Analyzing HTAN scRNASeq data accessible in BigQuery with CellTypist

Fabian Seidl and Dar'ya Pozhidayeva for ISB-CGC

2024-04-02

ISB-CGC

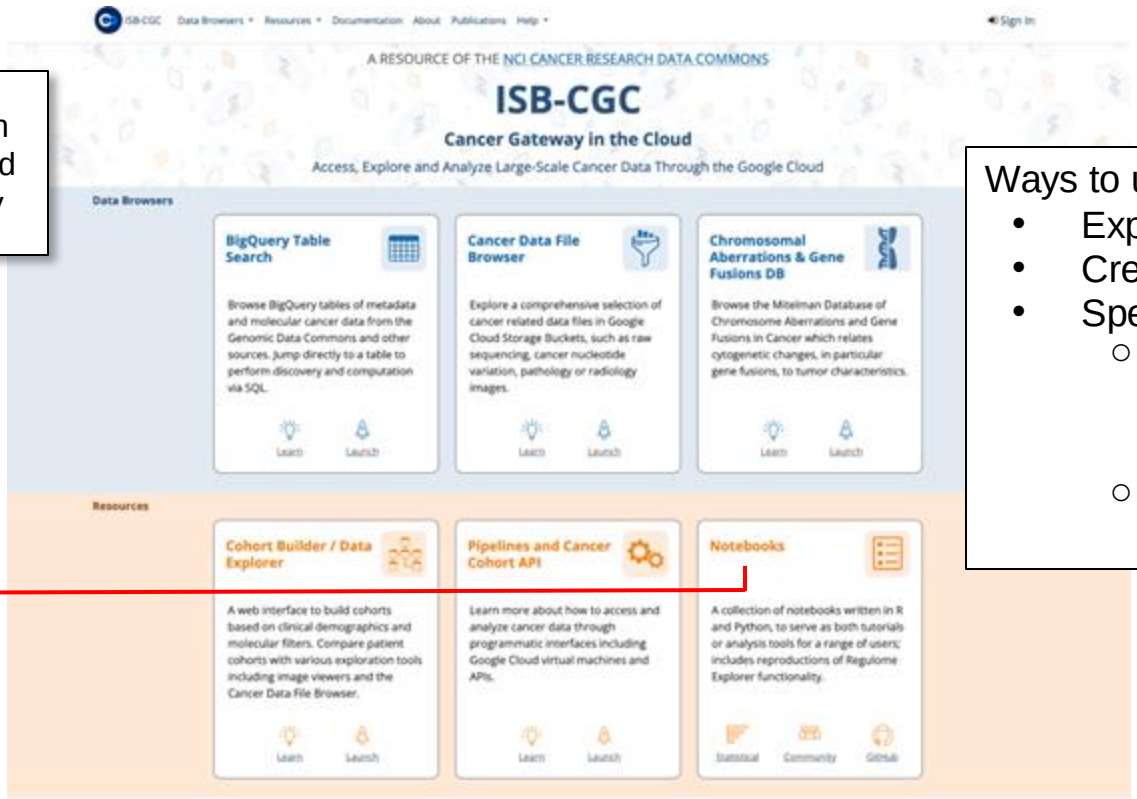
Outline of today’s agenda

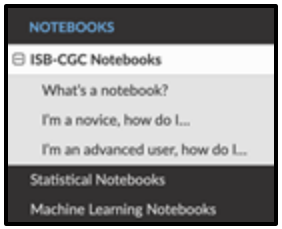
- Short intro to CRDC, ISB-CGC and the Data Commons
- Data exploration in BigQuery (‘Excel-like data tables in the cloud’)
- Demonstration of HTAN investigation with BigQuery
- Live demo on notebook with Celltypist in MyBinder

The ISB-CGC homepage

isb-cgc.org

Focus today is on analysis of derived data in BigQuery

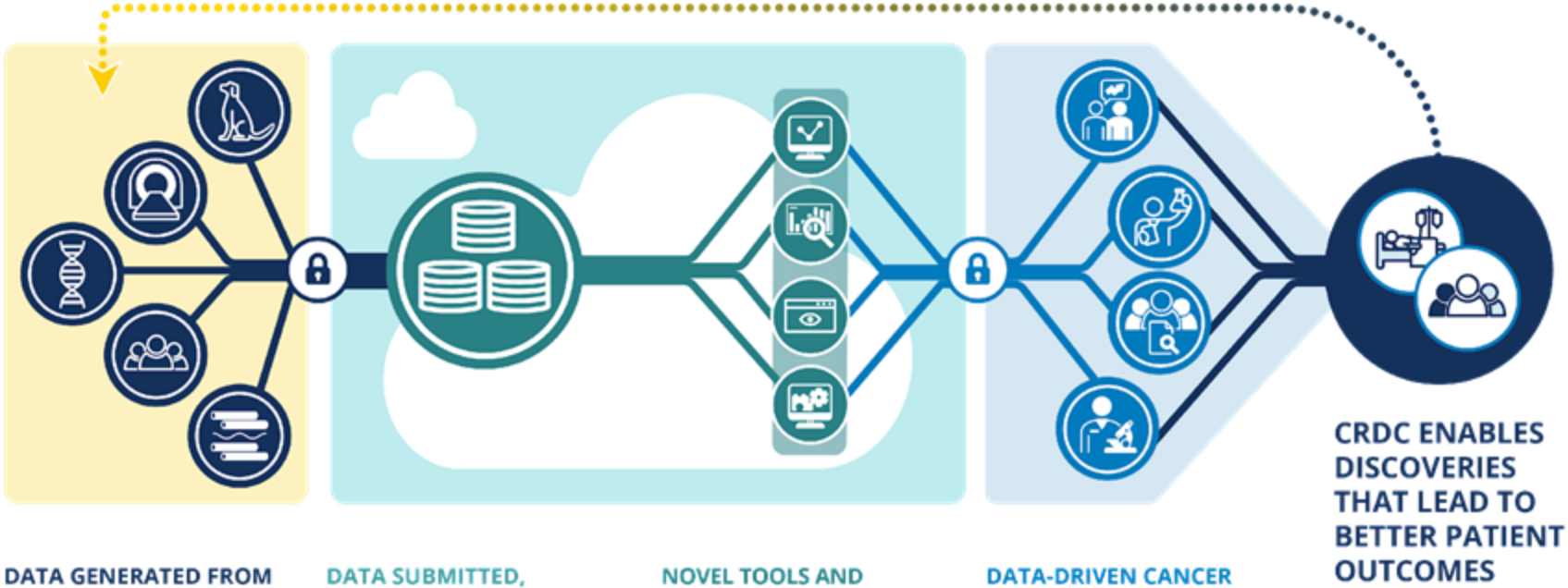




Ways to use data and tools:

- Explore datasets
- Create cohorts
- Specialized DBs
  - Mitelman DB of Chromosomal Aberrations & Gene Fusions
  - caNanoLab

NCI's Cancer Research Data Commons (CRDC)



https://datacommons.cancer.gov/

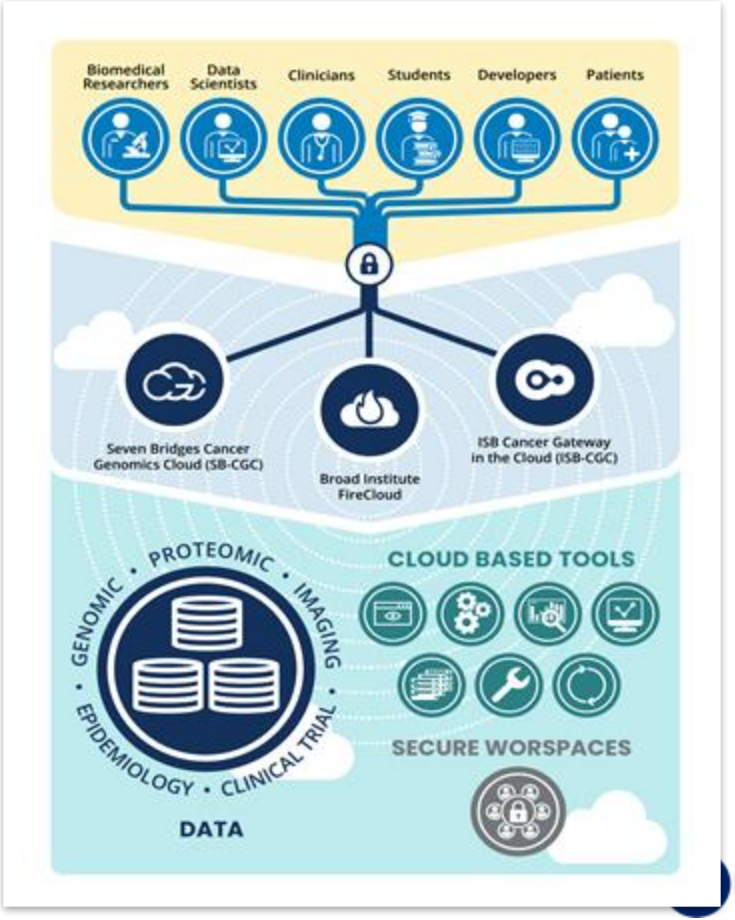
NIH NATIONAL CANCER INSTITUTE

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

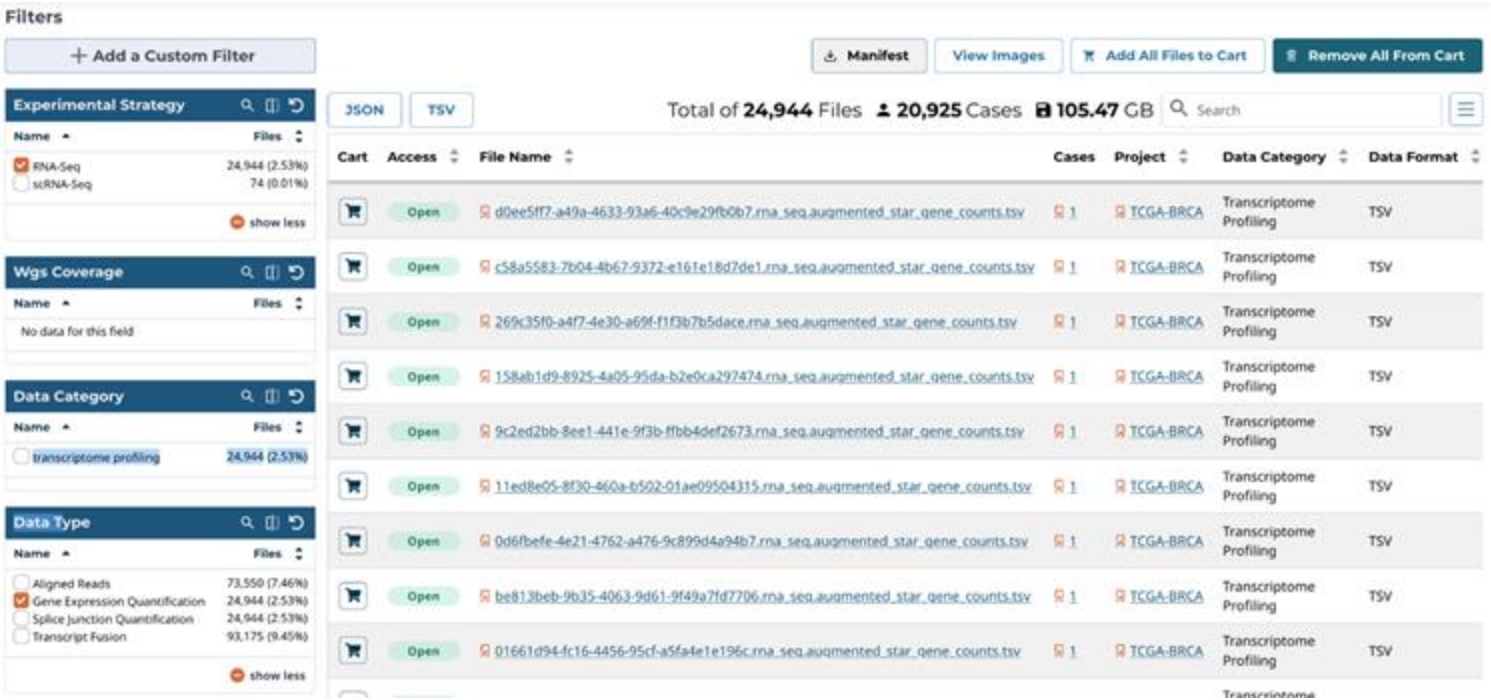
ISB-CGC

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

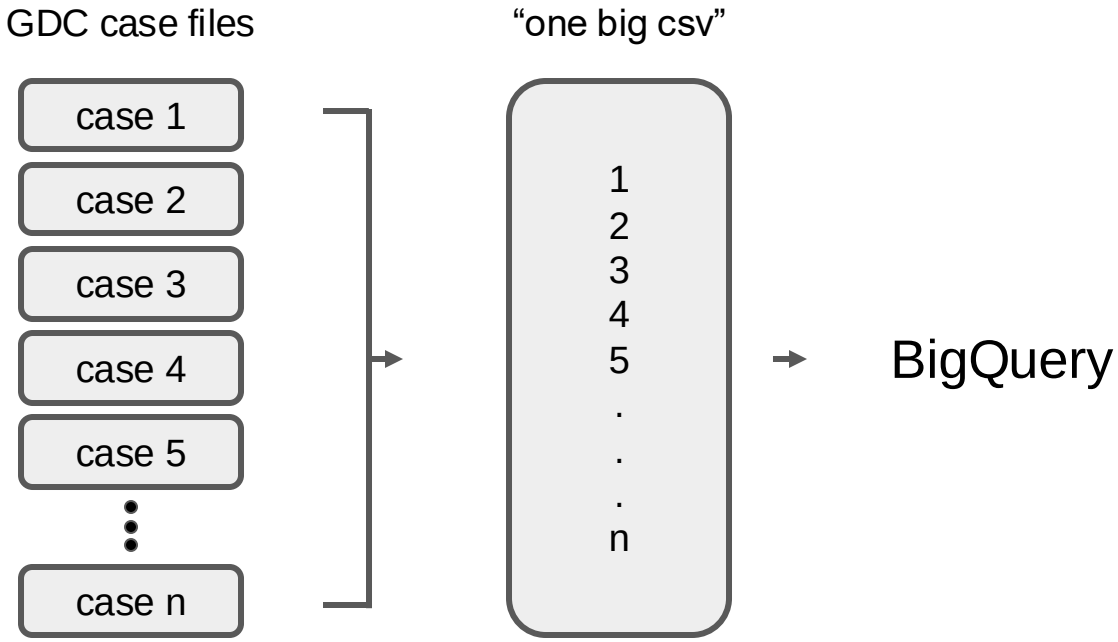


Data wrangling can be onerous, for example GDC has 24,944 individual transcriptome files for just TCGA



ISB-CGC

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

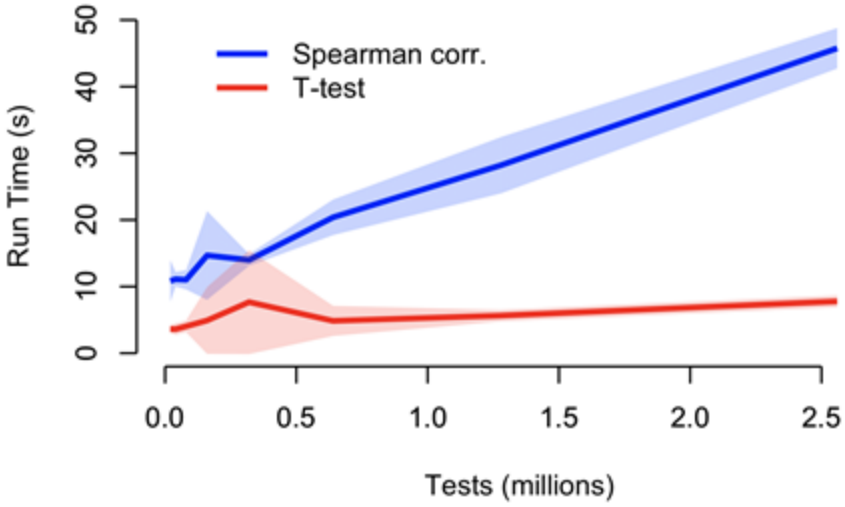


ISB-CGC

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

Testing BigQuery compute time with statistical tests

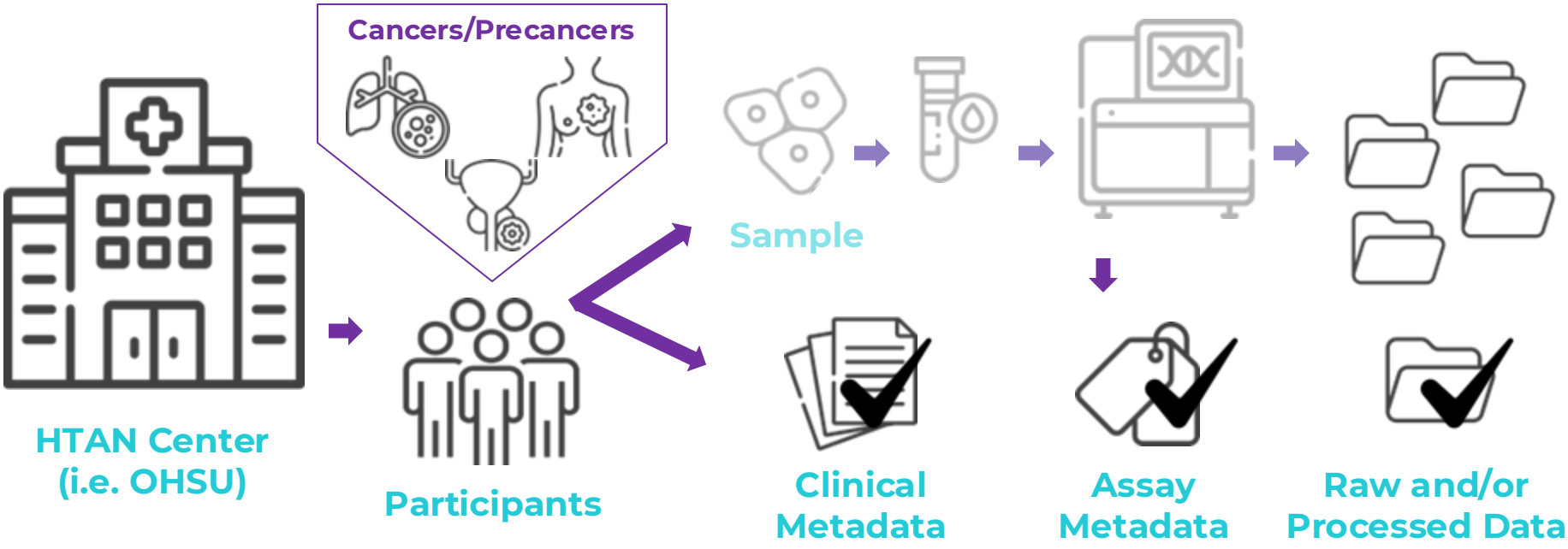
- Millions of tests in 40 seconds
- 6.6 billion correlations for $1.16



ISB-CGC

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

DATA LEVELS IN HTAN EXPLAINED

Example: Sequencing Data

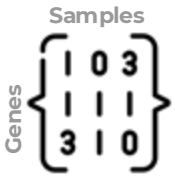
Raw Data

FASTQs
(Paired or Single)

Aligned Data

BAMs
(Aligned)

Derived Biomolecular Data



Expression Matrix
(Raw and/or filtered)

Sample Level Summary Data

Cell
Annotations

HTA1_203_3321026 .fastq.gz
HTA1_203_3321025 .fastq.gz

HTA1_203_332102501 .bam

HTA1_203_3321025013 matrix.mtx.gz
HTA1_203_3321025011 barcodes.tsv.gz
HTA1_203_3321025012 features.tsv.gz

HTA1_203_33210250111 .csv

RAW DATA → **PROCESSED DATA**

LEVEL 1
Data File ID(s)

→ Parent Data File ID(s) For L1

LEVEL 2
Data File ID(s)

→ Parent Data File ID(s) For L2

LEVEL 3

→

LEVEL 4 ...

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

ISB-CGC