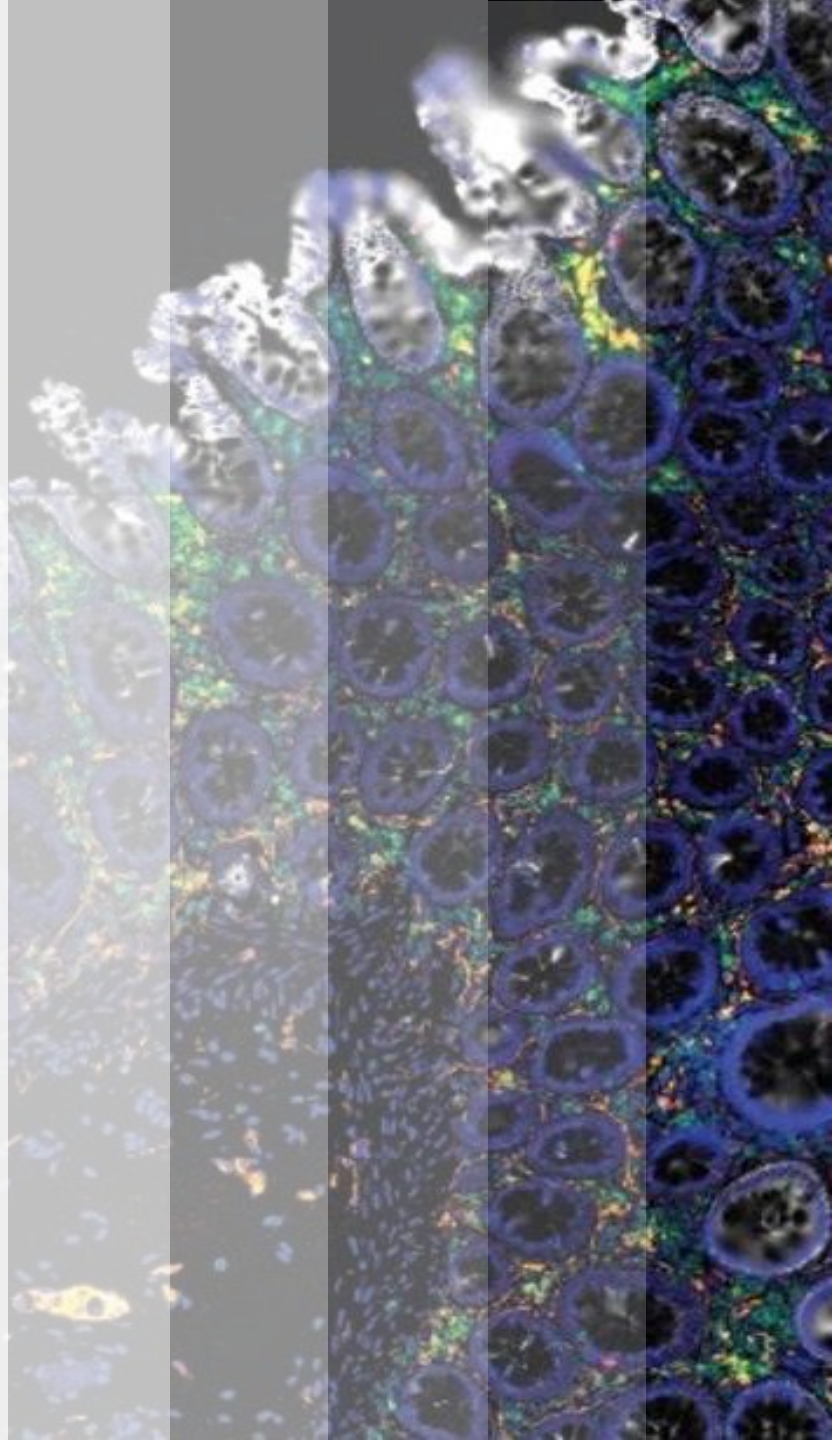
VELSERA.

HTAN via The CRDC's Seven Bridges Cancer Genomics Cloud

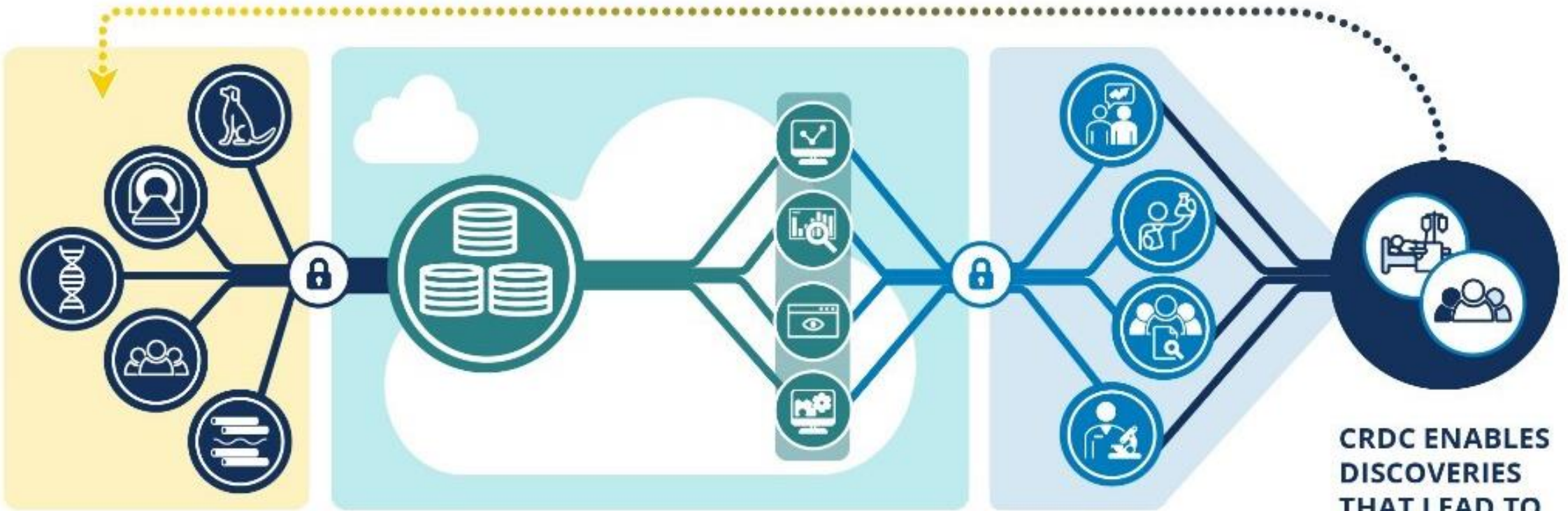
Rowan Beck
Seven Bridges SB-CGC
Bioinformatics Trainers and
Education Program



Agenda

- **Overview of the Seven Bridges Cancer Genomics Cloud (SB-CGC), powered by Velsera**
- **How to use SB-CGC for HTAN Integration**

Cancer Research Data Commons



CANCER GENOMICS CLOUD
SEVEN BRIDGES

3+	1600+	1000+	8000+	80000+
Petabytes Public Data	Years of Compute	Public Tools & Workflows	Users	User-Created Workflows

Provides powerful, yet easy to use interfaces to empower cancer researchers to draw new insights from petabyte scale data.

Stable, secure, and highly customizable cloud storage and computing platform.

Who are the SB-CGC Users?

The SB-CGC is designed to serve a wide range of scientists and users with varying skill sets

ADMINISTRATORS

- Manage and Control Users
- Monitor and Control Institutional Assets
- Manage and Monitor Projects
- Monitor and Control Costs
- Create Reports

BIOINFORMATICIANS

- Store, Manage, and Share Data
- Access Public and Proprietary Datasets
- Query, Build, and Investigate Cohorts of Interest
- Access Optimized Tools and Workflows
- Create, Optimize, Maintain, and Distribute New Tools and Workflows
- Create Push-button Automation Solutions
- Analyze Data at Scale with Tools and Workflows
- Conduct Interactive Exploratory Analyses
- Explore/Visualize Results and Gather Insights
- Easily Collaborate with Other Stakeholders
- Integrate with External Systems

BENCH SCIENTISTS

- Store, Manage, and Share Data
- Run Optimized Tools/ Workflows at Scale
- Conduct Defined Analyses via Push-button Solutions
- Investigate/Visualize Results
- Easily Collaborate with Other Stakeholders

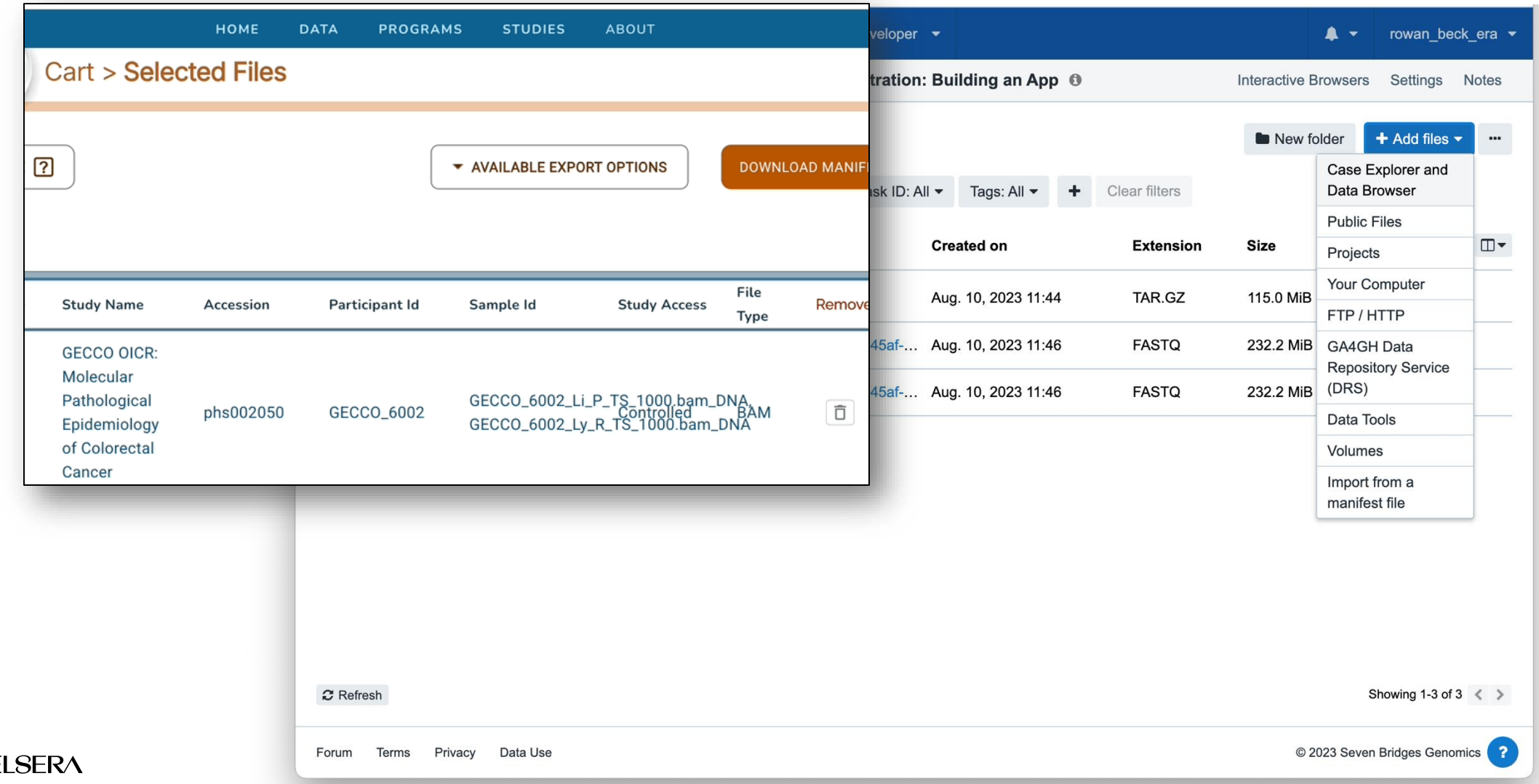
CLINICIANS

- Conduct Validated Analyses via Push-button Solutions
- Query, Build, and Investigate Cohorts of Interest
- Create Reports
- Investigate/Visualize Results
- Easily Collaborate with Other Stakeholders

DEVELOPERS

- Create, Optimize, and Maintain New Tools and Workflows
- Create Push-button Automation Solutions
- Create Custom Interfaces for Specific Use Cases
- Distribute Proprietary Tools/ Workflows
- Integrate with Upstream/ Downstream Systems

Access Public Data, or Use Your Own



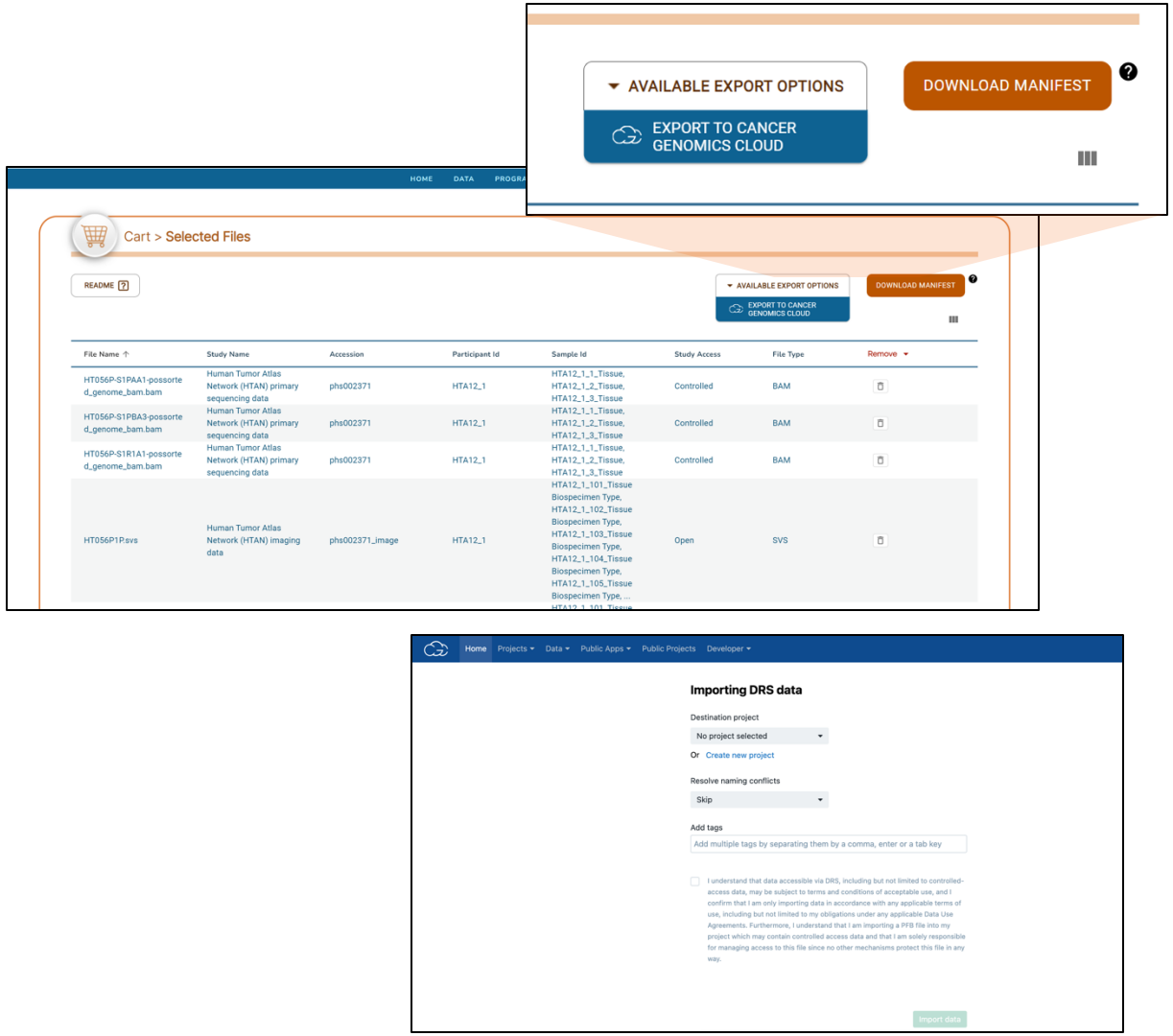
Access HTAN Data Directly on the SB-CGC

Level 1 & 2 on the SB-CGC

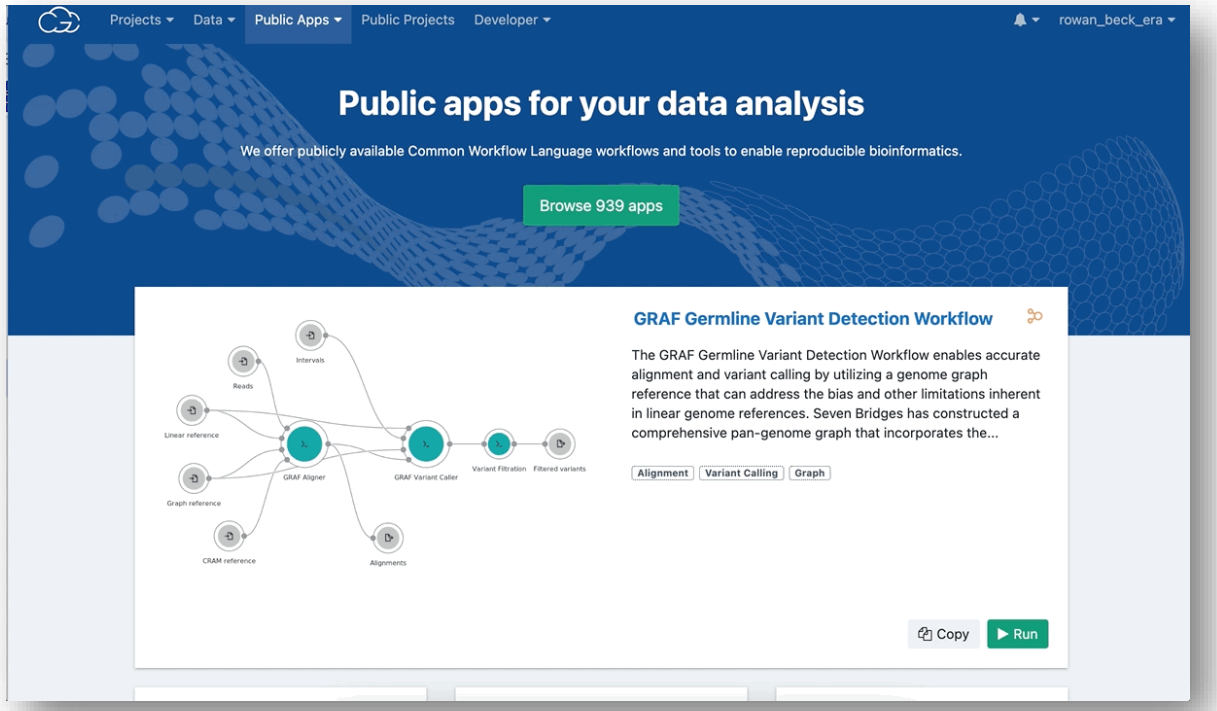
- Access Controlled Sequencing data and Open Access Imaging data (CC BY 4.0).

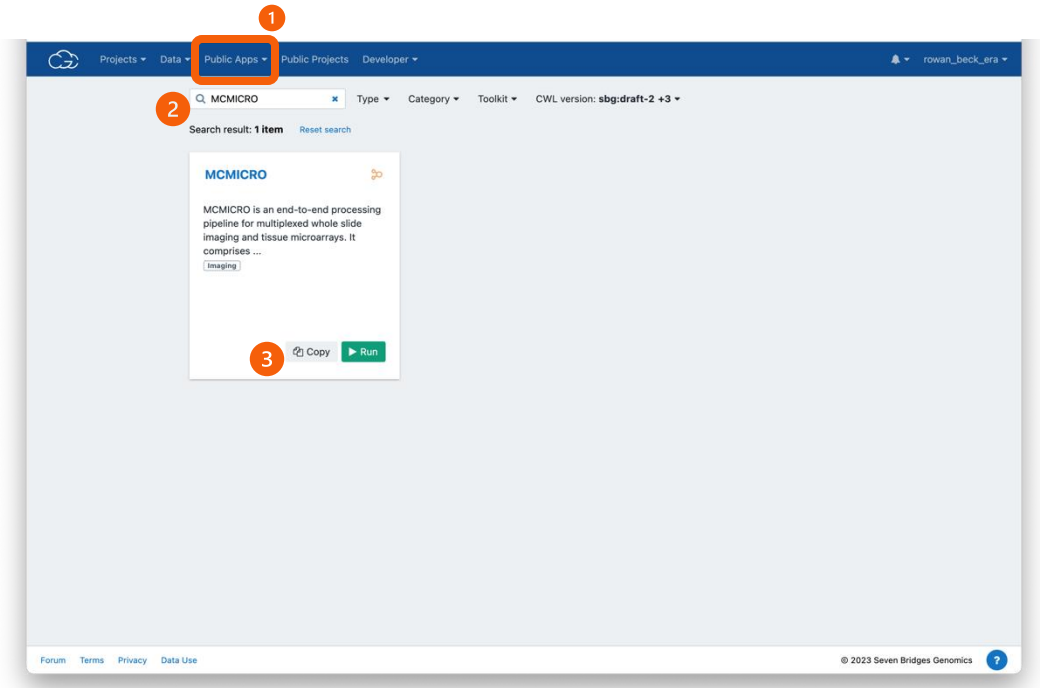
Synapse - Open Access Processed level 3 and level 4 data

Imaging Data Commons (IDC) - Open Access Imaging data (CC BY 4.0) in DICOM-TIFF format



Browse Hundreds of Tools and Workflows





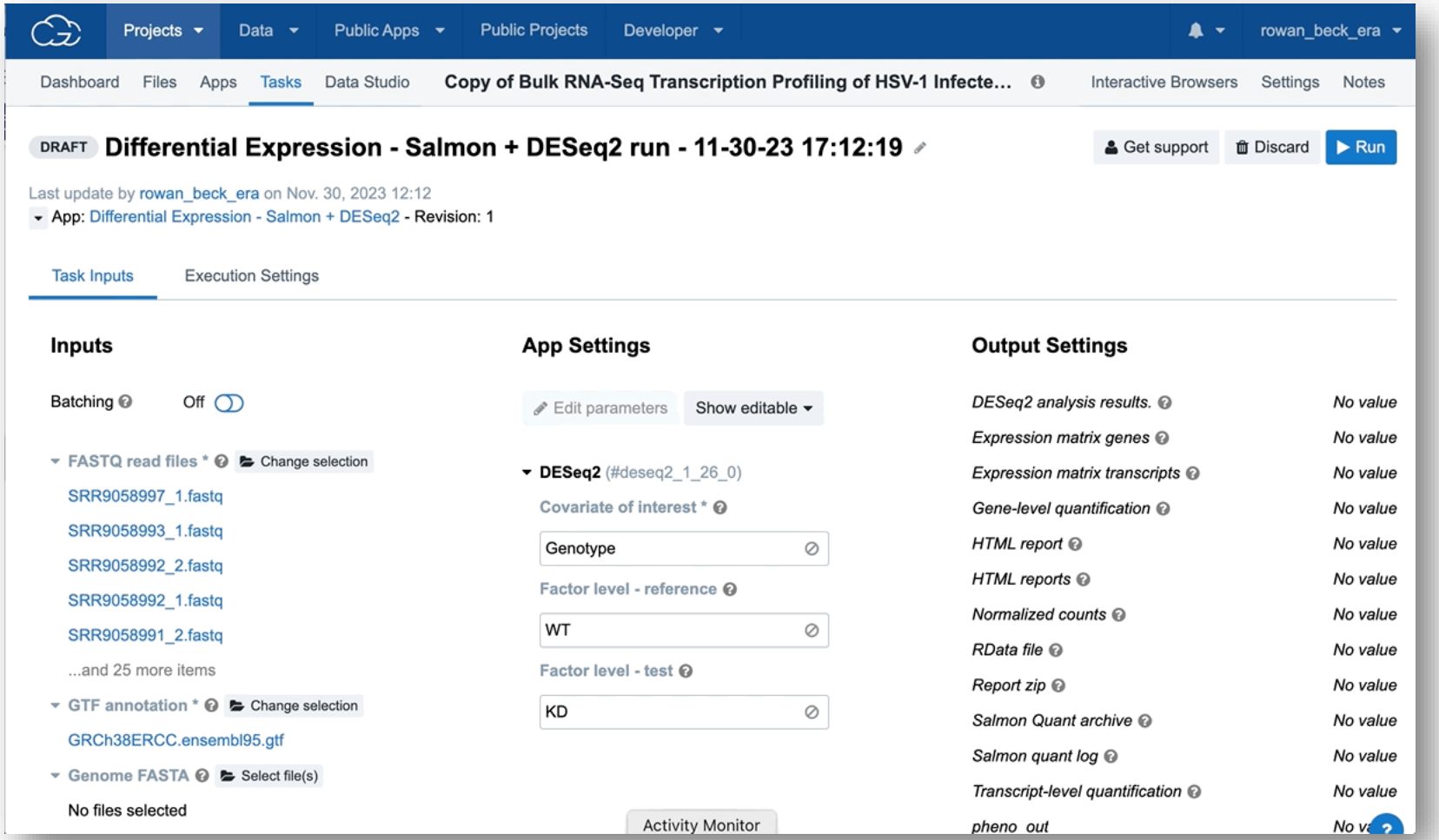
MCMICRO

Enables the processing of multiplexed tissue images.

Transform whole-slide images into single-cell data using this simple workflow.

VELSERA

No Coding Required to Run an Analysis



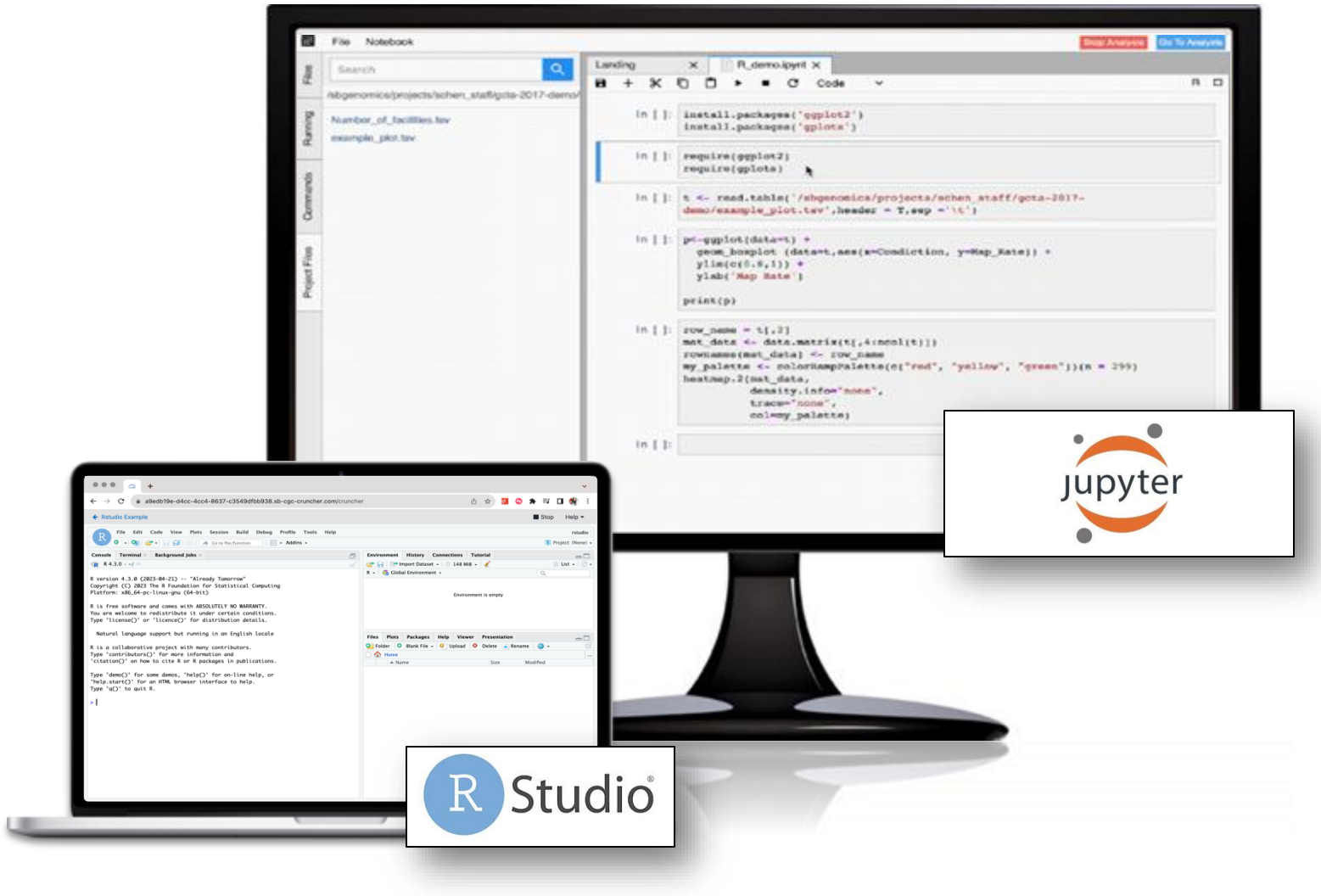
Integrated Custom Tertiary Analysis Tools

Data Science Workbench

Derive new insights using interactive analysis environments with **JupyterLab, SAS, Galaxy**, and **RStudio** environments.

Code in Python and create **Jupyter Notebooks** to record and share your analyses.

Similarly, turn your analyses into high quality documents, reports, presentations and dashboards using **R Markdown**.



Estimate Cloud Costs

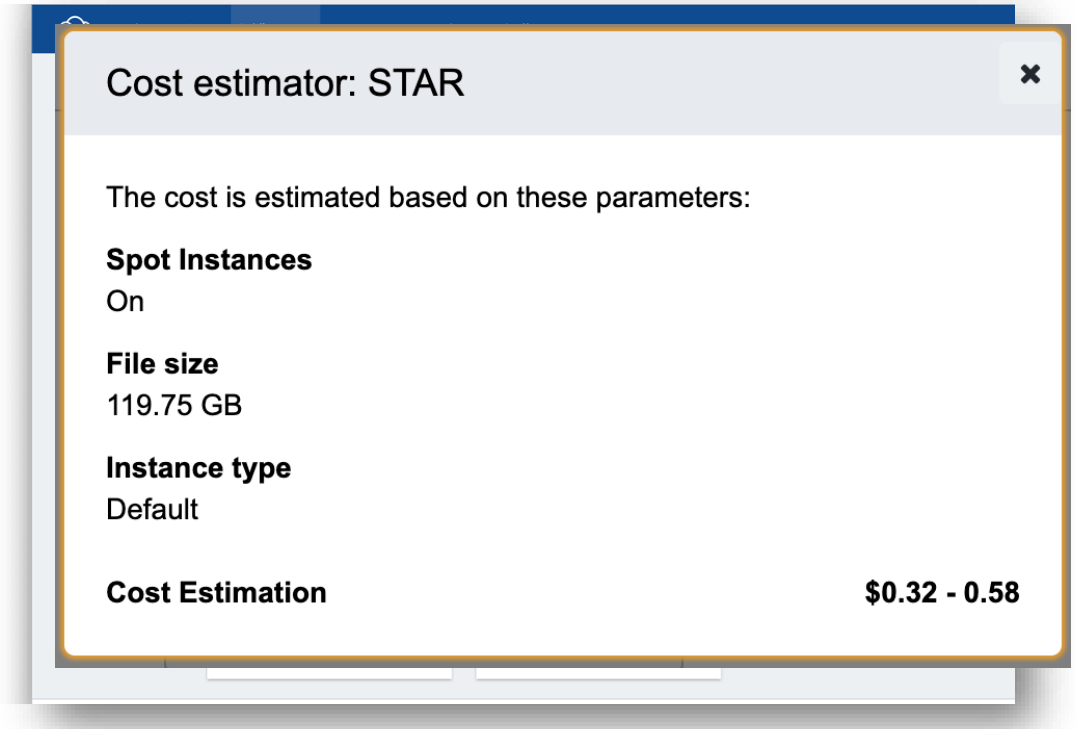
Performance Benchmarking

- Runtimes
- Task Costs
- Various file sizes

Experiment type	Input size	Paired-end	# of reads	Read length	Duration	Cost	Instance (AWS)
RNA-Seq	2 x 230 MB	Yes	1M	101	18min	$0.40	c4.8xlarge
RNA-Seq	2 x 4.5 GB	Yes	20M	101	30min	$0.60	c4.8xlarge
RNA-Seq	2 x 17.4 GB	Yes	76M	101	64min	$1.20	c4.8xlarge

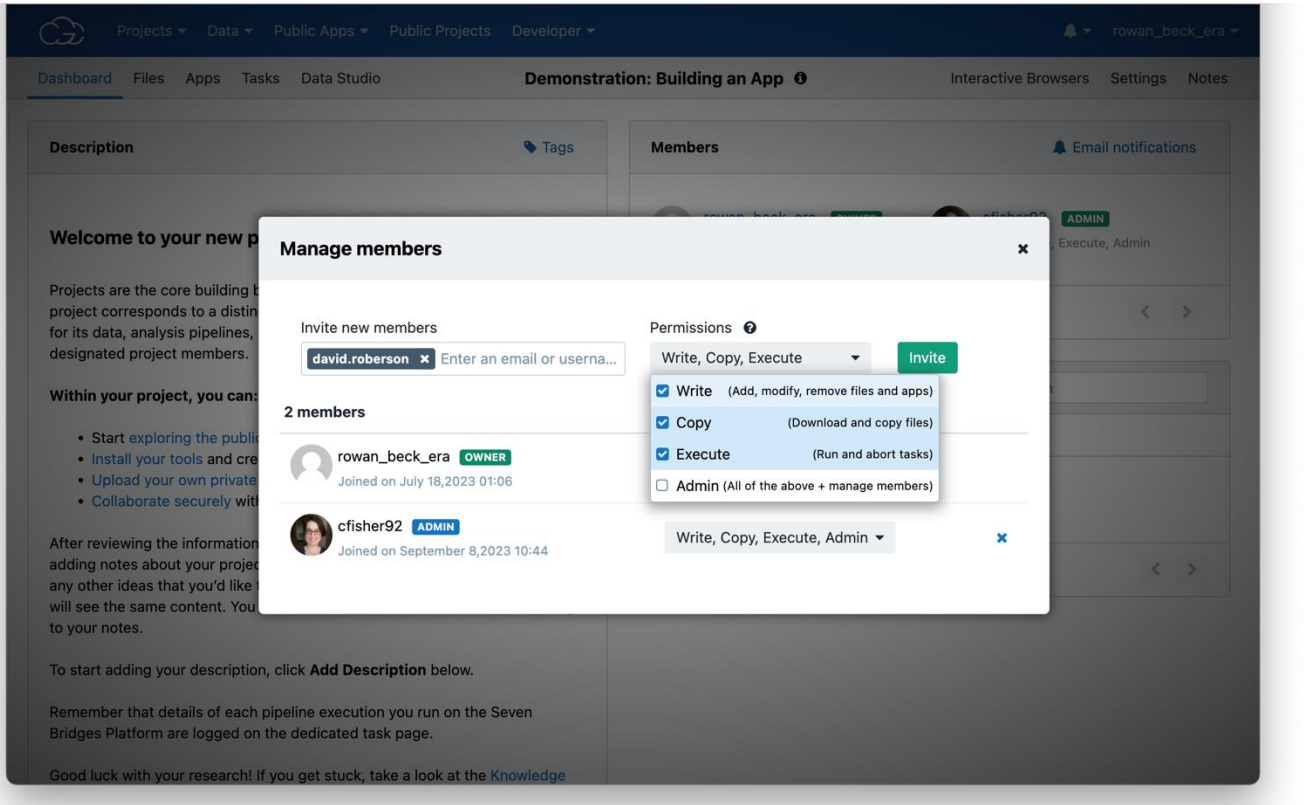
Cloud Cost Estimator

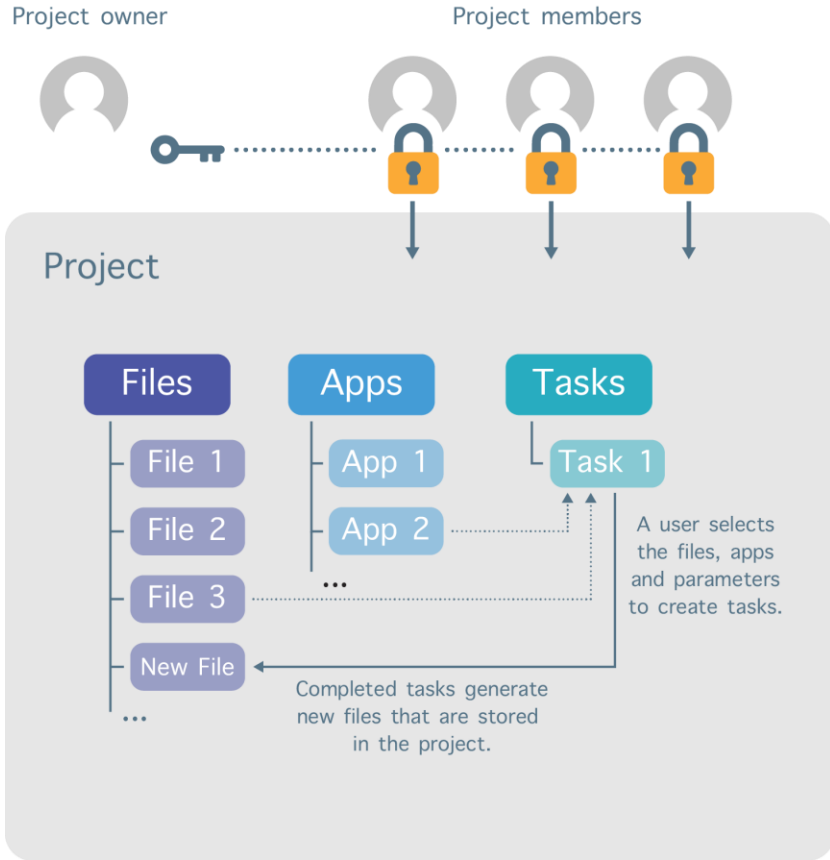
- Available for a limited number of apps
- Create an estimate for the cost of your specific use case
- Compare workflows to save on computing costs



VELSERA

Collaborating Has Never Been Easier



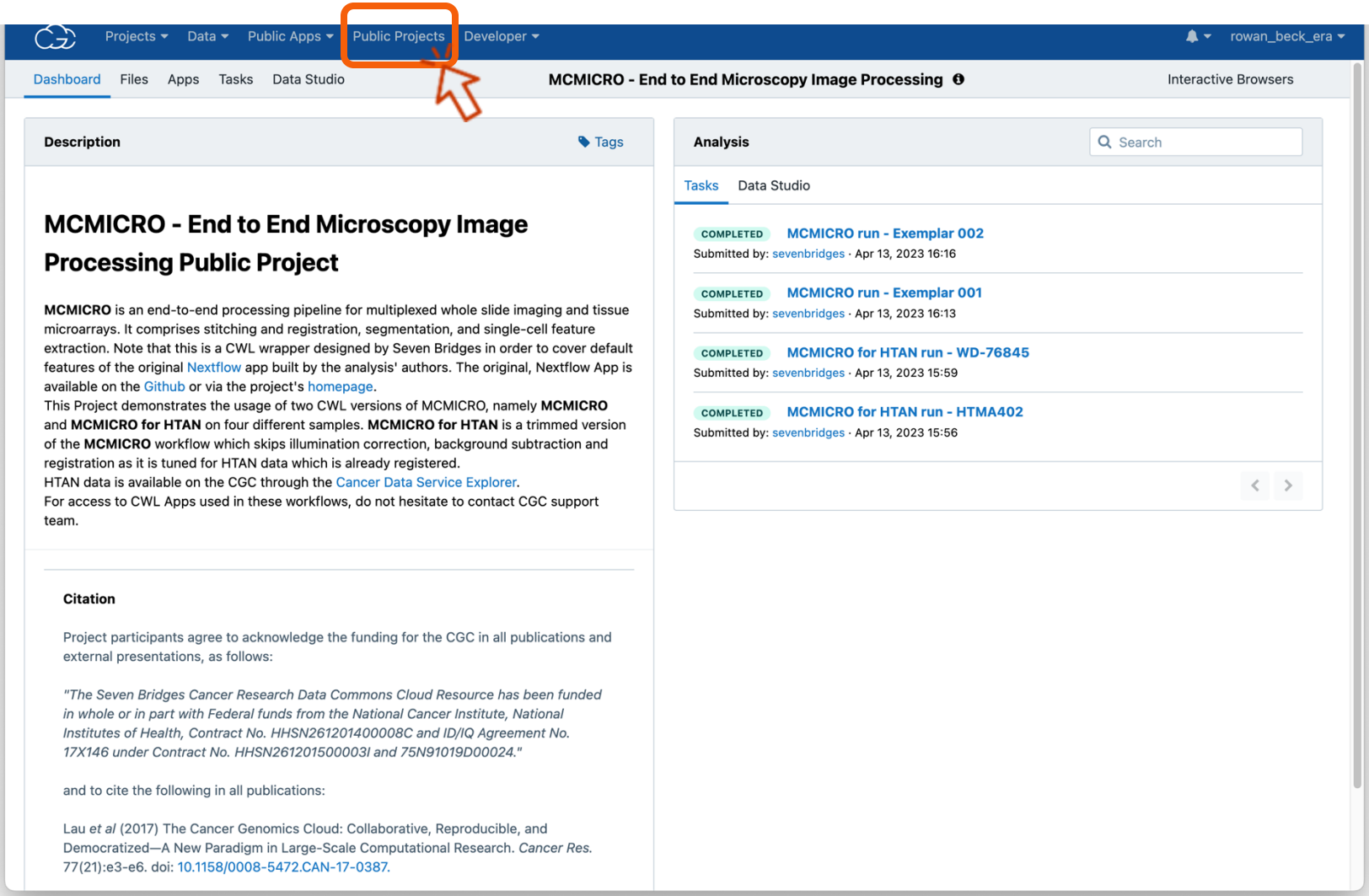


Collaborating Has Never Been Easier

Pre-loaded with

- input and output files
- pre-run tasks
- recommended settings

Detailed descriptions on what the workflow does and how to use the pipeline



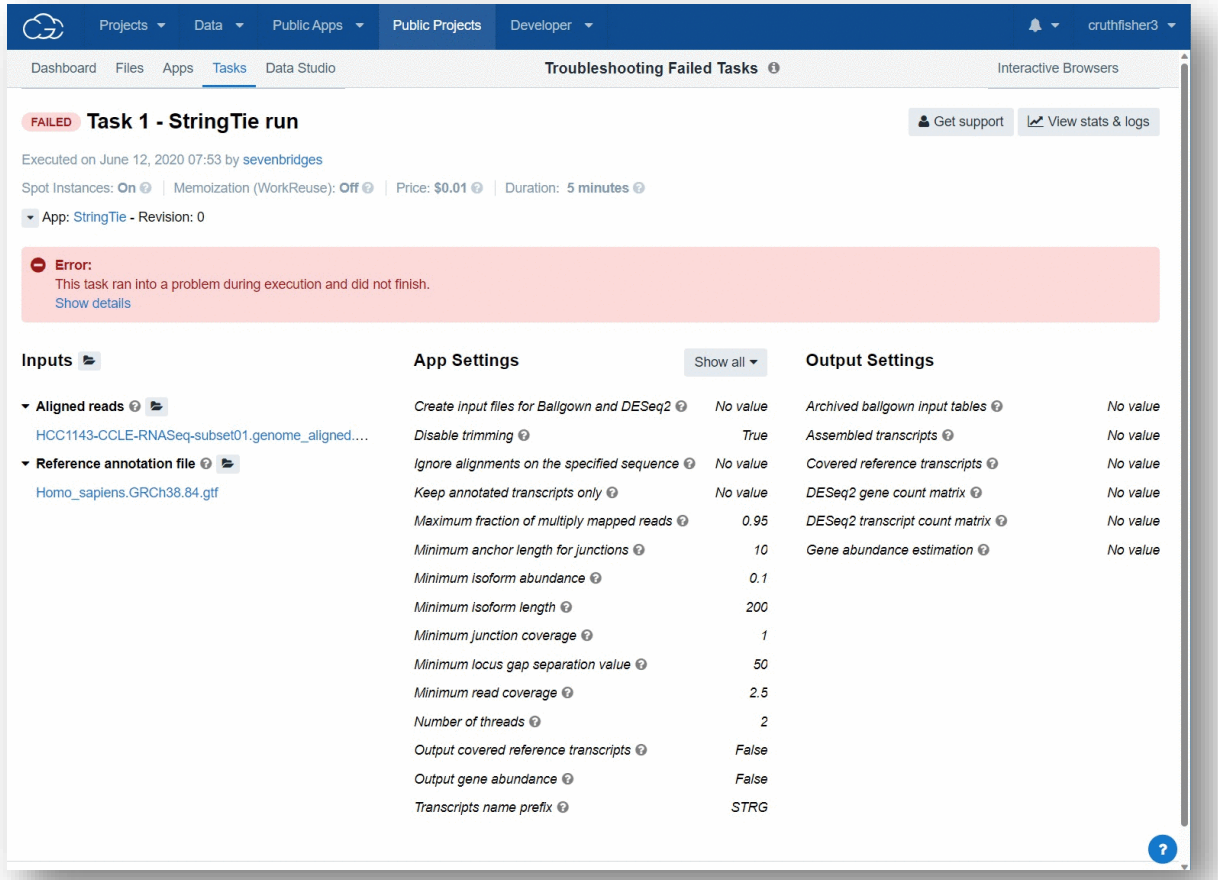
Collaborating Has Never Been Easier

Harvard Medical School	Artem Sokolov Clarence Yapp Jeremy Muhlich Yu-An Chen Clemens Hug Greg Baker Juha Ruokonen Edward Novikov Robert Krueger
Heidelberg University	Denis Schapiro Florian Wünnemann Miguel Ibarra Krešimir Beštak
Oregon Health and Sciences University	Allison Creason Jeremy Goecks Daniel Persson Qiang Gu Luke Sargent Cameron Watson Luke Strgar
Dana-Farber Cancer Institute	Ajit Johnson
Vanderbilt University	Darren Tyson
Brigham and Women's Hospital	Giorgio Gaglia
The Jackson Laboratory	Brian White
Broad Institute	Huan Wang
Brigham Young University	Matthew Hodgman
Indica Labs	Erik Burlingame
University of Macau	Yimin Zheng
Sage Bionetworks	Adam Taylor
Leidos Biomedical Research	Shirish (Sam) Pathak
Cancer Research Data Commons (CRDC)	Erin Beck

Get The Support You Need

Every Week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday



VELSERA

Funding Is Available on the SB-CGC

Diverse approaches and engagement strategies tailored to community needs

Pilot Credit Funds

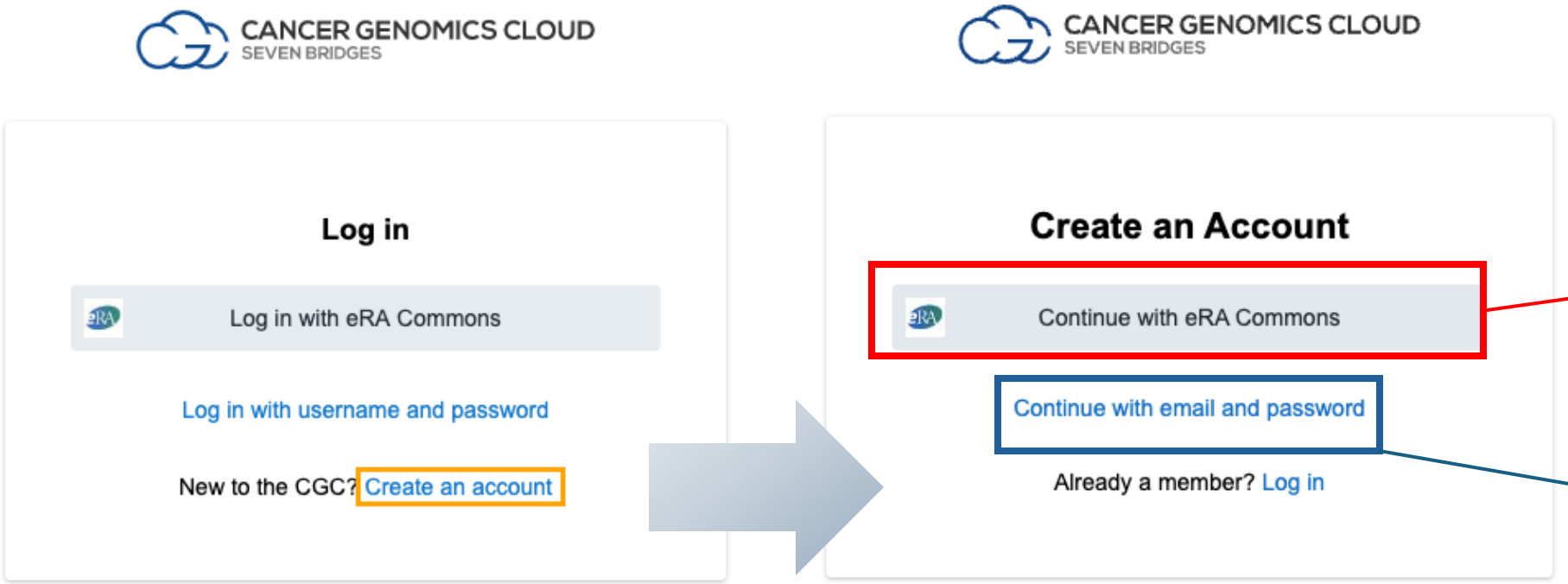
- **$300 of cloud credits**
- Free for new SB-CGC users
- Easy to request when signing up
- Fast approval

Collaborative Projects

- Cost estimation, optimization, and planning support
- Great for researchers new to bioinformatics and cloud approaches
- **Up to $10k compute/storage costs**
- Fast, rolling applications
- To date > 60 projects

How to use the SB-CGC for HTAN Integration

Register An Account



If you plan to access any **controlled** data, register your eRA Commons account

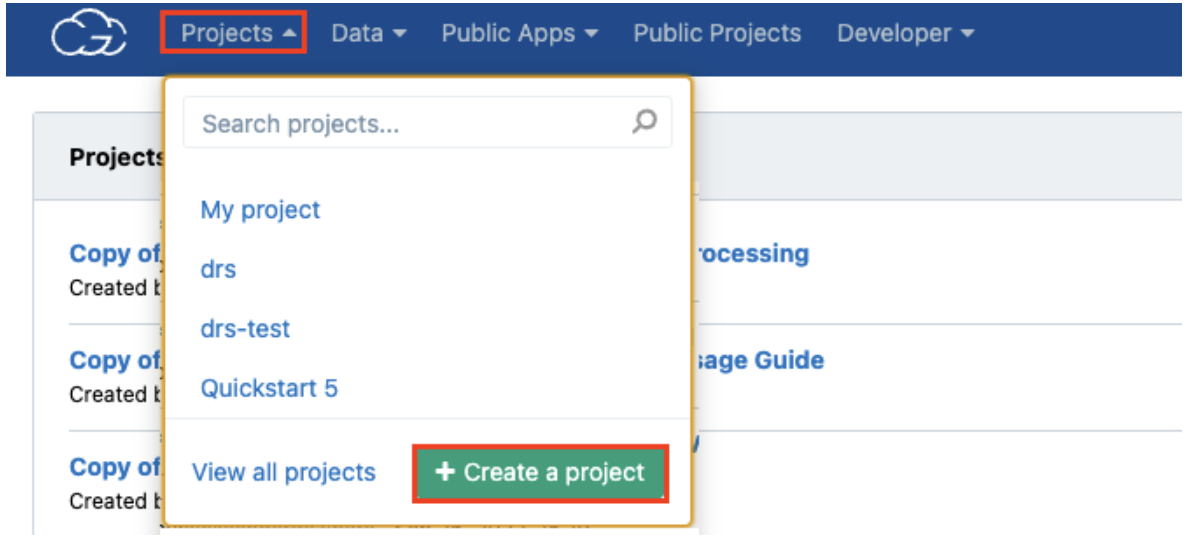
If you plan to work only with **open access** data, register with this option.

Register at:

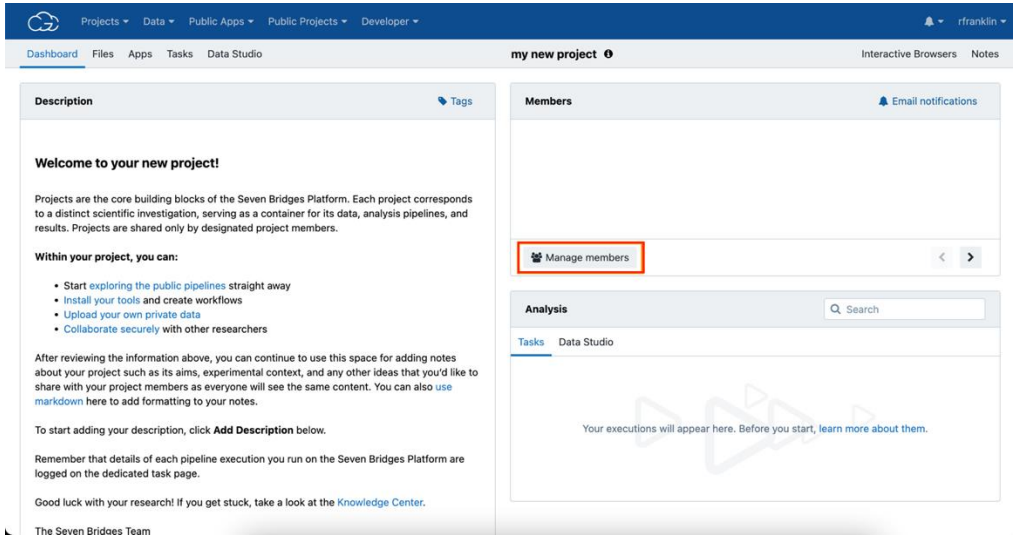
CGC.sbgenomics.com

Analysis Flow

1. Create a Project

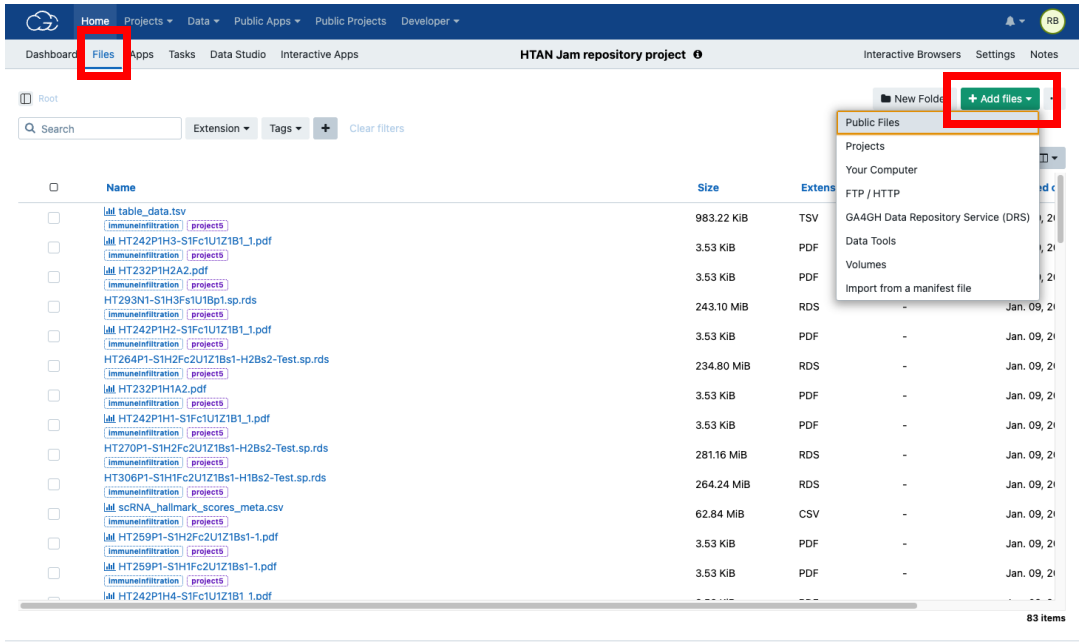


2. Add members

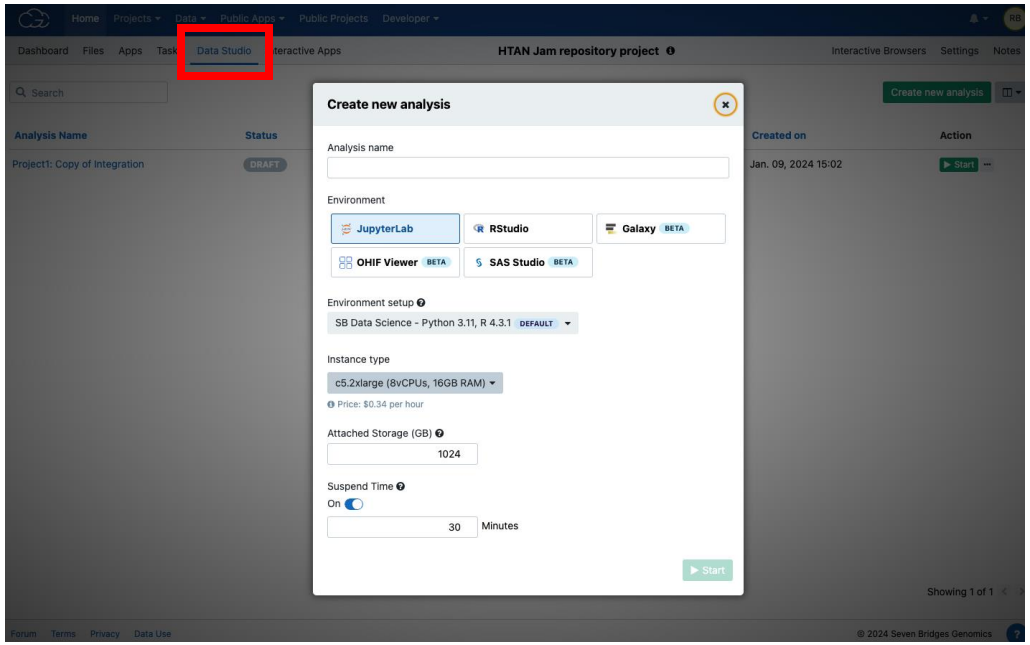


Analysis Flow

3. Add data (More information on next slide)



4. Conduct your analysis
(More information available: Helpful Information – Accessing Files in Data Studio)



Data Access

The HTAN Portal leverages several repositories to provide access to data:

- **Synapse** - Open Access Processed level 3 and level 4 data
- **Imaging Data Commons (IDC)** - Open Access Imaging data (CC BY 4.0) in DICOM-TIFF format
- **Seven Bridges Cancer Genomics Cloud** - Level 1 & 2 Access- Controlled Sequencing data and Open Access Imaging data (CC BY 4.0). Access control for the sequencing data is managed through dbGaP (Study Accession: phs002371).

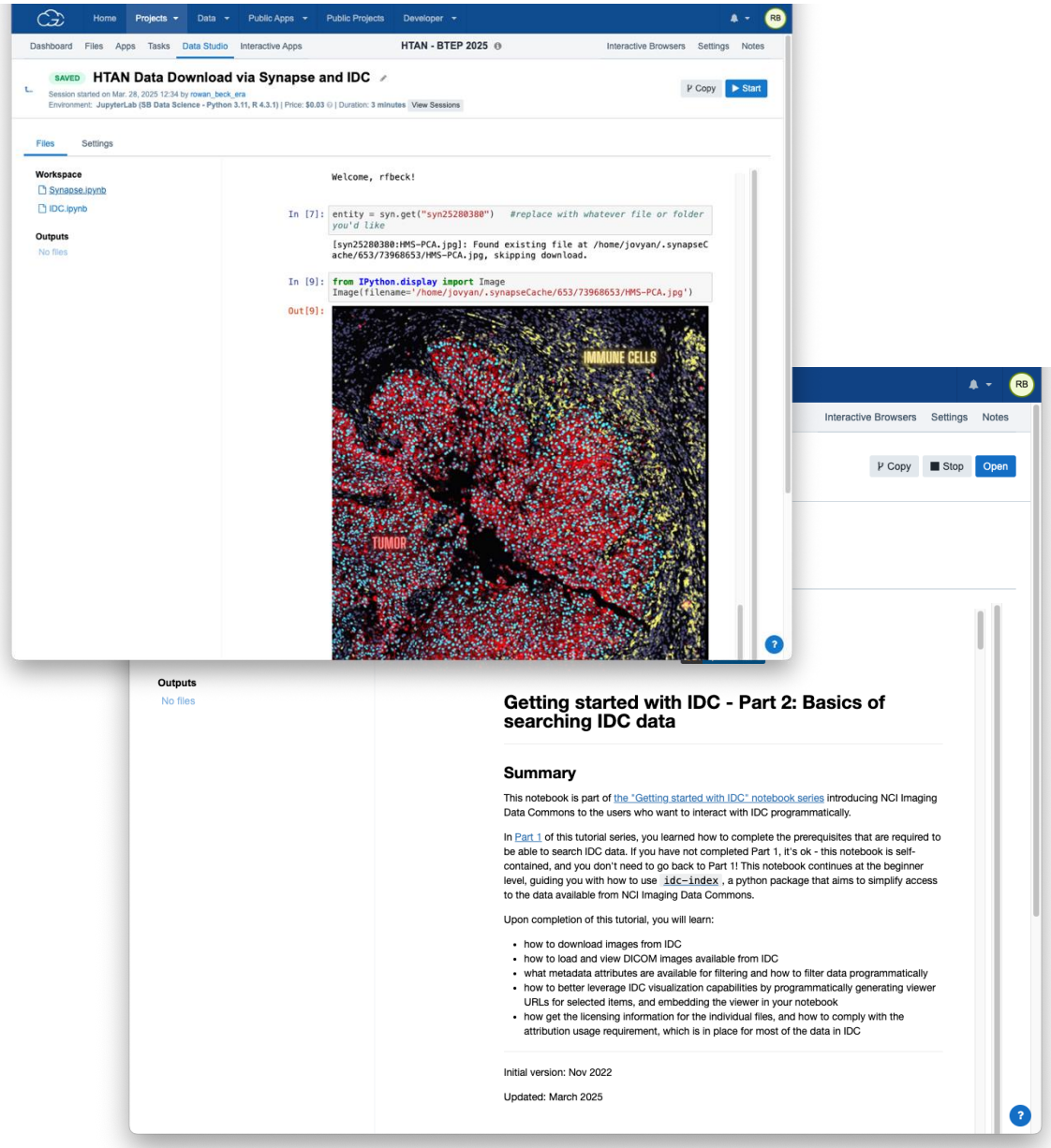
Instructions for accessing data from these repositories can be found directly from the Explore Page by making a selection of files and clicking on the download button. More information about accessing data can be found here.

Data Access

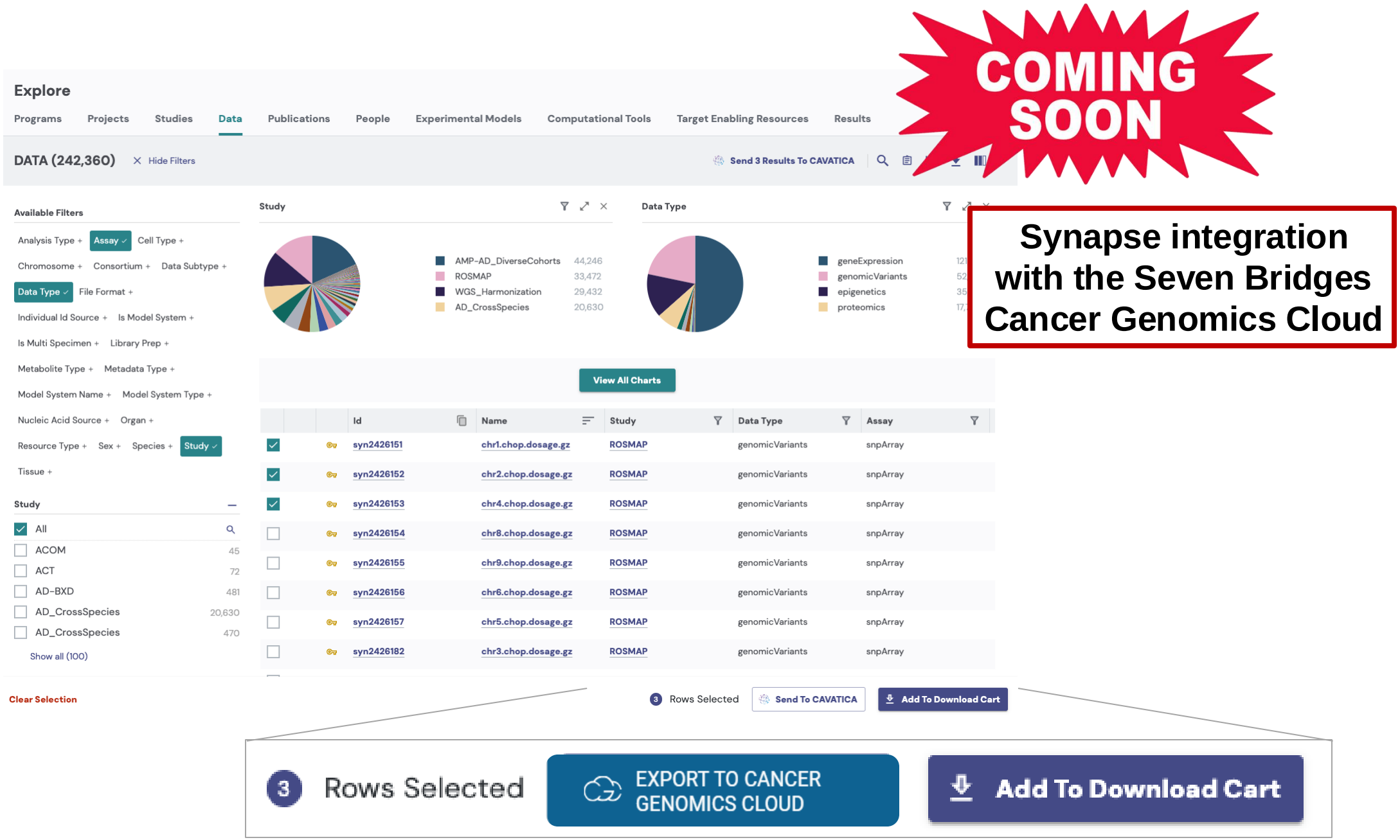
Synapse - Open Access
Processed level 3 and level 4 data

Imaging Data Commons (IDC) - Open Access
Imaging data (CC BY 4.0) in DICOM-TIFF format

Pre-built Jupyter notebooks to obtain data through Synapse or the Imaging Data Commons



https://learn.canceridc.dev/data/downloading-data
https://docs.humantumoratlas.org/open_access/synapse_to_cds/

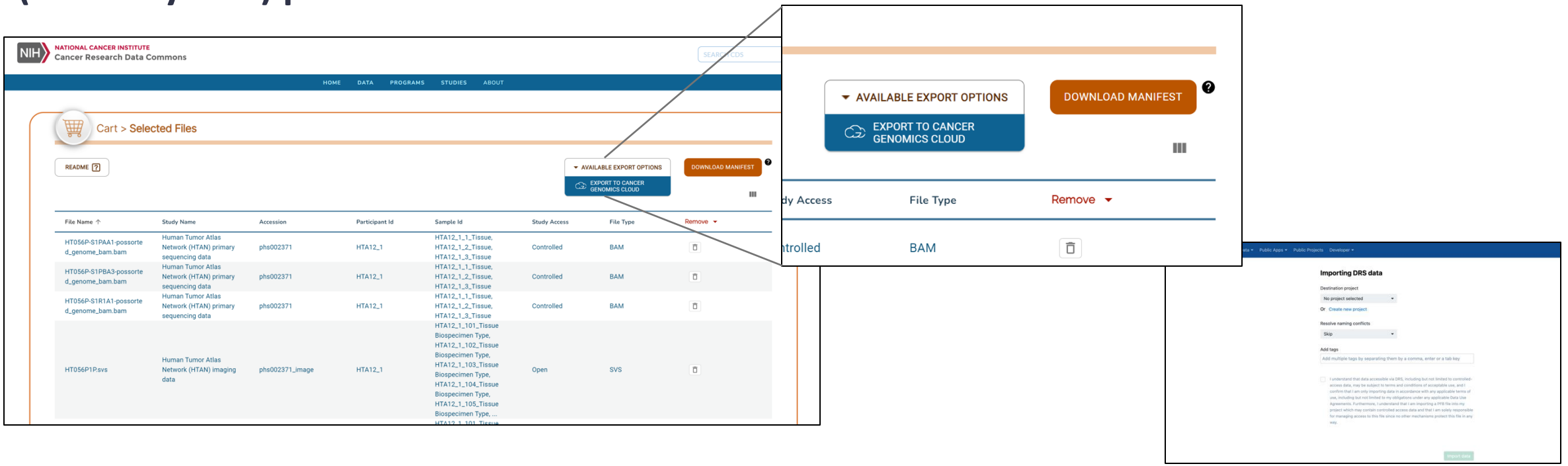


Synapse integration with the Seven Bridges Cancer Genomics Cloud

Data Access through SB-CGC

Seven Bridges Cancer Genomics Cloud - Level 1 & 2 Access- Controlled Sequencing data and Open Access Imaging data (CC BY 4.0). Access control for the sequencing data is managed through dbGaP (Study Accession: phs002371).

Direct export from the General Commons (formerly CDS) portal.



https://docs.humantumoratlas.org/open_access/cds_imaging/

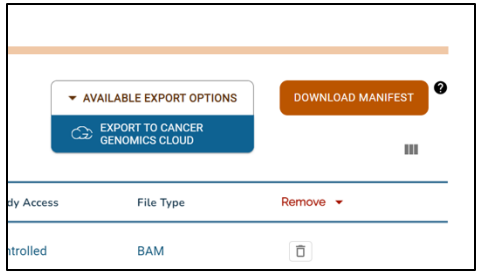
Data Access through SB-CGC

Seven Bridges Cancer Genomics Cloud - Level 1 & 2 Access- Controlled Sequencing data and Open Access Imaging data (CC BY 4.0). Access control for the sequencing data is managed through dbGaP (Study Accession: phs002371).

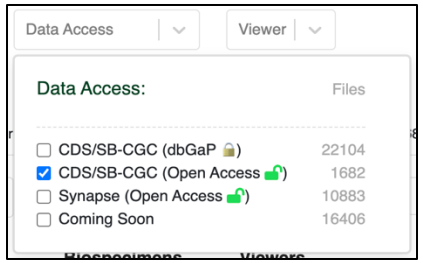
Export via a Data Repository Service (DRS) Manifest.

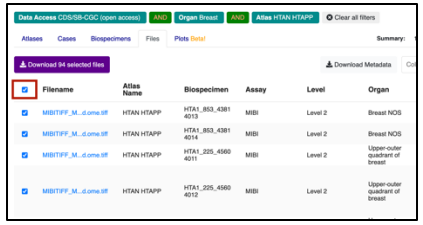
1. Generate a DRS Manifest

General Commons (formerly CDS) Portal

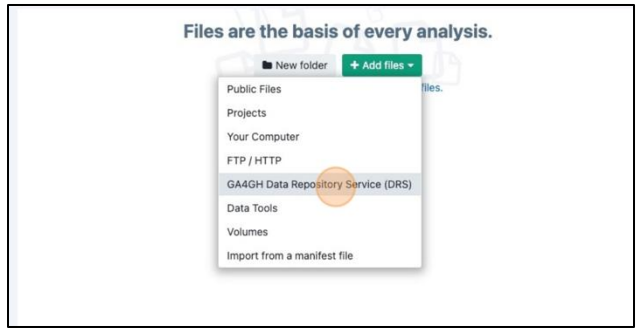


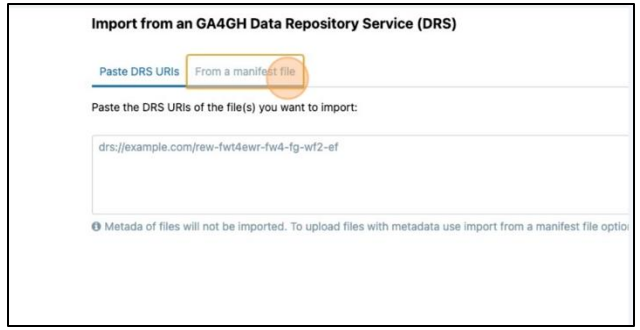
HTAN Data Portal





2. Import Manifest into SB-CGC





https://docs.humantumoratlas.org/open_access/cds_imaging/

State of the art.

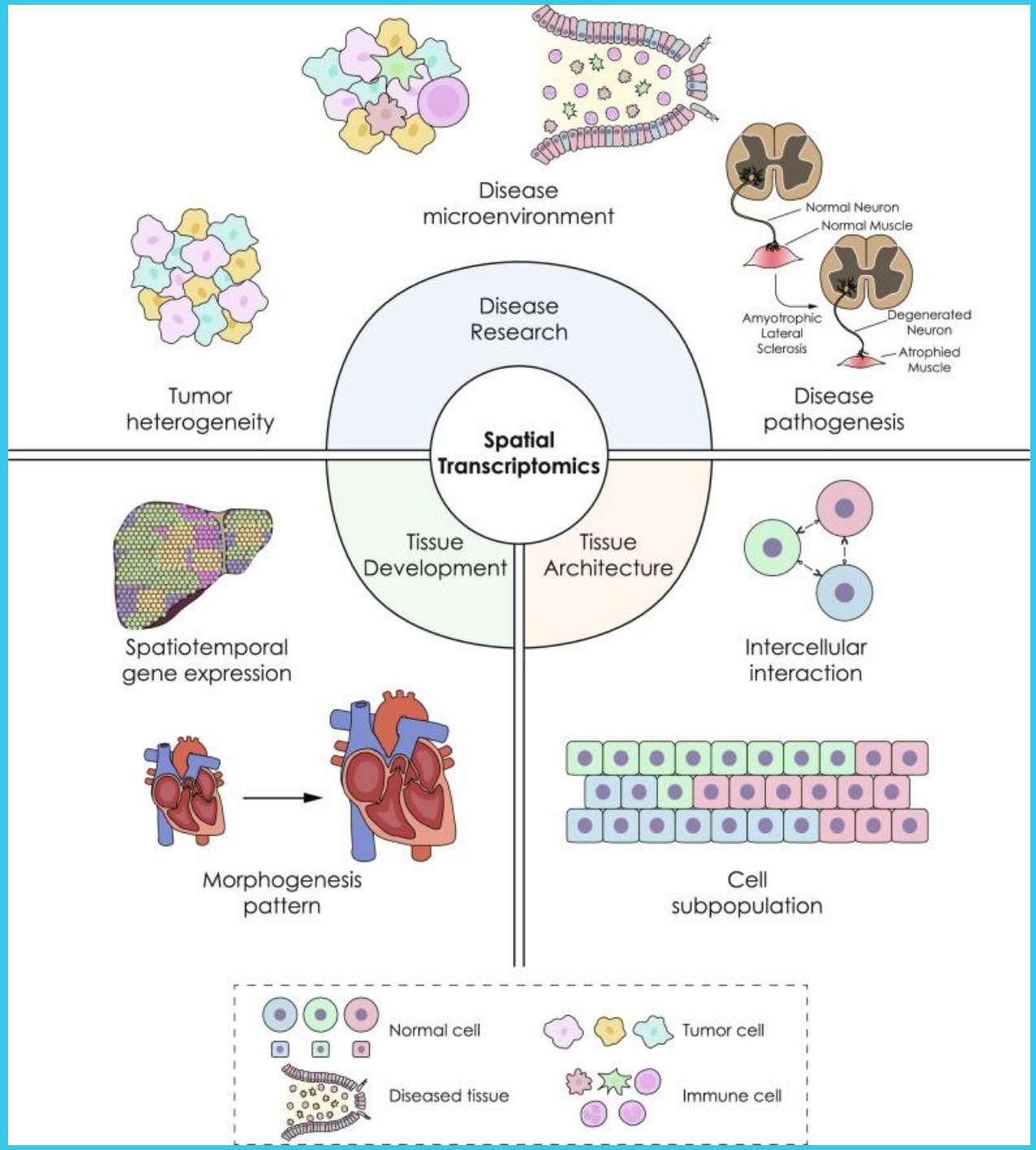
- **134** abstracts in AACR 2023 used spatial transcriptomic analysis
- **>160** abstracts in AACR 2024
- Focus of topics of discussion in sessions and posters
- **25k+** research papers mentioning spatial omics (transcripts, proteomics, metabolomics)

"Single-cell RNA sequencing (scRNA-seq) cannot provide spatial information, while spatial transcriptomics technologies allow gene expression information to be obtained from intact tissue sections in the original physiological context at a spatial resolution.

Various biological insights can be generated into tissue architecture and further the elucidation of the interaction between cells and the microenvironment.

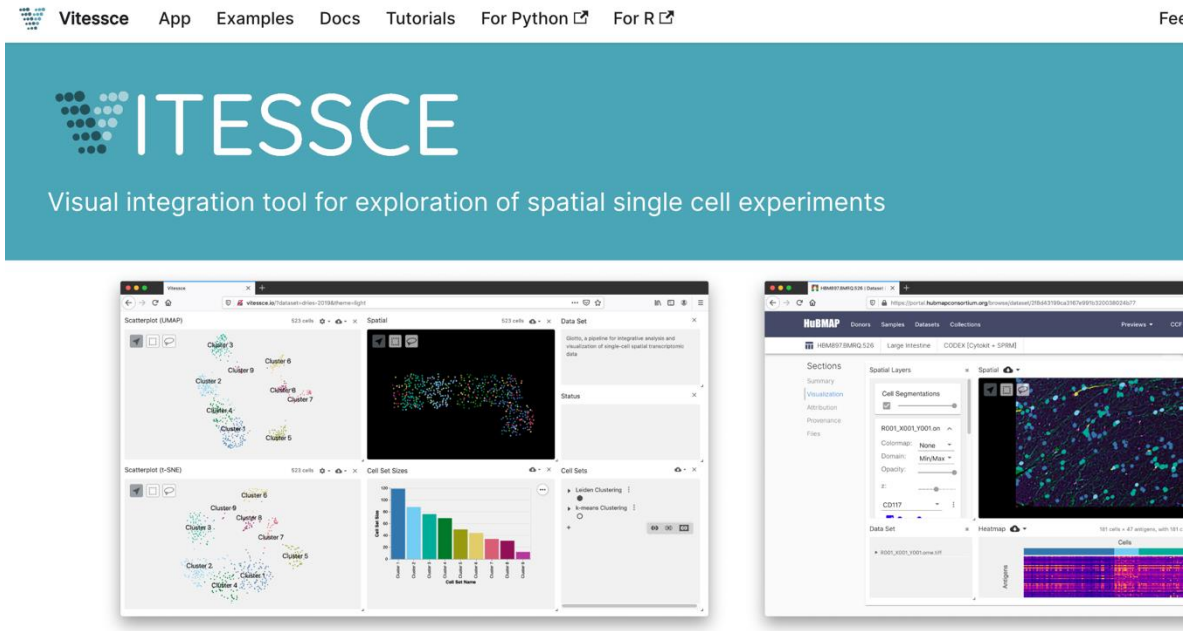
Thus, we can gain a general understanding of histogenesis processes and disease pathogenesis, etc. "

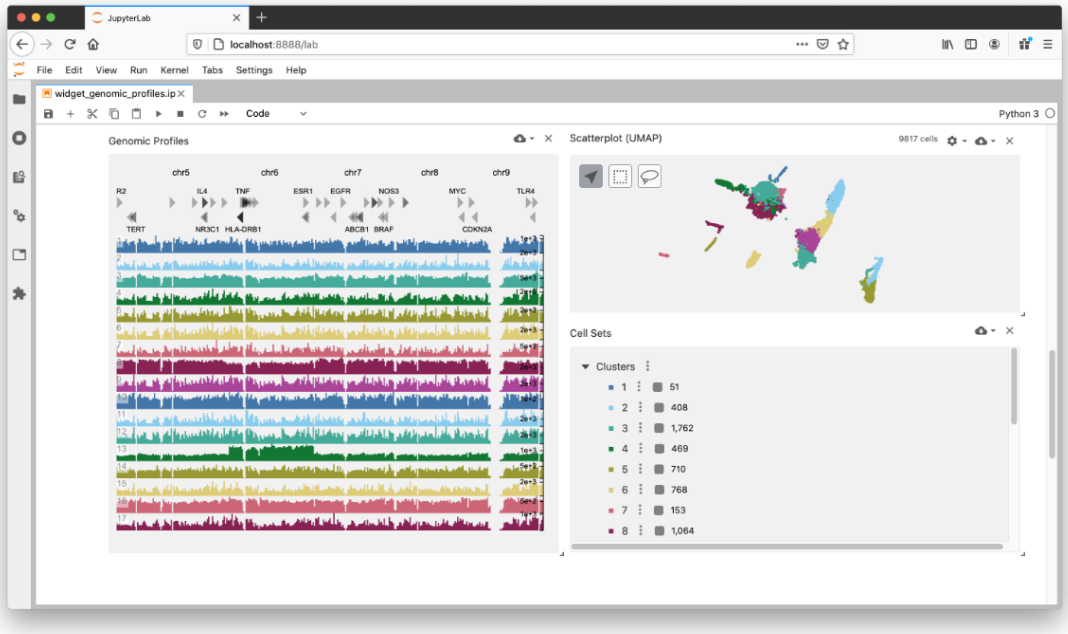
Du J, Yang YC, An ZJ, Zhang MH, Fu XH, Huang ZF, Yuan Y, Hou J. Advances in spatial transcriptomics and related data analysis strategies. J Transl Med. 2023 May 18;21(1):330. doi: 10.1186/s12967-023-04150-2. PMID: 37202762; PMCID: PMC10193345.



What is Vitessce?

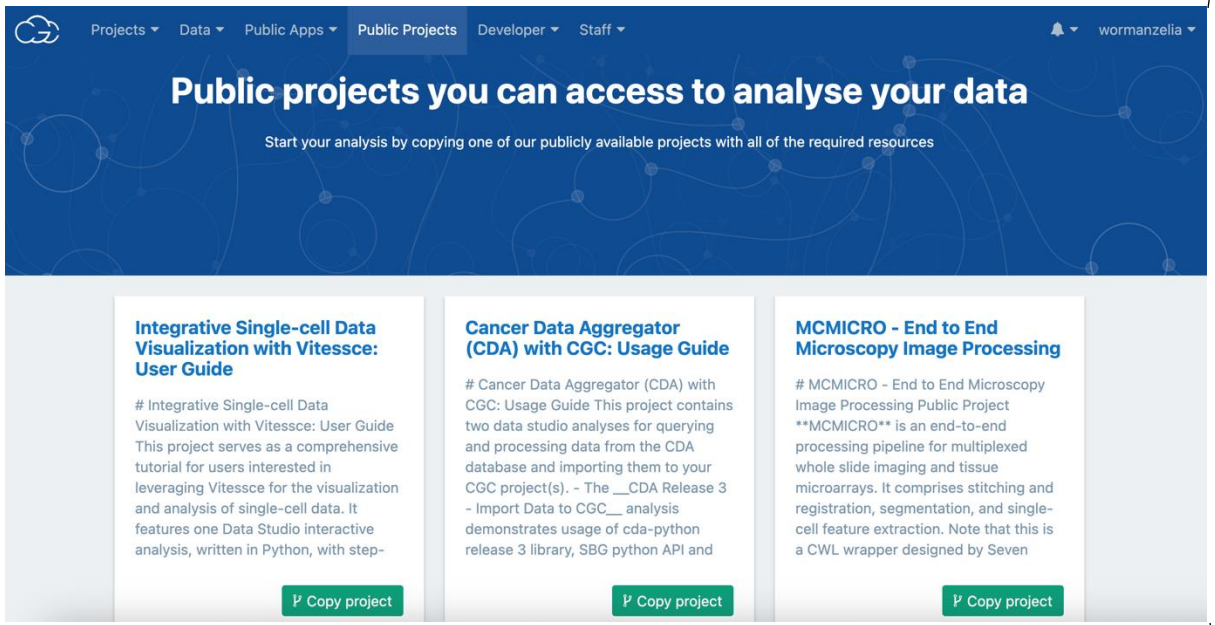
A visual integration tool for exploration of spatial single cell experiments - http://vitessce.io/





Vitessce as an ipywidget in JupyterLab

Vitessce on the SB-CGC.



Projects Data Public Apps Public Projects Developer Staff

Integrative Single-cell Data Visualization with Vitessce: User Guide

This project serves as a comprehensive tutorial for users interested in leveraging Vitessce for the visualization and analysis of single-cell data. It features one Data Studio interactive analysis, written in Python, with step-by-step demonstrations and examples showcasing the integrative capabilities of Vitessce Python API.

Vitessce is a powerful tool for visualizing and analyzing single-cell data, offering integrative capabilities for exploring multimodal and spatially-resolved datasets. With Vitessce, users can interactively visualize their single-cell data in a spatial context, enabling deeper insights into cellular interactions and spatial relationships.

More information about Vitessce can be found in the official documentation, while the latest `vitessce-python` package, along with additional tutorial notebooks, can be found on the GitHub page.

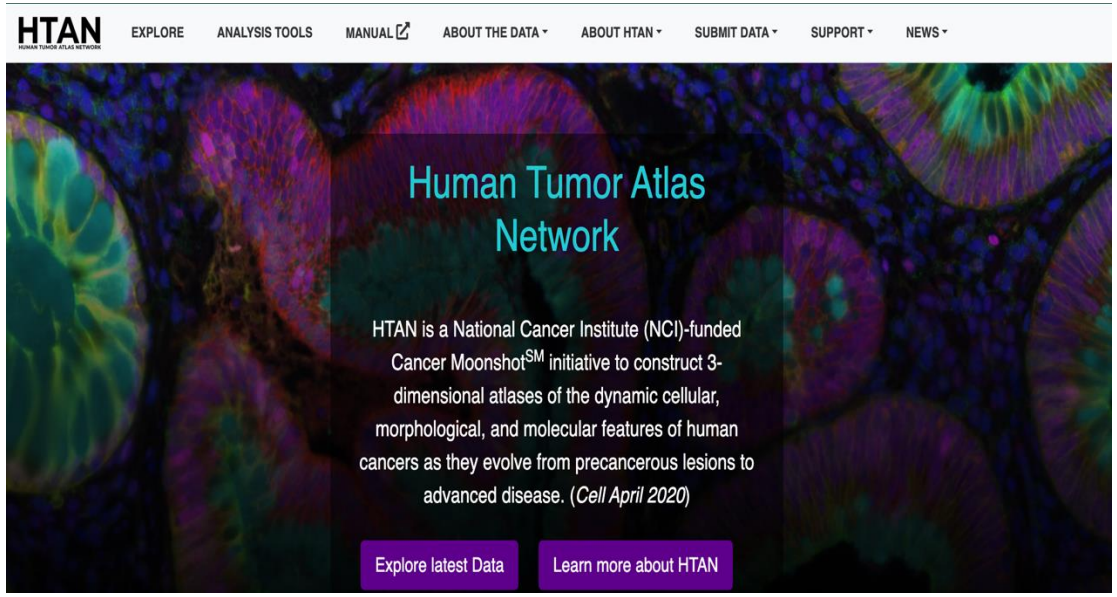
Copying the analysis to your project.

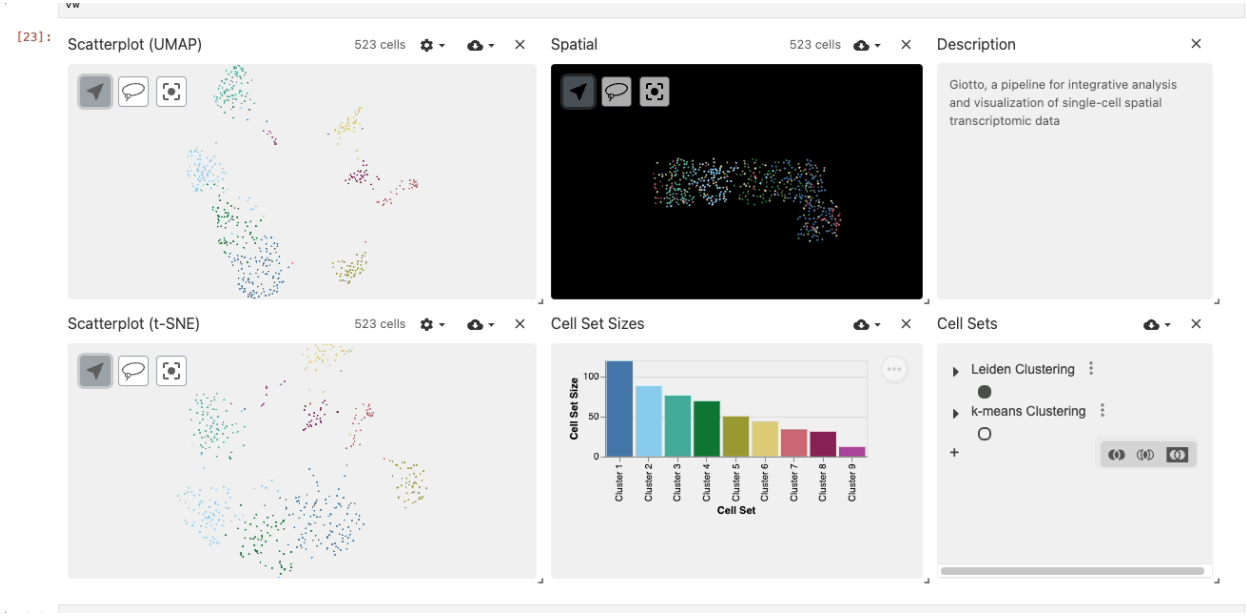
Data Studio analyses are available under the Data Studio tab in this project. To copy the analysis to one of your projects, click the additional actions (three dots) button to the right of **Vitessce Demo Notebook**, click **Copy** and select the target project. You should be able to choose any of the projects you are a part of and in which you have at least *Write* access.

Copying the entire project.

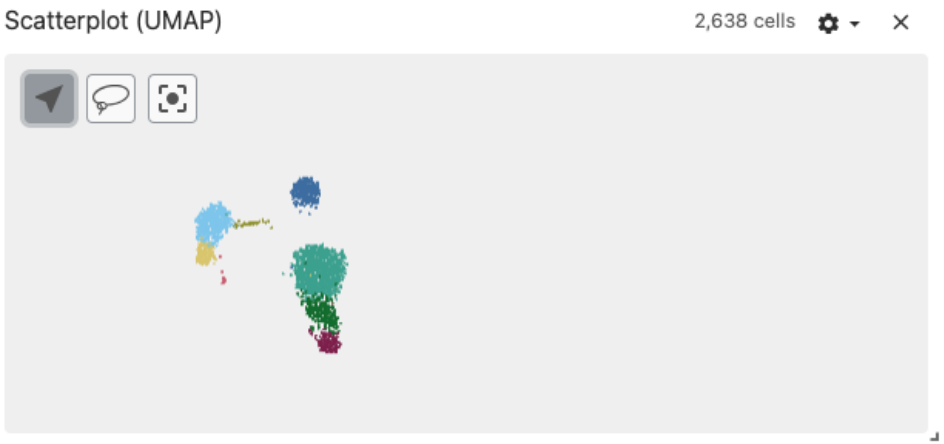
It is also possible to copy the entire project. That way, all Data Studio analyses, as well as Apps, Tasks and Files will be copied. To copy the entire project, click on the information circle next to the public project title (in the top-center of the screen). Then, enter the name for your copy of the project, and select the billing group it will be assigned to.

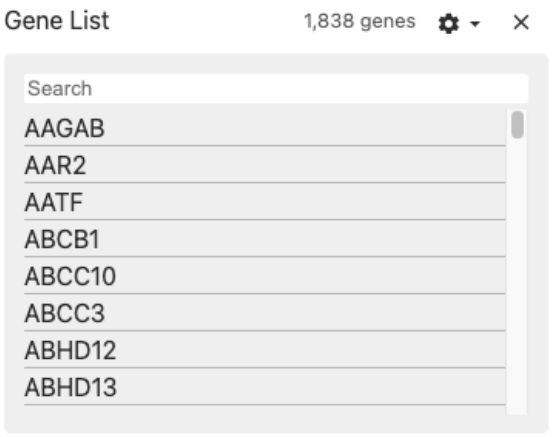
What can we use Vitessce for?

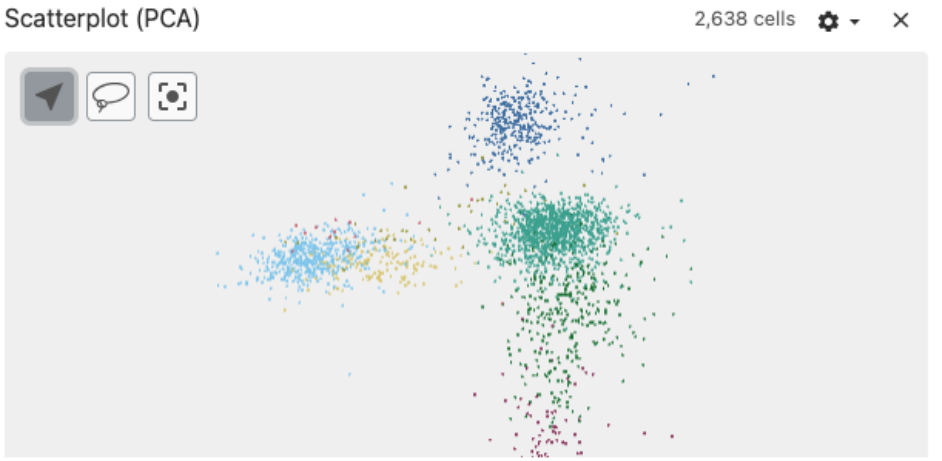


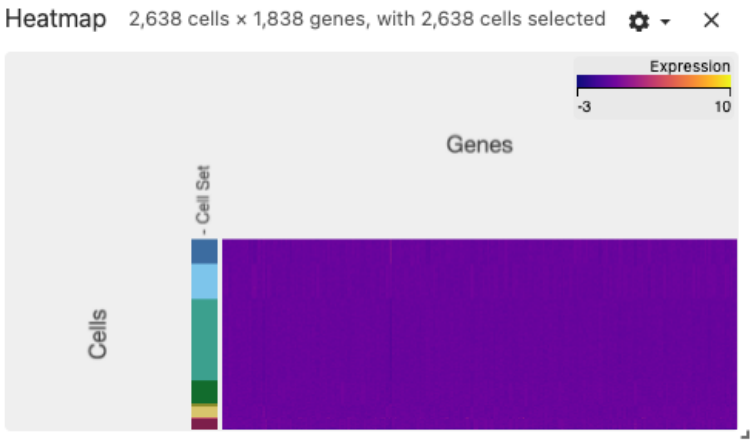


3K PBMC UMAP, PCA, and Scatter Plots

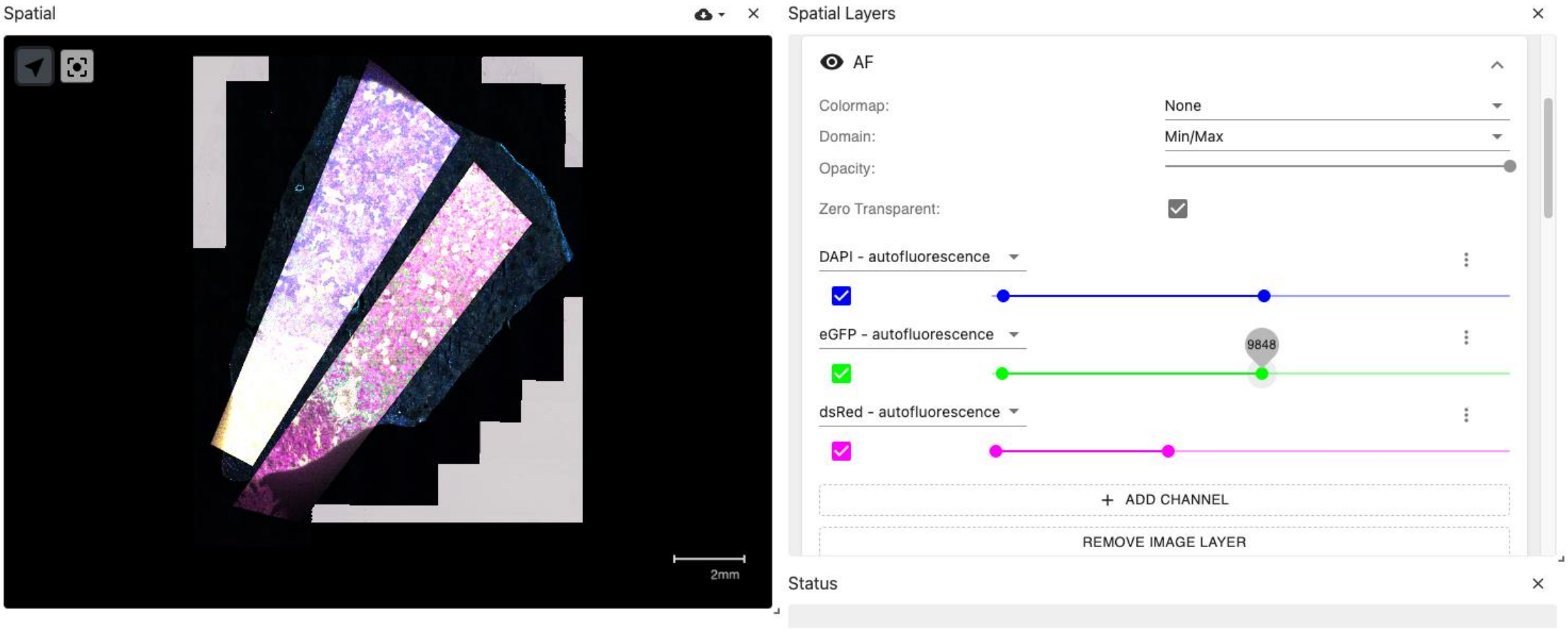






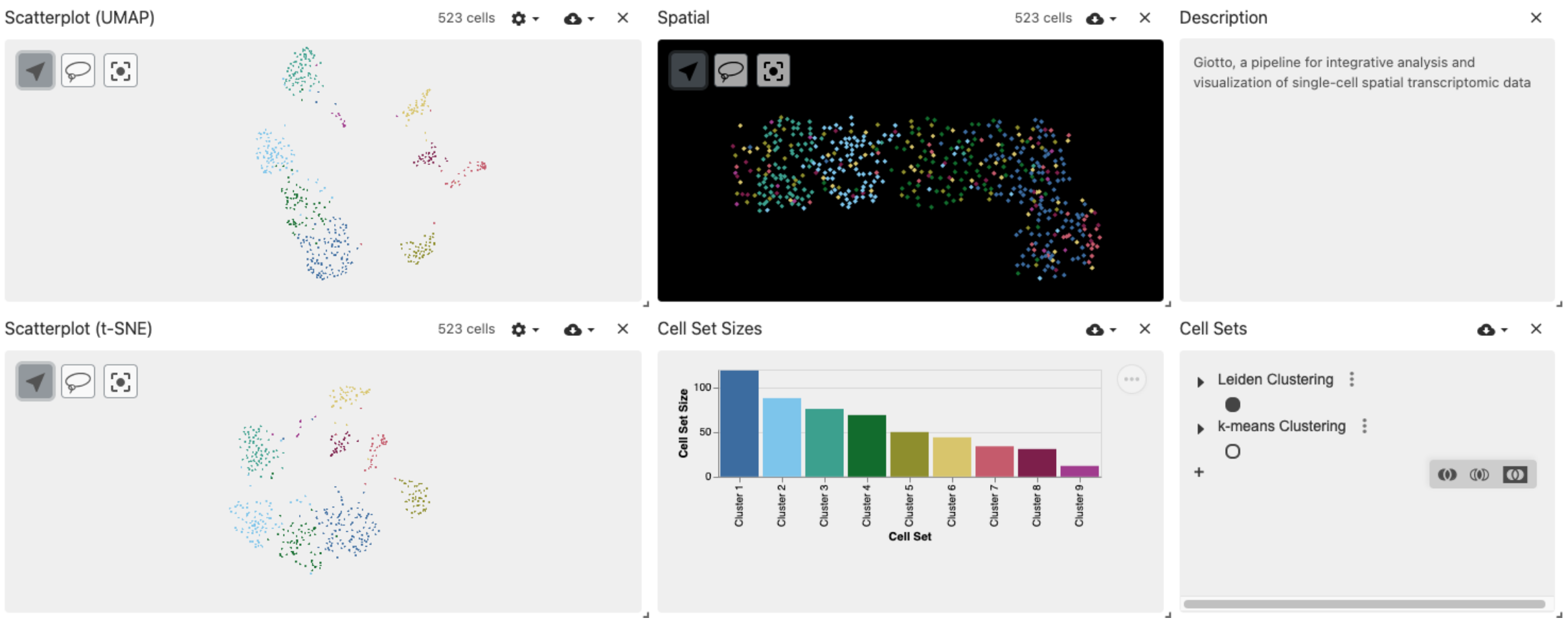


Multimodal Data Images (Spriggins Lab)



Spatial Transcriptomics

[39]:



Want to learn more?

Explore our Public Projects
Request access to our MCMICRO on SB-CGC Walkthrough

Contact us:

- **Email:** cgc-SB@velsera.com
- **Website:** CancerGenomicsCloud.org