

Poster 1067: The ISB Cancer Gateway in the Cloud (ISB-CGC): Access, explore and analyze large-scale cancer data through the Google Cloud



Fabian Seidl¹; Jacob Wilson¹; Boris Aguilar²; Lauren Hagen²; Deena Bleich¹; Lauren Wolfe²; Poojitha Gundluru¹; George White²; Suzanne Paquette²; Elaine Lee²; Danna Huffman¹; David Pot¹; William Longabaugh²

¹General Dynamics Information Technology; ²Institute for Systems Biology

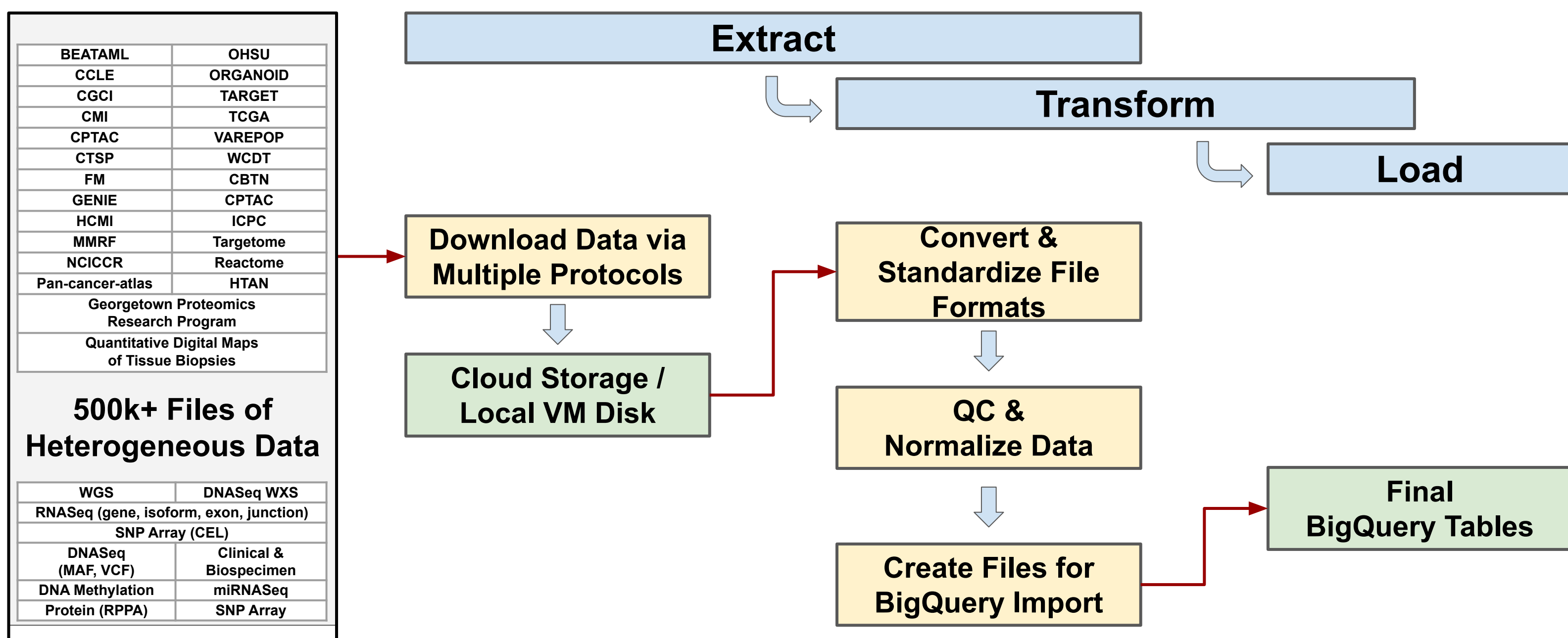
GENERAL DYNAMICS
Information Technology

isb-cgc.org

Abstract

- Rapid growth of cancer data in recent decades has made data discovery and analysis difficult for cancer researchers. ISB-CGC's mission is to **democratize access to large NCI cancer datasets**.
- Data from NCI genomic and proteomic projects such as **TCGA**, **TARGET**, and **CPTAC** are populated into hundreds of **tabular compute ready tables of mutations, gene expression, and protein abundance**, which enable integrative data analysis in the cloud via SQL.
- The data can be accessed affordably from Google Cloud VMs where researchers can develop analysis pipelines in **Python, R, and Bioconductor**.
- 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**.
- In this poster, we highlight some of the features of ISB-CGC and the hosted data, and present a collaborative example of using the cloud to **analyze cytogenetics data**.

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 populate the program metadata through the GDC API including subject clinical data, experimental strategies, etc. We then use these metadata to aggregate derived processed tables such as RNAseq, DNA methylation, etc.

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

Row	sample_barcode	gene_name	gene_type	unstranded	fpkm_unstranded	sample_type_name	primary_site
1	TCGA-OR-A5J1-D1A-11R-A2S5...	AKT2	protein_coding	25802	35.5365	Primary Tumor	Adrenal gland
2	TCGA-OR-A5J1-D1A-11R-A2S5...	IL13RB	protein_coding	27	0.1784	Primary Tumor	Adrenal gland
3	TCGA-OR-A5J1-D1A-11R-A2S5...	CLTA	protein_coding	9355	80.7183	Primary Tumor	Adrenal gland
4	TCGA-OR-A5J1-D1A-11R-A2S5...	CHN1	protein_coding	1352	3.655	Primary Tumor	Adrenal gland
5	TCGA-OR-A5J1-D1A-11R-A2S5...	C3orf93	protein_coding	189	2.0007	Primary Tumor	Adrenal gland
6	TCGA-OR-A5J1-D1A-11R-A2S5...	PRK2	protein_coding	1592	6.6302	Primary Tumor	Adrenal gland
7	TCGA-OR-A5J1-D1A-11R-A2S5...	ZNF10	protein_coding	573	4.2871	Primary Tumor	Adrenal gland
8	TCGA-OR-A5J1-D1A-11R-A2S5...	CYP2C8	protein_coding	14	0.1188	Primary Tumor	Adrenal gland
9	TCGA-OR-A5J1-D1A-11R-A2S5...	AL333751.1	lncRNA				
10	TCGA-OR-A5J1-D1A-11R-A2S5...	FOK1	protein_coding				
11	TCGA-OR-A5J1-D1A-11R-A2S5...	ABHD144	protein_coding				
12	TCGA-OR-A5J1-D1A-11R-A2S5...	SUPT10T	lncRNA				
13	TCGA-OR-A5J1-D1A-11R-A2S5...	OTUS5	protein_coding				
14	TCGA-OR-A5J1-D1A-11R-A2S5...	TNFRSF10	protein_coding				

Below the table is a detailed view of the data structure, showing columns for Projects, Genes, and various data types. It includes a legend for the data types and a list of projects.

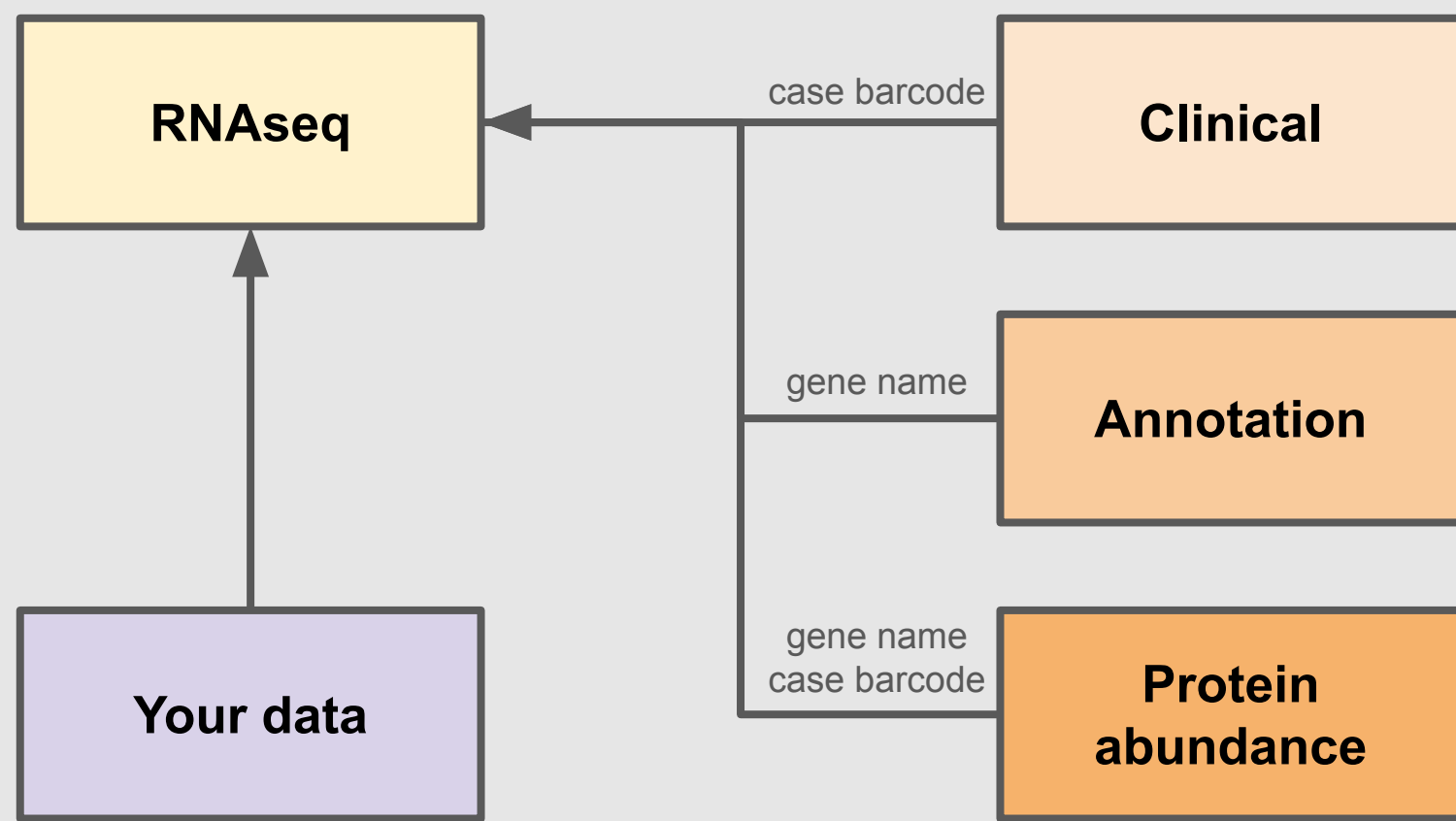
ISB-CGC hosts a wide variety of derived and annotation data, our catalogue is always expanding.
Requests welcomed.

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 IDs, gene name, cancer type and more. These joins can generally be made in less than a minute for cents worth of cloud costs.



Derived data in BigQuery enables rapid data discovery and quick comparisons 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`
JOIN `isb-cgc-bq.TCGA.protein_expr_hg38_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.

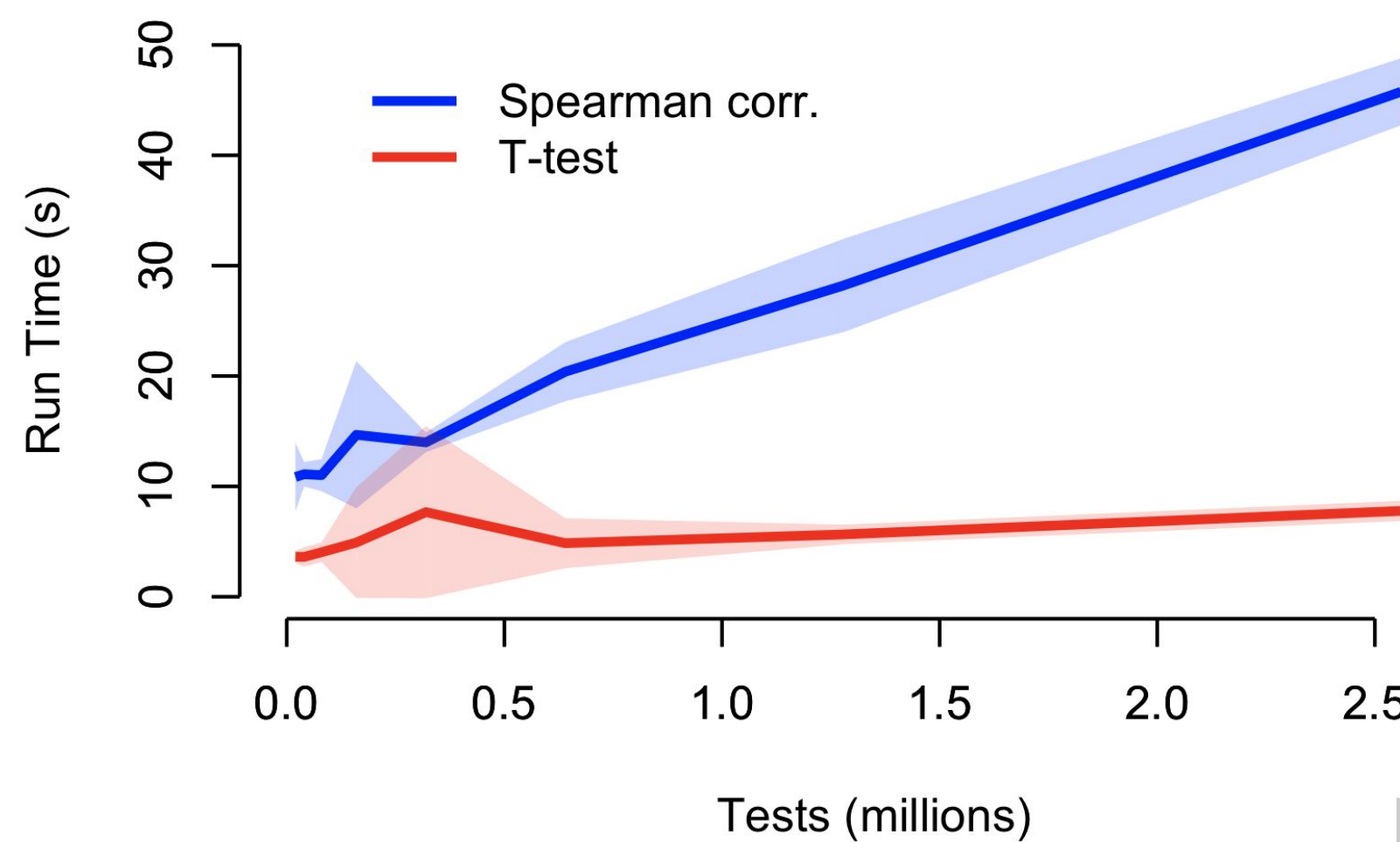
```
SELECT
<fields>
FROM `isb-cgc-bq.TCGA.RNAseq_hg38_gdc_current`
JOIN `isb-cgc-bq.TCGA.clinical_gdc_current`
JOIN `isb-cgc-bq.TCGA.protein_expr_hg38_gdc_current`
WHERE <conditionals>
```

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	-0.174617102
4	TCGA-86-A4P8	1.947	null	-0.4681542825
5	TCGA-38-4629	40.1784	5.479452054794...	0.799331331
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

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. BigQuery also allows for custom functions written in SQL or JavaScript and features automated compute scaling that allow it to outperform traditional High Performance Compute clusters.



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/

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**.

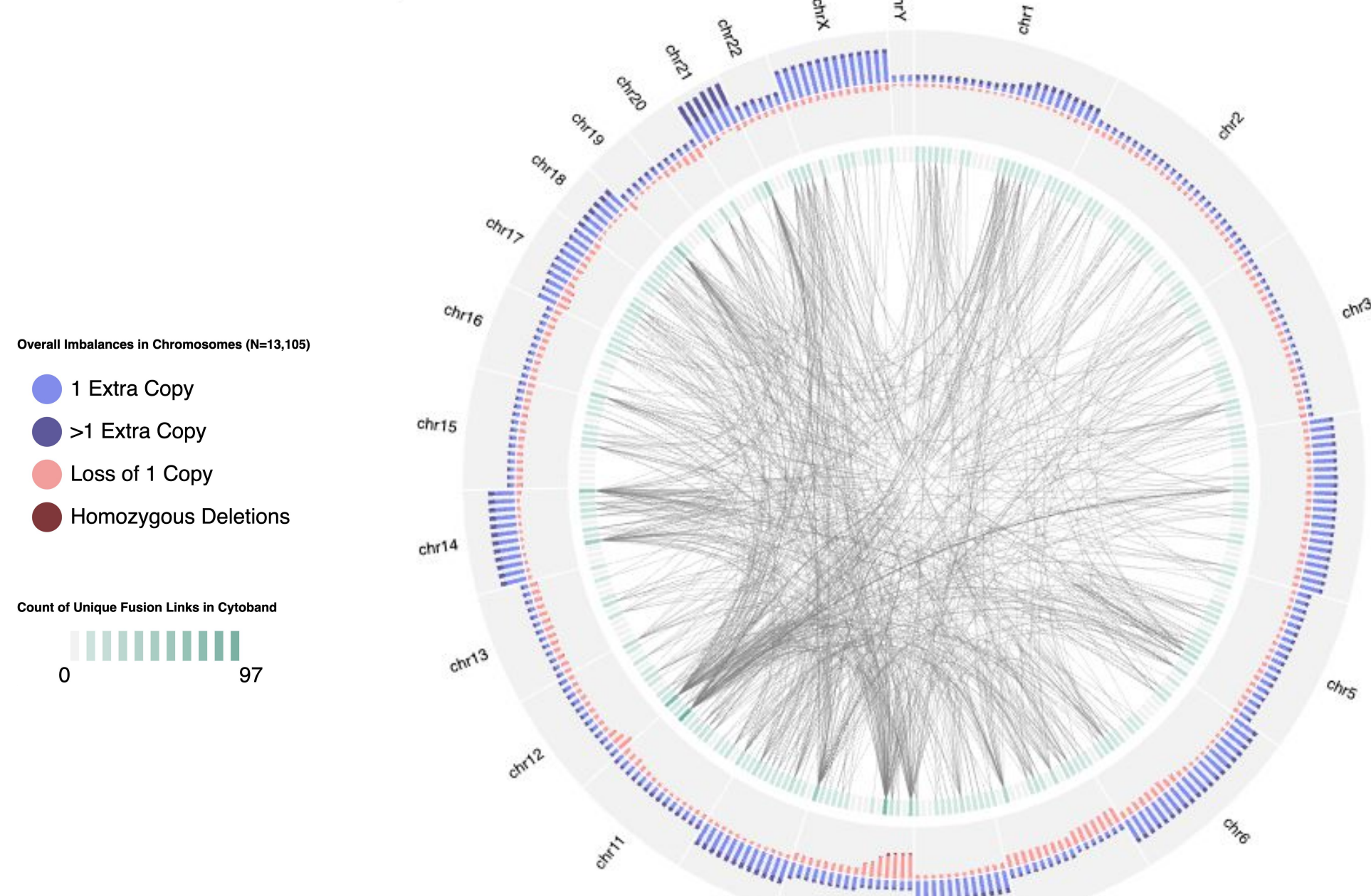
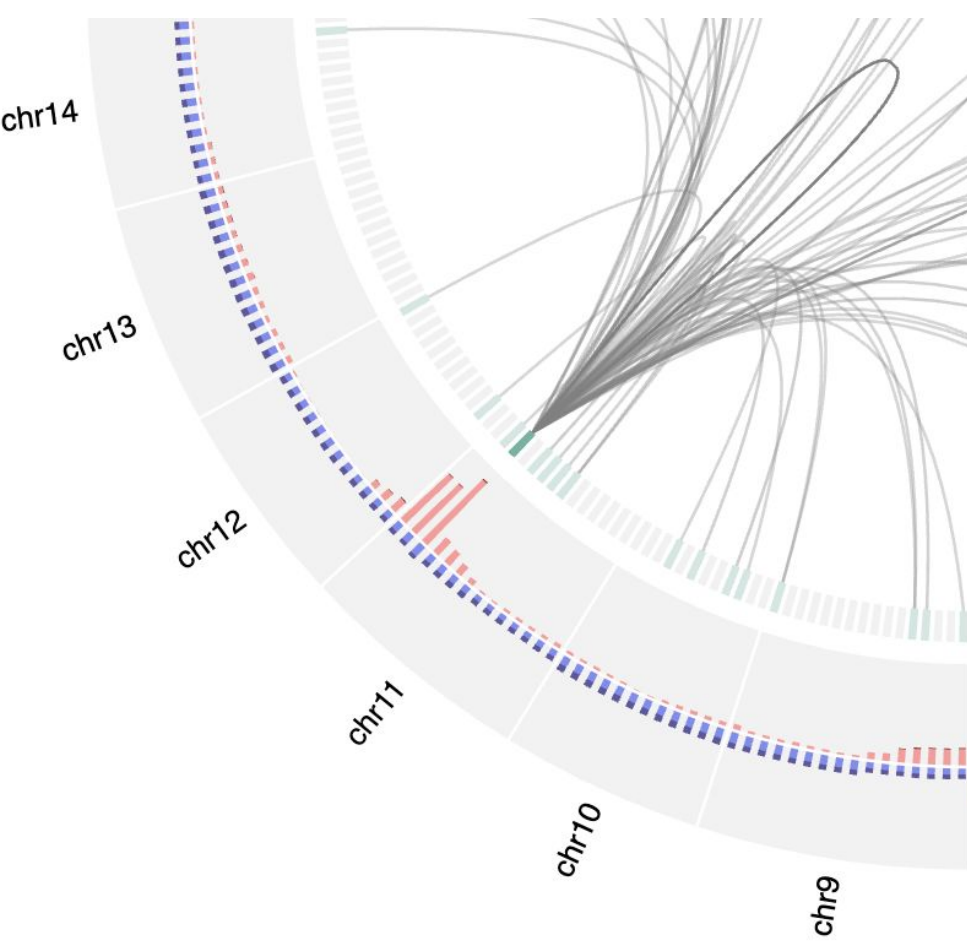
Data type 1	Data type 2	Statistical test/notebook
Gene expression	Clinical	Kruskal Wallis score
Gene expression	Somatic mutation	T-test score
Gene expression	Gene expression	Spearman Correlation
Somatic mutation	Clinical	Chi Square test
Somatic mutation	Somatic mutation	Fisher's exact test

Recent upgrades

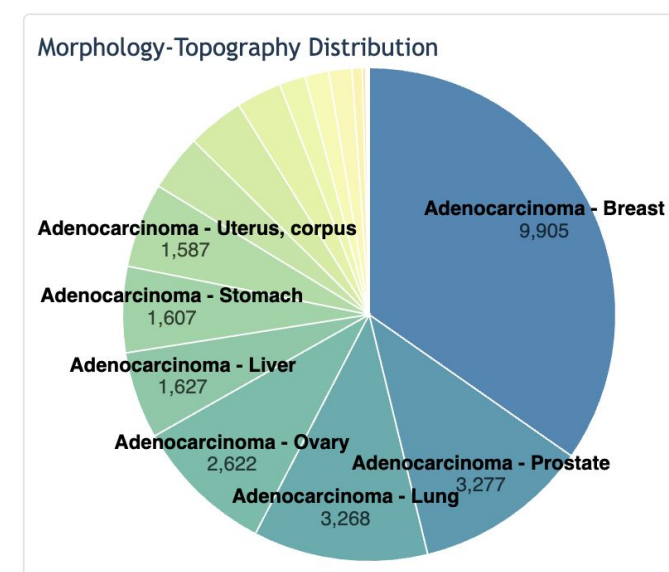
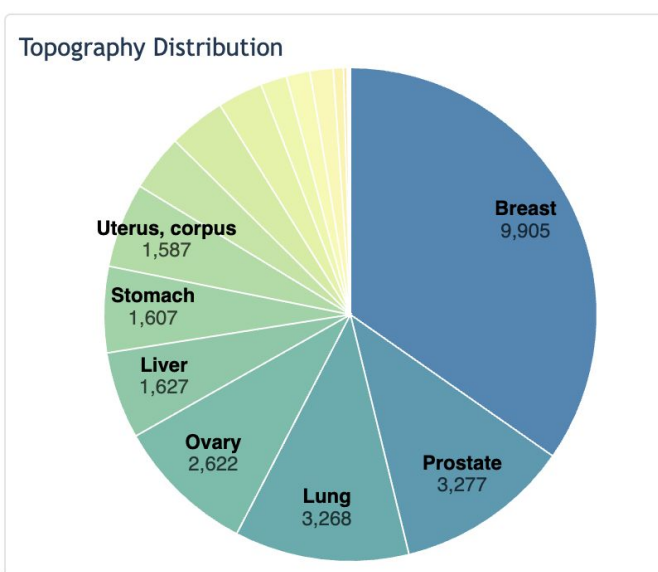
The Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer is a free-access database devoted to chromosomes, genes, and cancer. The information in the Mitelman Database relates cytogenetic changes and their genomic consequences, in particular gene fusions, to tumor characteristics, based either on individual cases or associations.

New Circos plots to visualize database contents.

The 11q23 chromosome band showing a high number of gene fusions in Acute Lymphoblastic Leukemia cases.



New charts and tables to view and compare morphology and topography distributions with gene fusions.



Example topography distributions for adenocarcinoma cases, and the table displays the most common gene fusions.

Fusion Gene	Tumor Morphology	Tumor Location	Count
TPM52::ERG	Adenocarcinoma	Prostate	11
KIF5B::RET	Adenocarcinoma	Lung	8
TPM3::NTRK1	Adenocarcinoma	Lung	8
EZR::ROS1	Adenocarcinoma	Lung	6
IRF2BP2::NTRK1	Adenocarcinoma	Lung	6
TPM3::NTRK1	Adenocarcinoma	Large intestine	6
CCDC6::RET	Adenocarcinoma	Lung	5
CD74::ROS1	Adenocarcinoma	Lung	5
ETV6::NTRK3	Adenocarcinoma	Lung	5
ETV6::NTRK3	Adenocarcinoma	Thyroid	5

Quarterly updates add thousands of new cases per year to the database. As of January 2025, the database consists of:

- 79,082 cytogenetic cases
- 49,643 unique cytogenetic aberrations
- 34,303 unique gene fusions
- 14,096 genes involved

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 unique 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 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. HHSN2612015000031.

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

Ask for help - we will get on a call.

feedback@isb-cgc.org

@isb_cgc