Poster 3547: The ISB Cancer Gateway in the Cloud (ISB-CGC): Access, explore and analyze large-scale cancer data through the Google Cloud

Fabian Seidl[1]; Lauren Hagen[2]; Jacob Wilson[1]; Boris Aguilar[2]; Deena Bleich[1]; Lauren Wolfe[2]; Poojitha Gundluru[1]; Prema Venkatesan[1]; Mi Tian[2]; Suzanne Paquette[2]; Elaine Lee[2]; Danna Huffman[1]; David Pot[1]; William Longabaugh[2]

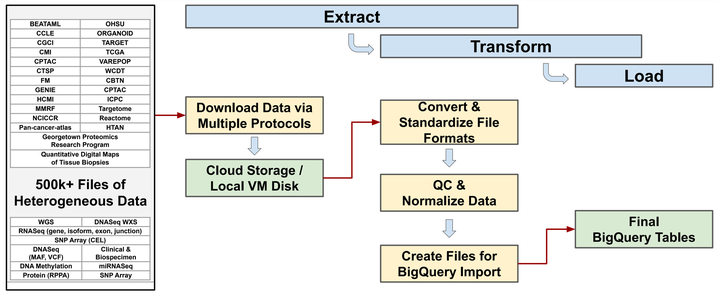
[1]General Dynamics Information Technology; [2]Institute for Systems Biology

isb-cgc.org

Abstract

Rapid growth of cancer data in recent decades has made data discovery and wrangling difficult for the average cancer research lab. Our mission at the ISB Cancer Gateway in the Cloud (ISB-CGC), part of the NCI's Cancer Research Data Commons ecosystem, is to democratize access to large cancer datasets. Funded by the NCI, we have performed ETL processes on data from GDC and PDC projects such as TCGA, TARGET, and CPTAC. We generated hundreds of BigQuery tables containing data such as mutations, gene expression, and protein abundance, which enable data analysis in the cloud via SQL. BigQuery analyses are inexpensive and rapid even when scaled to petabyte sized inputs, for example we ran 6.6 billion correlations in 2.5 hours with a total cost of about one dollar. These data can also be accessed affordably from Google Cloud VMs where researchers can develop analysis pipelines in Python, R, and workflow languages such as CWL. We present two recent collaborations: In one BigQuery was used to develop machine learning algorithms that calculated genetic risk scores from TCGA glioblastoma and ovarian cancer copy number variation. In another example researchers combined SQL queries of our BQ tables with data from the ISPY2 Trial initiative and generated an R shiny app that can dynamically create data visualizations for genes of interest in different TCGA cohorts.

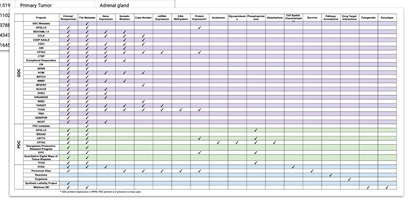
A BigQuery Ecosystem housing cancer data



A diagram of our Extract, Transform, and Load process used to populate our BigQuery Ecosystem with data sourced from the Genomic Data Commons. We first call the GDC API and populate the program metadata including subject clinical data, experimental strategies, file locations, etc. We then use these metadata to aggregate derived metadata tables such as RNAseq, DNA methylation, etc.

An example of a Subset of RNA sequencing data stored in BigQuery.

Row	aliquot_barcode	gene_name	gene_type	unstranded	fpkm_unstranded	sample_type_name	primary_site
1	TCGA-OR-A5J1-01A-11R-A295-...	AKT2	protein_coding	25902	35.5365	Primary Tumor	Adrenal gland
2	TCGA-OR-A5J1-01A-11R-A295-...	IL10RB	protein_coding	37	0.1784	Primary Tumor	Adrenal gland
3	TCGA-OR-A5J1-01A-11R-A295-...	CLTA	protein_coding	9305	80.7183	Primary Tumor	Adrenal gland
4	TCGA-OR-A5J1-01A-11R-A295-...	CRBN	protein_coding	1352	3.656	Primary Tumor	Adrenal gland
5	TCGA-OR-A5J1-01A-11R-A295-...	C3orf33	protein_coding	189	2.0007	Primary Tumor	Adrenal gland
6	TCGA-OR-A5J1-01A-11R-A295-...	PRPS2	protein_coding	1092	8.6082	Primary Tumor	Adrenal gland
7	TCGA-OR-A5J1-01A-11R-A295-...	ZW10	protein_coding	573	4.2871	Primary Tumor	Adrenal gland
8	TCGA-OR-A5J1-01A-11R-A295-...	CYP2C8	protein_coding	14	0.1188	Primary Tumor	Adrenal gland
9	TCGA-OR-A5J1-01A-11R-A295-...	AL353751.1	lncRNA	34	0.3017	Primary Tumor	Adrenal gland
10	TCGA-OR-A5J1-01A-11R-A295-...	PDK1	protein_coding	1714	2.519	Primary Tumor	Adrenal gland
11	TCGA-OR-A5J1-01A-11R-A295-...	ABHD14A	protein_coding	448	4.1102	Primary Tumor	Adrenal gland
12	TCGA-OR-A5J1-01A-11R-A295-...	SUGT1-DT	lncRNA	151	3.3788	Primary Tumor	Adrenal gland
13	TCGA-OR-A5J1-01A-11R-A295-...	GTF2E1	protein_coding	3533	21.4342	Primary Tumor	Adrenal gland
14	TCGA-OR-A5J1-01A-11R-A295-...	TNFSF10	protein_coding	1130	7.1645	Primary Tumor	Adrenal gland



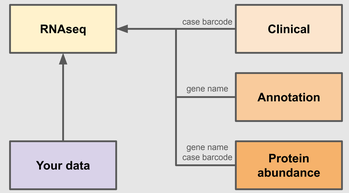
ISB-CGC hosts a wide variety of derived and annotation data, this is a snapshot of our current catalogue that is always expanding. **Requests welcomed.**

An example of the TCGA RNA expression table in our ecosystem. Data can be browsed in a columnar format, similar to Excel. You can also query the tables in SQL to perform operations such as filtering, set operations, or overview level statistics such as mean, minimum, and maximum.

ISB-CGC hosts more than one thousand tables from a large variety of large cancer initiatives all open access. We have a Search Tool on our Portal (isb-cgc.org) to assist in finding the right tables for your uses.

Co-analysis of multiple data types in BigQuery

Within our BigQuery ecosystem it is possible to join between the many data types based on key fields such as case identifiers, gene name, cancer type and more. These joins can generally be made in less than a minute for cents worth of cloud costs.



Storing derived data in BigQuery enables rapid data discovery and quick comparisons between data types via "table joins". Below is an example of a set of queries slightly modified for simplicity joining gene expression data to clinical data and protein abundance.

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current`
WHERE <conditionals>
```

The initial query selects specific fields from the table of interest, in this case upper quartile normalized fpkm.

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current`
JOIN `isb-cgc-bq.TCGA.clinical_gdc_current`
WHERE <conditionals>
```

The second query performs a join between the expression and clinical tables, in this case selecting for cigarettes smoked per day.

```
SELECT
 <fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current`
JOIN `isb-cgc-bq.TCGA.clinical_gdc_current`
JOIN `isb-cgc-bq.TCGA.protein_expr_hg38_gdc_current`
WHERE <conditionals>
```

Query results — SAVE RESULTS — EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

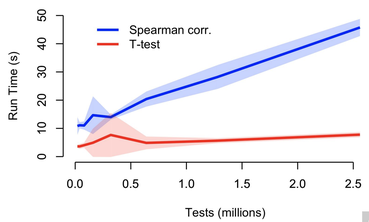
Row	case_barcode	fpkm_uq_unstranded	exp._cigarettes_per_	protein_expression
1	TCGA-86-A4P7	8.9616	null	-0.6541634745
2	TCGA-91-6829	14.4956	5.178082191780...	0.399470062
3	TCGA-91-6828	12.558	null	-0.1746171102
4	TCGA-86-A4P8	1.947	null	-0.4681542825
5	TCGA-38-4629	40.1784	5.479452054794...	0.7393313331
6	TCGA-38-6178	10.7342	null	-0.013338784
7	TCGA-78-7166	32.8658	2.082191780821...	-0.0290081565
8	TCGA-78-7167	3.7353	3.506849315068...	-0.438169383

Results per page: 50 — 1 – 50 of 378

Output of the final query joining between RNA expression, protein abundance, and a selected clinical field, ready for statistical tests and Machine Learning

BigQuery is a powerful statistical tool

One reason to host bioinformatic data from the Cancer Research Data Commons in BigQuery is to improve accessibility and ease of exploration. However, BigQuery also allows for custom functions written in SQL or JavaScript and features automated compute scaling that allow it to outcompete traditional High Performance Compute clusters.



Shown is the average run time required to calculate Spearman correlations and T-tests in BigQuery. Even **2.5 million tests** completed in **less than 50 seconds** in this trial.

BigQuery compute is cheap and we offer **$300 cloud credits** for exploration and setup. Charges are based on data read rather than compute used. In a trial running **6.6 billion tests** cost **$1.16**.

We have generated custom user defined functions for commonly used statistical tests and incorporated these tests into example notebooks.

These notebooks as well as many others are a public teaching resource for researchers. isb-cancer-genomics-cloud.readthedocs.io/

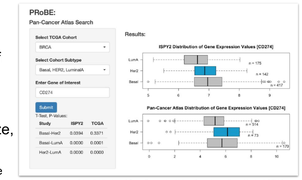
Data type 1	Data type 2	Statistical test/notebook
Gene expression	Clinical	Kruskal Wallis score
Gene expression	Somatic mutation	T-test score
Gene expression	Gene expression	Spearman Correlation
Somatic mutation	Clinical	Chi Square test
Somatic mutation	Somatic Mutation	Fisher's exact test

Recent collaborations

ISB-CGC has enabled many researchers with moving their Workflows to the cloud and leveraging its unique capabilities to answer their research questions. Here we show two examples of such collaborations.

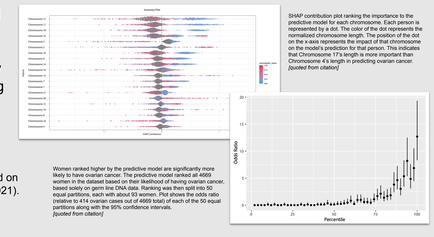
Shown is a screenshot of a web tool developed using ISB-CGC and a BigQuery cloud-based platform to collect, visualize, analyze, and share data in the context of a clinical trial.

O'Grady N, et al. PRoBE the cloud toolkit: finding the best biomarkers of drug response within a breast cancer clinical trial. *JAMIA Open*. (2021) 10.1093/jamiaopen/ooab038



An example plot of the distribution of Gene Expression Values from data collected in the I-SPY 2 Trial and TCGA Cohorts. *[quoted from citation]*

Another project we are highlighting developed genetic risk scores based on chromosomal-scale length variation of germline DNA, using Affymetrix SNP 6.0 array data and Copy Number Variation for predicting whether or not a woman will develop ovarian cancer.

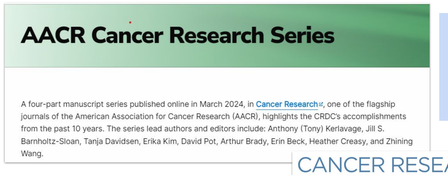


SHAP contribution plot ranking the importance to the predictive model for each chromosome. Each person is represented by a dot. The color of the dot represents the normalized chromosome length. The position of the dot on the x-axis represents the impact of that chromosome on the model's prediction for that person. This indicates that Chromosome 17's length is more important than Chromosome 4's length in predicting ovarian cancer. *[quoted from citation]*

Toh, C., Brody, J.P. Genetic risk score for ovarian cancer based on chromosomal-scale length variation. *BioData Mining* **14**, 18 (2021). 10.1186/s13040-021-00253-y

Women ranked higher by the predictive model are significantly more likely to have ovarian cancer. The predictive model ranked all 4669 women in the dataset based on their likelihood of having ovarian cancer, based solely on germ line DNA data. Ranking was then split into 50 equal partitions, each with about 93 women. Plot shows the odds ratio (relative to 414 ovarian cases out of 4669 total) of each of the 50 equal partitions along with the 95% confidence intervals. *[quoted from citation]*

Cancer Research Data Commons wide publications



A recent set of four publications describing the CRDC with detailed information about the organization and capabilities can be found at:

datacommons.cancer.gov/publications/aacr-cancer-research

The article highlighted here specifically focuses on the Cloud Resources (CRs) with a plethora of detailed information in the Supplementary Materials.

Summary and Conclusions

- BigQuery is a tool with cloud-powered scaling of Microsoft Excel functionality
  - Affordable storage and sharing of YOUR tabular data
  - Rapid data exploration and quick statistics natively
  - Derived data from well known reference NCI sets and annotations for co-analysis
  - Fast links between diverse data types
  - Advanced statistical analyses using Python, R, SQL, and Bioconductor
  - Rapidly able to expand to Machine Learning
- Easy exploration of existing GDC and PDC data
- Access Virtual Machines and controlled data for customized pipelines
- Multiple specialized databases such as Mitelman DB of Chromosomal Aberrations and Gene Fusions in Cancer
- Receive $300 pilot funding and more available for your pilot project

Funding

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

If you have any questions about our resources or would like to collaborate please email us:

Office hours twice weekly feedback@isb-cgc.org @isb_cgc