC.
- To access controlled data, users must first be authenticated by NIH (via the ISB-CGC web-app). Upon successful authentication, user dbGaP authorization will be verified. These two steps are required before the user's Google identity is added to the access control list (ACL) for the controlled data. At this time, this access must be renewed every 24 hours.
- Summary of data types and format available

- Working with large-scale data hosted by the ISB-CGC in Google Cloud Storage requires some familiarity with tools such as the Google Cloud SDK, Google Compute Engine, Virtual Machines and Docker.
- Cheat-sheets and slides on computing in the cloud including how to access files stored on GCS can be found here.