Analyzing HTAN scRNASeq data accessible in BigQuery with CellTypist

Fabian Seidl and Dar'ya Pozhidayeva for ISB-CGC 2024-04-02



Outline of today's agenda

- Short intro to CRDC, ISB-CGC and the Data Commons
- Data exploration in BigQuery ('Excel-like data tables in the cloud')
- Demonstration of HTAN investigation with BigQuery
- Live demo on notebook with Celltypist in MyBinder

The ISB-CGC homepage isb-cgc.org



NCI's Cancer Research Data Commons (CRDC)



DATA GENERATED FROM BASIC, CLINICAL, AND POPULATION RESEARCH DATA SUBMITTED, HARMONIZED, STORED, AND MADE PUBLICLY ACCESSIBLE NOVEL TOOLS AND APPLICATIONS FOR USE IN COLLABORATIVE RESEARCH DATA-DRIVEN CANCER RESEARCH FOR BETTER DETECTION, TREATMENT, AND CARE

https://datacommons.cancer.gov/

OUTCOMES

Multiple Data Commons host and control access to different types of cancer data





ISB-CGC's approach to enabling data science in the cloud

- Moving Excel files into the cloud
- Derived molecular data available for query as you need, updated frequently
- Tooling examples provided to enable data mining and Machine Learning of your data
- Sharing of results with those you choose
- Maximum flexibility of scripting and compute for those who desire it



Data wrangling can be onerous, for example GDC has 24,944 individual transcriptome files for just TCGA

Filters								
+ Add a Custom Filter				👌 Manifest View Images	×	Add All Files to	Cart E Remov	e All From Cart
Experimental Strategy	C D P	JSON	TSV	Total of 24,944 Files ± 20,925 Cases	105.4	TGB Q Se	sarch	
Name +	Files 🗘							
SIRNA-Seq SIRNA-Seq	24,944 (2.53%) 74 (0.01%)	Cart	Access 🗘	File Name 🌐	Cases	Project ‡	Data Category 🗘	Data Format 🗘
	O show less	×	Open	g d0ee5ff7-a49a-4633-93a6-40c9e29fb0b7.ma_seq.augmented_star_gene_counts.tsv	81	S TCGA-BRCA	Transcriptome Profiling	TSV
Wgs Coverage	C [] P	×	Open	S c58a5583-7004-4b67-9372-e161e18d7de1.rna_seq.augmented_star_gene_counts.tsy	81	R TCGA-BRCA	Transcriptome Profiling	TSV
Name No data for this field	Files 🛟		Open	R 269c35f0-a4f7-4e30-a69f-f1f3b7b5dace.ma_seq.augmented_star_gene_counts.tsv	81	R TCGA-BRCA	Transcriptome Profiling	TSV
Data Category	C 11 2	×	Open	😡 158ab1d9-8925-4a05-95da-b2e0ca297474.ma, seq.augmented_star_gene_counts.tsy	81	S TCGA-BRCA	Transcriptome Profiling	TSV
Name +	Files 1		Open	S 9c2ed2bb-8ee1-441e-9f3b-ffbb4def2673.ma_seq.augmented_star_gene_counts.tsv	91	R TCGA-BRCA	Transcriptome Profiling	TSV
Uanscriptome profiling	24,944 (2,53%)	×	Open	5 11ed8e05-8530-460a-8502-01ae09504315.ma_seq.augmented_star_gene_counts.tsv	81	R TCGA-BRCA	Transcriptome Profiling	TSV
Data Type	C D P		Open	G 0d6/befe-4e21-4762-a476-9c899d4a94b7.rna_seq.augmented_star_gene_counts.tsv	81	R TCGA-BRCA	Transcriptome Profiling	TSV
Aligned Reads Gene Expression Quantification Splice junction Quantification Transcript Fusion	73,550 (7.46%) 24,944 (2.53%) 24,944 (2.53%)		Open	S be813beb-9b35-4063-9d61-9f49a7fd7706 ma_seq.augmented_star_gene_counts.tsv	81	9 TCGA-BRCA	Transcriptome Profiling	TSV
	93,175 (9.45%)		Open	S 01661d94.fc16-4456-95cFa5fa4e1e196c.ma_seq.augmented_star_gene_counts.tsv	<mark>9</mark> 1	TCGA-BRCA	Transcriptome Profiling	TSV
	Show less	-					Transminterne	



The Google Cloud offers tools to simply host derived data by concatenating these files into a single BQ table





BigQuery is a powerful statistical tool that can run hundreds of millions of tests in seconds

Testing BigQuery compute time with statistical tests

- Millions of tests in 40 seconds
- 6.6 billion correlations for \$1.16





Where does HTAN data come from?

https://humantumoratlas.org/research-network https://humantumoratlas.org/standards



Data in HTAN is contributed by centers on our Atlas Team.

Collection of data is often complex, comprising of many different cancers (or precancer) types, time points, and assays.

In HTAN we collect 3 key parts of the data from this process:

<u>Clinical Data</u> (i.e. Patient Information), <u>Metadata</u> (i.e. How was the data created), <u>All final Raw and Processed data</u> files.

DATA LEVELS IN HTAN EXPLAINED

Example: Sequencing Data



Navigating the Google Cloud Console and Mybinder

https://console.cloud.google.com

https://mybinder.org

ISB-CGC team



Bill Longabaugh Dar'ya Pozhidayeva Suzanne Paquette Elaine Lee Lauren Hagen Boris Aguilar Lauren Wolfe David Pot Danna Huffman Fabian Seidl Jacob Wilson Poojitha Gundluru Deena Bleich

GENERAL DYNAMICS

Information Technology

