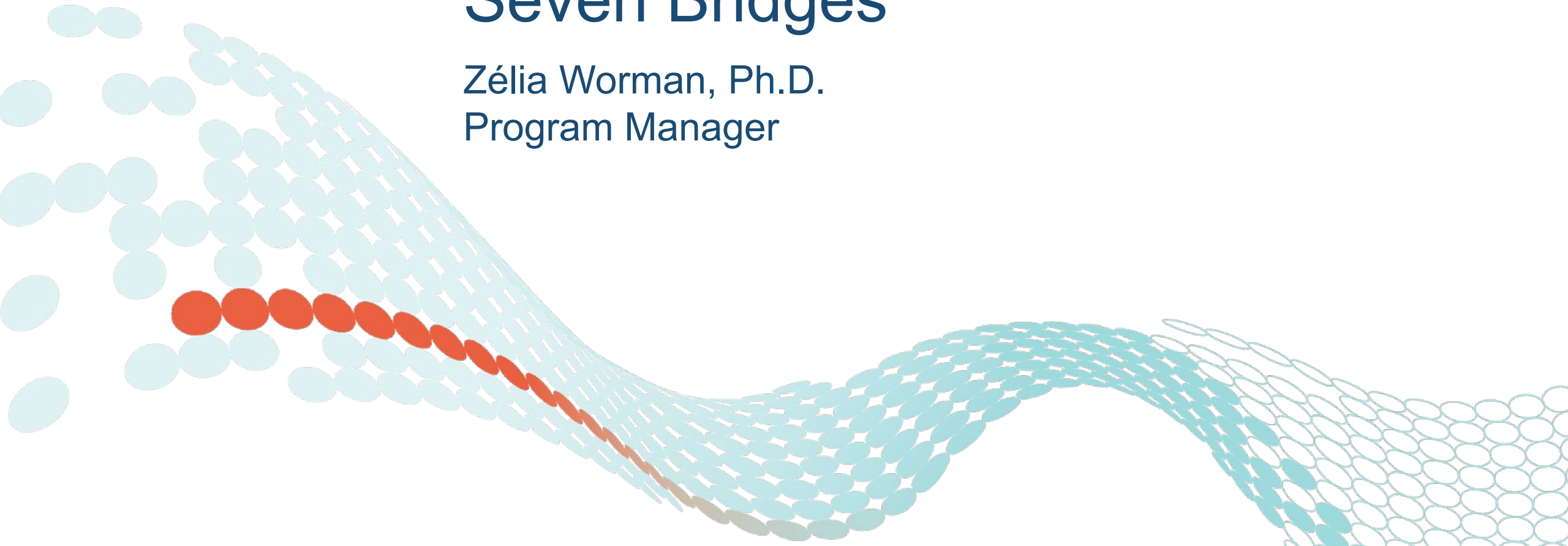
CANCER GENOMICS CLOUD
SEVEN BRIDGES

Introduction to the Cancer Genomics Cloud, powered by Seven Bridges

Zélia Worman, Ph.D.
Program Manager



CANCER GENOMICS CLOUD
SEVEN BRIDGES

Class Overview

- Lecture 1 - Overview
  Hands-on demo: your first project on the CGC
- Lecture 2 - Bulk RNA-seq workflow(s) and sequence alignment
  Hands-on demo: your CGC workflow
- Lecture 3 - Differential expression and visualizations
  Hands-on demo: your CGC visualization
- Lecture 4 - Single-cell RNA-seq overview and workflow
  Hands-on demo: Seurat

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Lecture Overview

- Overview of the Cancer Genomics Cloud, powered by Seven Bridges
- Highlight of key features of how to import data to the CGC and running an analysis
- Hands-on Demo – logging in setting up your first project

Explosion of ‘omics data with ease of sequencing

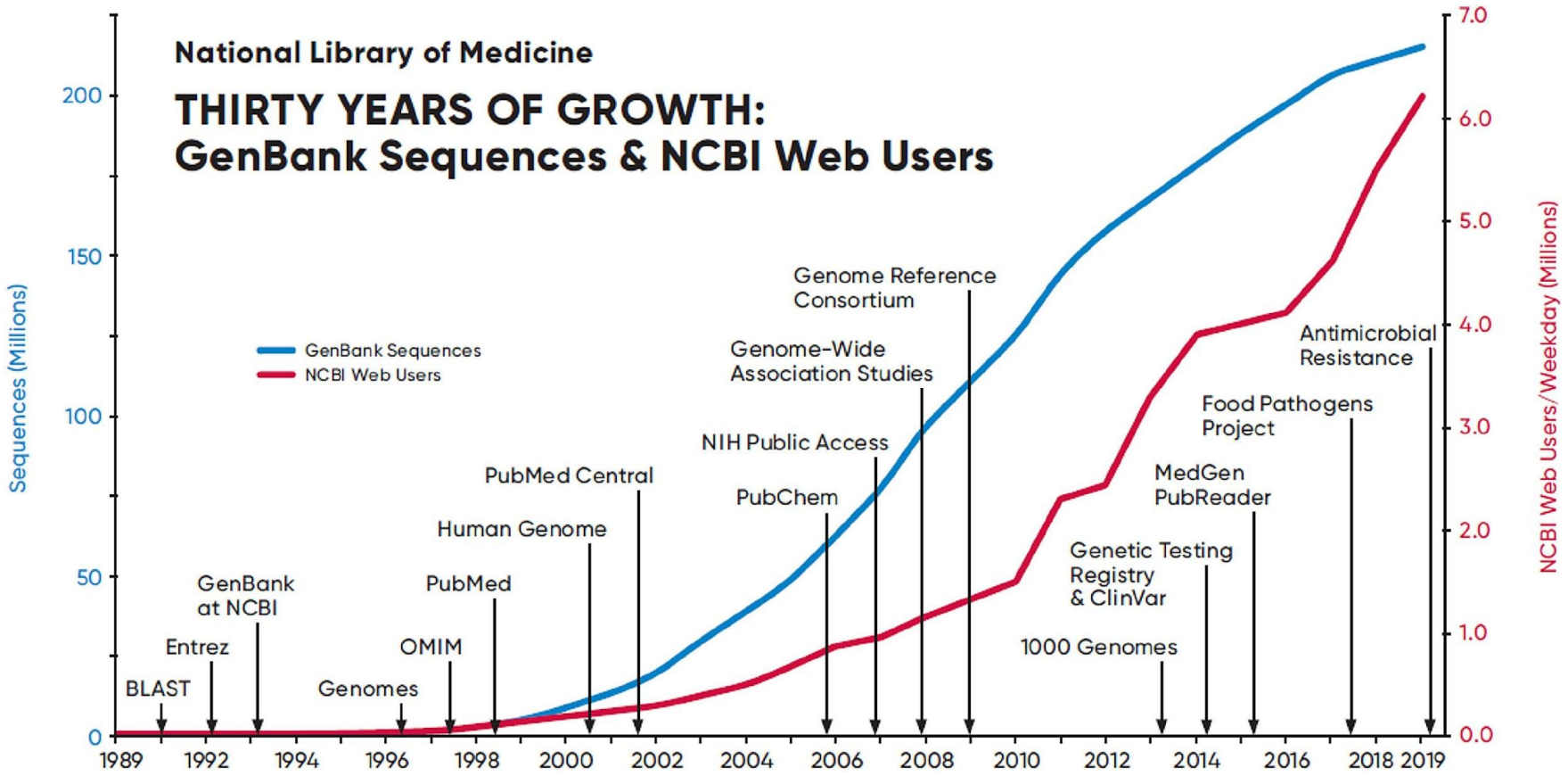
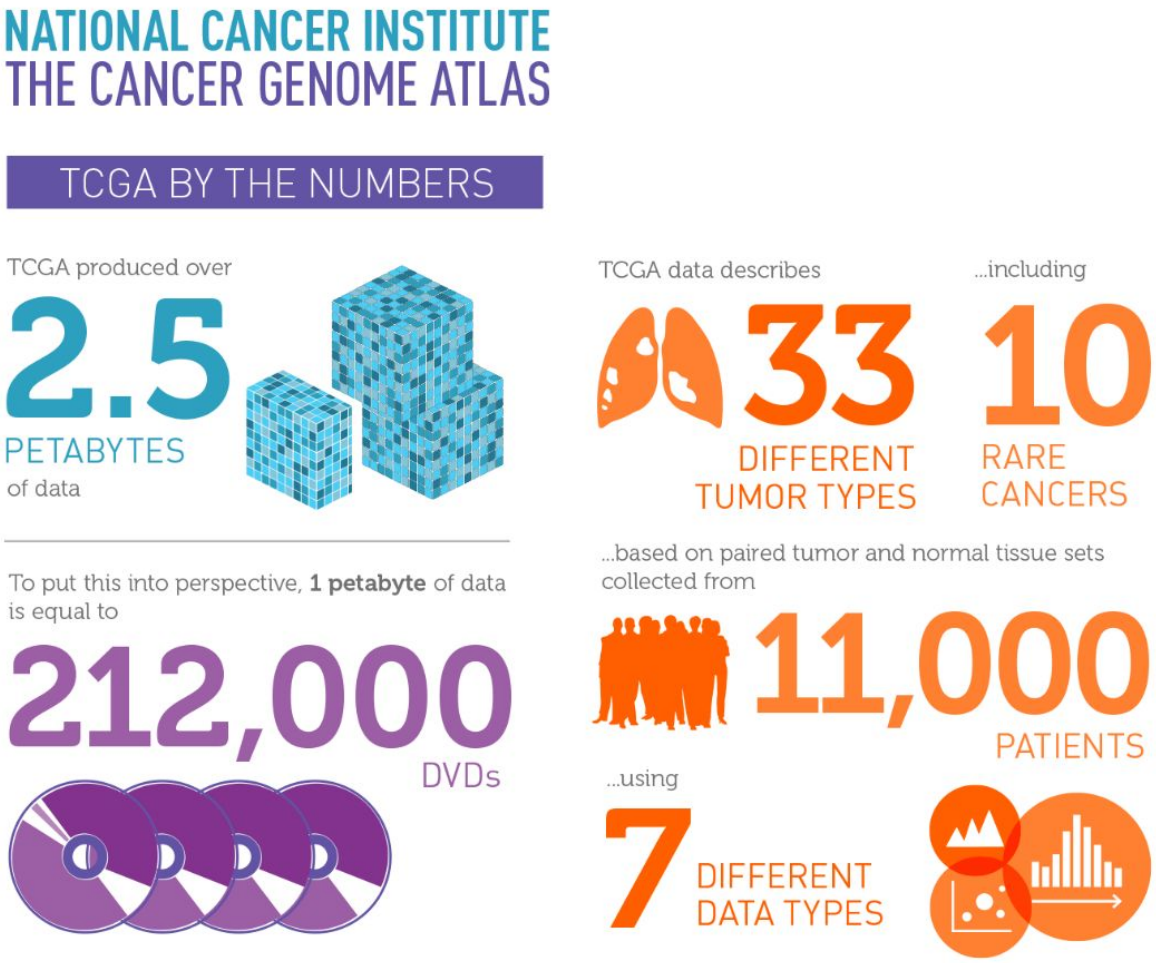
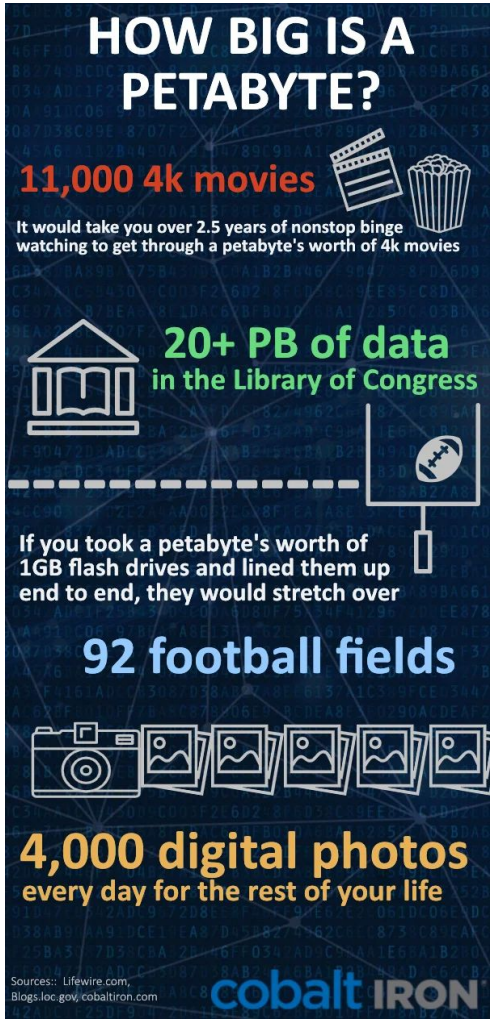


Fig. 1. **Growth of GenBank sequences and NCBI web users through 2019.** Figure from the Department of Health and Human Services National Institutes of Health. Jim Gaffney, et. al. Open access to genetic sequence data maximizes value to scientists, farmers, and society, Global Food Security, Volume 26, 2020.

To give you one example, the TCGA is just one of many datasets currently available!





Increasingly large datasets bring challenges to data analysis

What is the cloud?

https://blog.vsoftconsulting.com/blog/what-exactly-is-the-cloud

The CGC Democratizes Complex Analyses

- A stable, secure, and highly customizable cloud storage and computing platform
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines

Easy data management

Secure collaboration & managed billing

Flexible & fully reproducible methods

Optimized bioinformatics algorithms

Scalable computation

Extensible & developer friendly tools

The Seven Bridges Cancer Genomics Cloud (CGC)

LEARN

TEACH

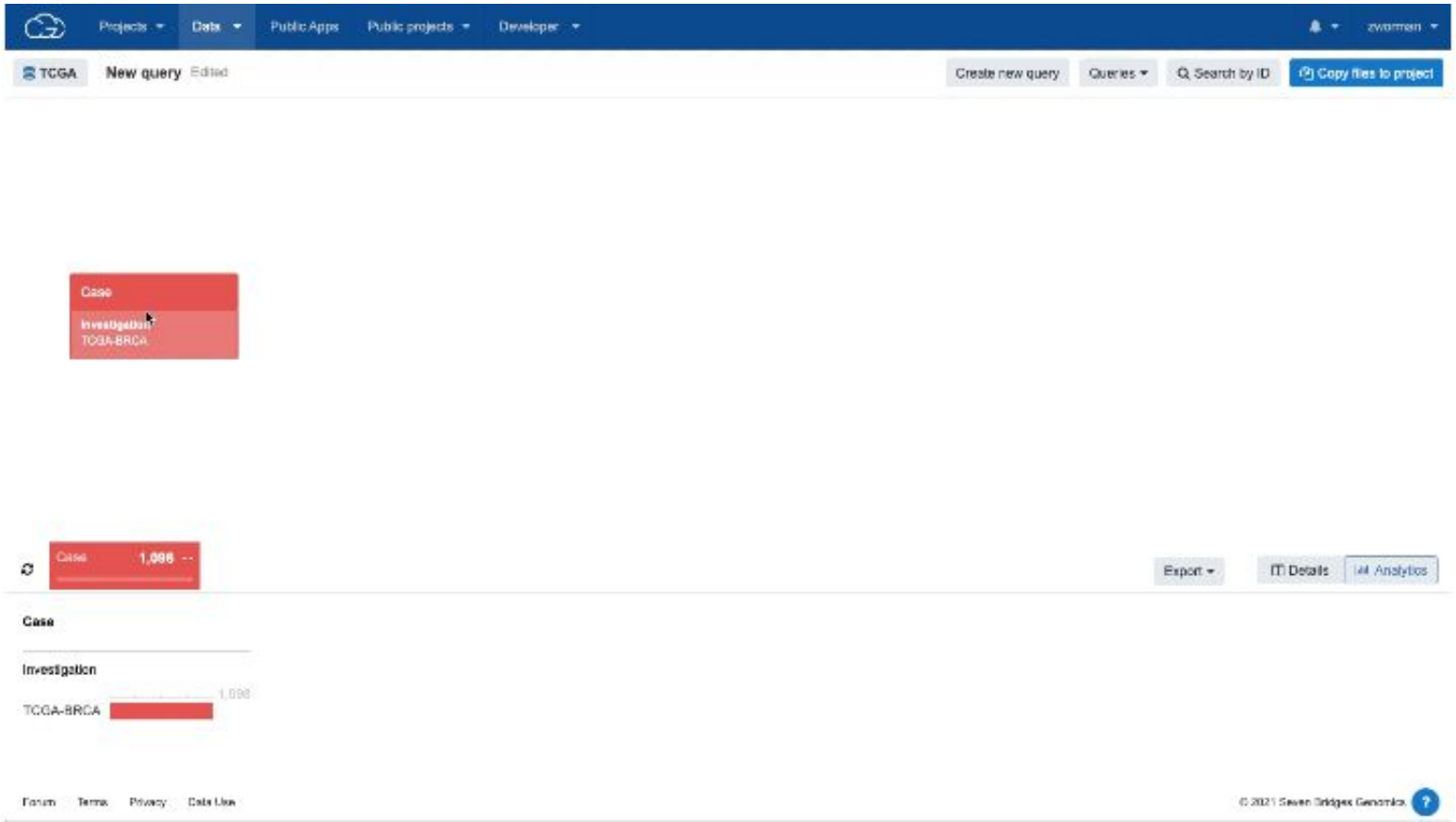
COLLABORATE

INNOVATE

For more information, visit us at https://www.cancergenomicscloud.org/

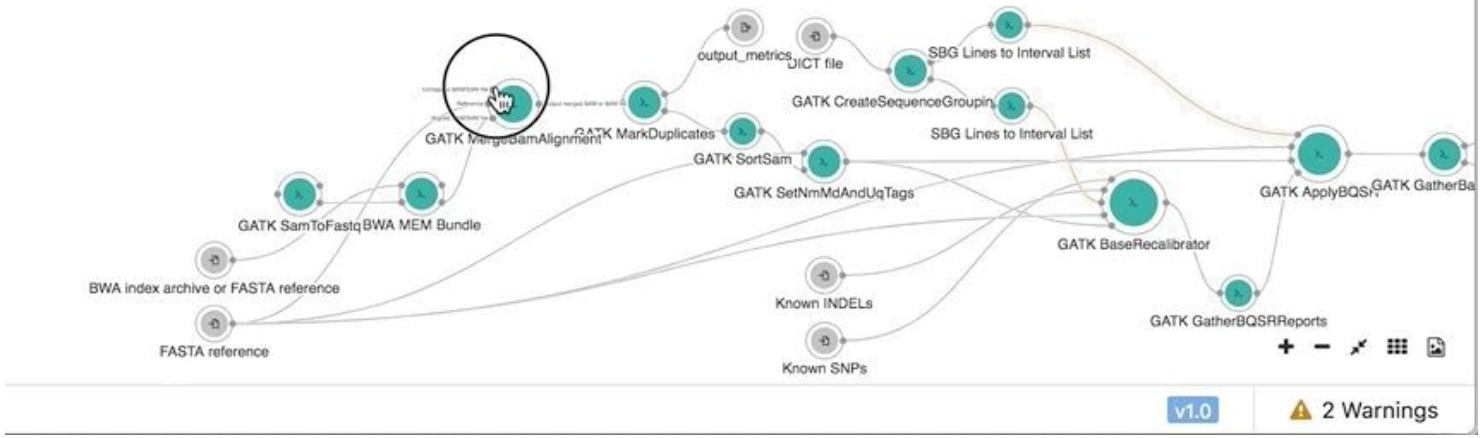
The CGC provides an easy way to find data

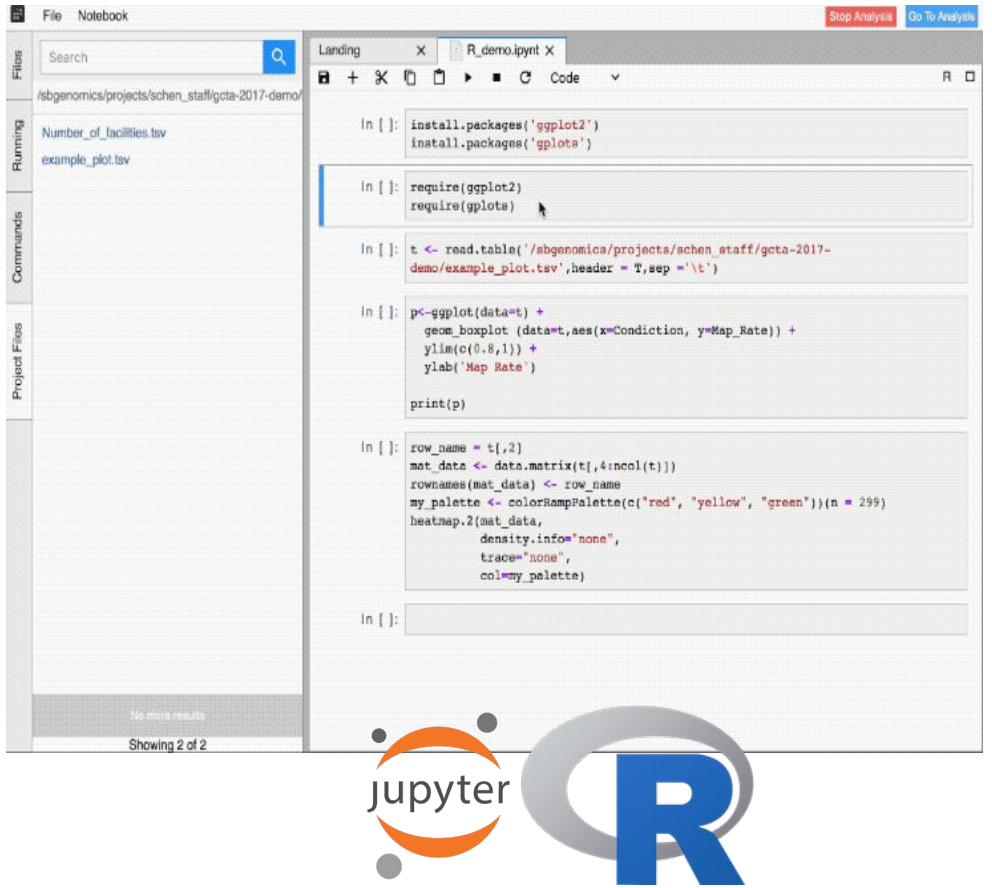
Visually explore and access **3+ PB** of multi-omic public data through our user-friendly portal and APIs.



The CGC provides an easy way analyze data

Use the **600+** cloud- and cost-optimized tools in our Public Apps library OR deploy custom tools using **Rabix Composer**, Jupyter notebooks or R packages





High impact publications on the CGC

PNAS
Proceedings of the National Academy of Sciences of the United States of America
Keyword, Author, or
Home | Articles | Front Matter | News | Podcasts | Authors

RESEARCH ARTICLE

Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers

Genome Biology
Home About Articles Submission Guidelines

Short Report | Open Access | Published: 05 August 2021

Specific splice junction detection in single cells with SICILIAN

Roozbeh Dehghannasiri, Julia Eve Olivieri, Ana Damljanovic & Julia Salzman

Genome Biology **22**, Article number: 219 (2021) | Cite this article

1336 Accesses | **30** Altmetric | Metrics

DATABASE
The Journal of Biological Databases and Curation

Bioinformatics tools developed to support BioCompute Objects

Janisha A Patel, Dennis A Dean, Charles Hadley King, Nan Xiao, Soner Koc, Ekaterina Minina, Anton Golikov, Phillip Brooks, Robel Kahsay, Rahi Navelkar ... Show more

Database, Volume 2021, 2021, baab008, https://doi.org/10.1093/database/baab008

Published: 30 March 2021 **Article history** ▾

RNAbiology

Novel, abundant Drosha isoforms are deficient in miRNA processing in cancer cells

Lisheng Dai, Lillian Hallmark, Xavier Bofill De Ros, Howard Crouch, Sean Chen, Tony Shi, ...show all

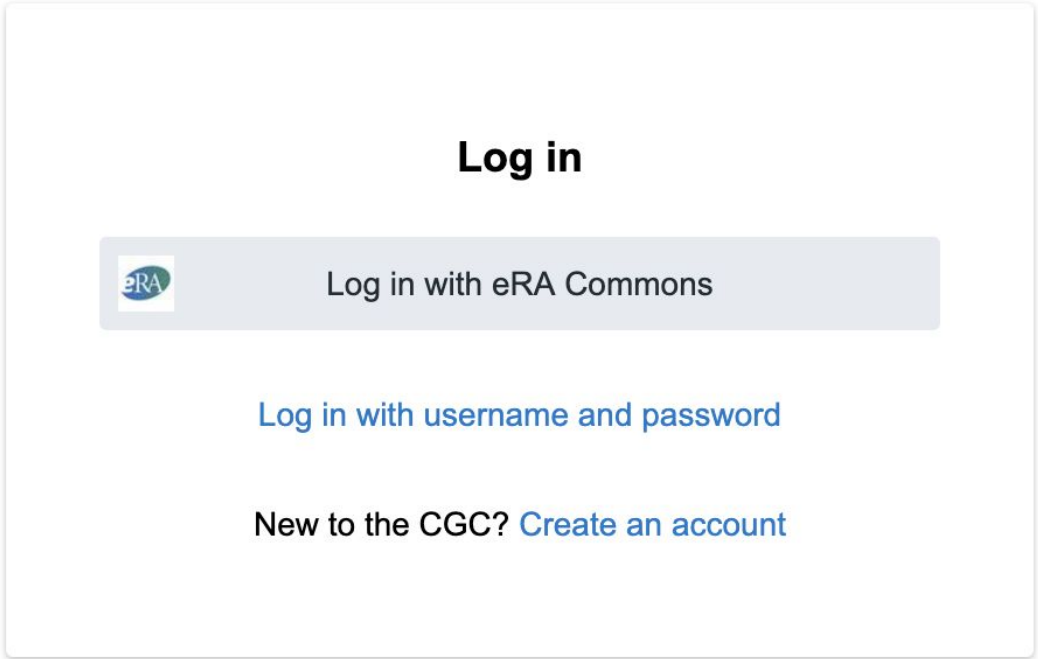
Pages 1603-1612 | Received 02 Jan 2020, Accepted 19 Aug 2020, Accepted author version posted online: 20 Aug 2020, Published online: 30 Aug 2020

Download citation https://doi.org/10.1080/15476286.2020.1813439 Check for updates

Quickly get an account on the CGC

CANCER GENOMICS CLOUD
SEVEN BRIDGES

- Sign up with your email
- Option to connect with eRA Commons to access controlled data
- $300 of pilot funding to get your project started
- Collaborative grants of up to $10.000 are available for larger projects



Join us at Office Hours!

- Comprehensive online documentation and training resources
- Technical support from a team of scientists, bioinformaticians, and engineers
- **Office Hours on at 10:00 am ET Tuesday and 2:00 pm ET Thursday office hours.**
- **All are welcome!**

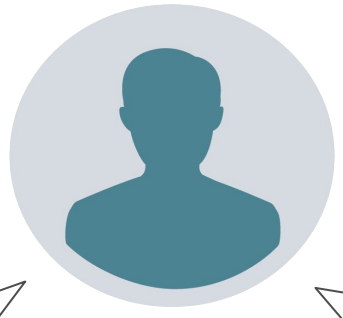
CGC Platform Components

- Key features of running an analysis on CGC
- Several ways to import data to the CGC

What's the research question you want to answer?

How do I check if two genes have different expression in tumor and normal tissues?

How do I test my hypothesis that variants impact gene expression and may increase cancer risk?



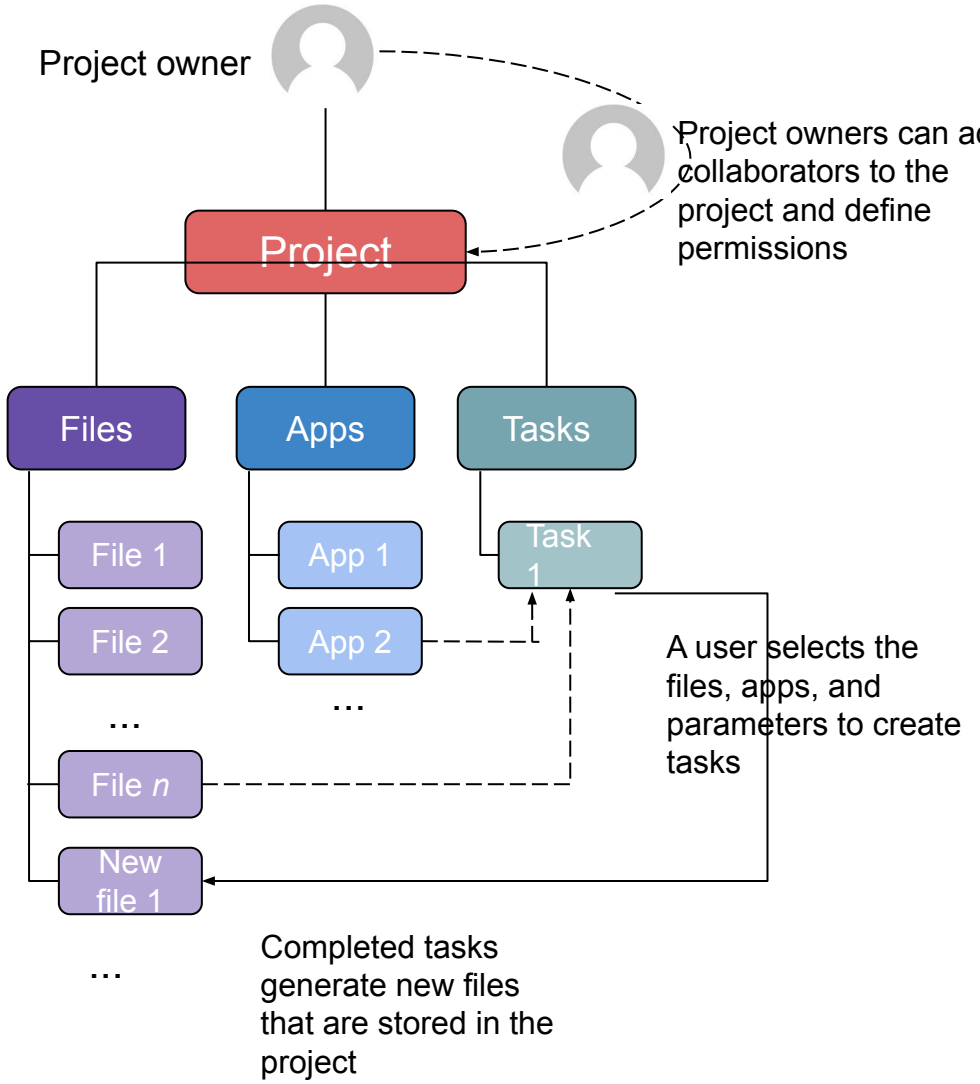
I don't have enough samples, how can I compare my data to publicly available databases?

How do I analyzed cancer types using imaging techniques such as MRI and CT scans?

Prototypical User Flow

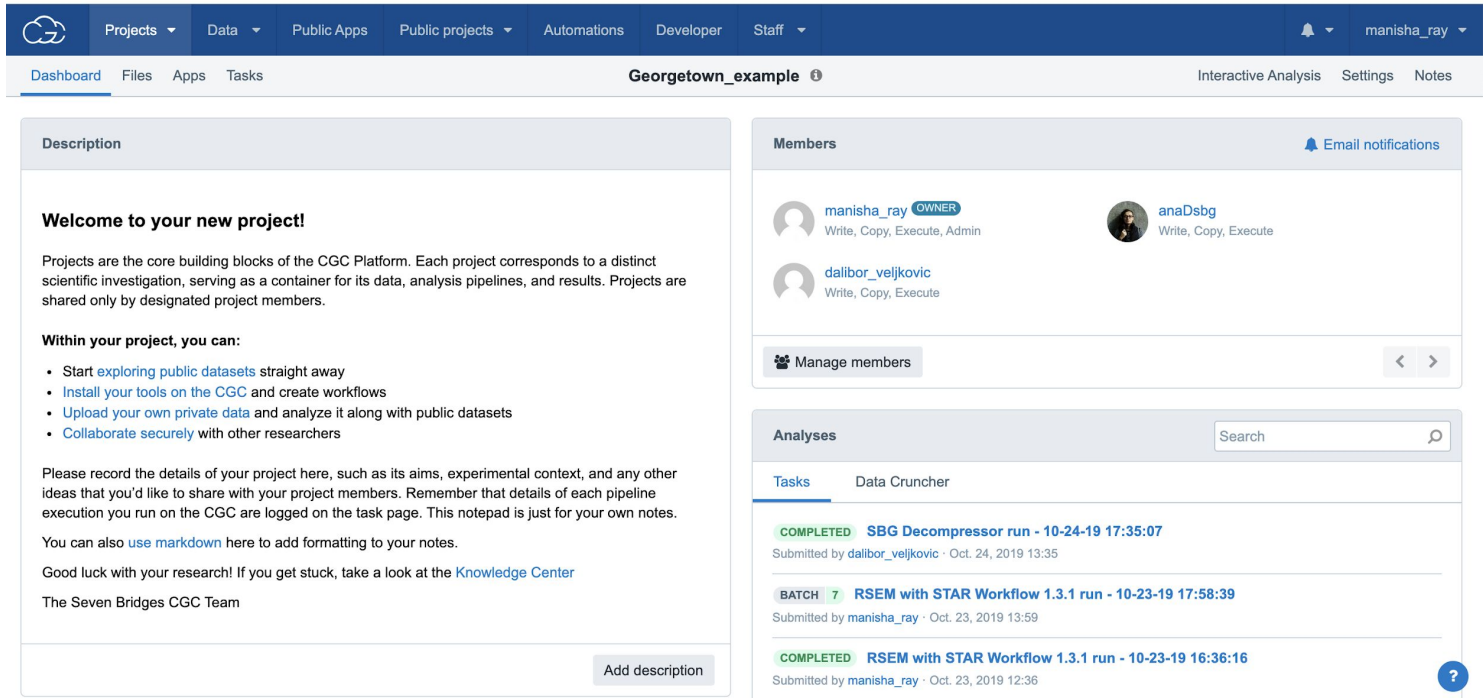
Create a Project	Select/access data	Select/create tools	Create and run analysis
Organizational unit within platform	Many ways to find and bring in data: • Data Browser • Desktop uploader • Command line uploader • Volumes	Tools, workflows, and software packages • Public Apps Gallery • Tools or workflows wrapped in CWL • R packages • Python libraries	Specify how an analysis will be run • Task page • Notebooks in RStudio or JupyterLab

Projects organize files, methods, and results

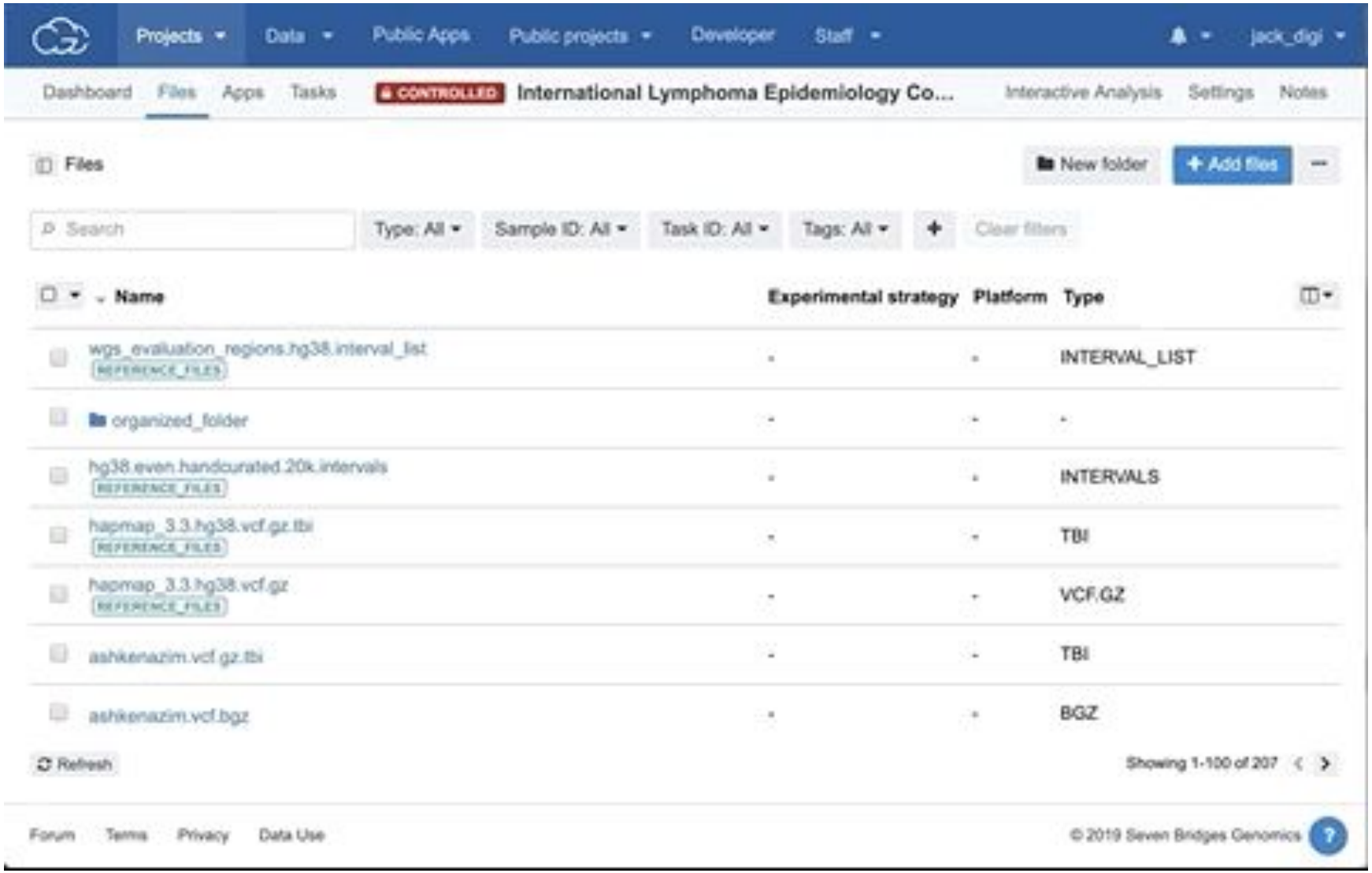


Also known as *workspaces* or *sandboxes*

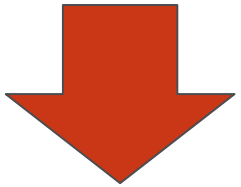
Easily manage collaborators and permissions



Collaborate and share results quickly and easily



Prototypical User Flow



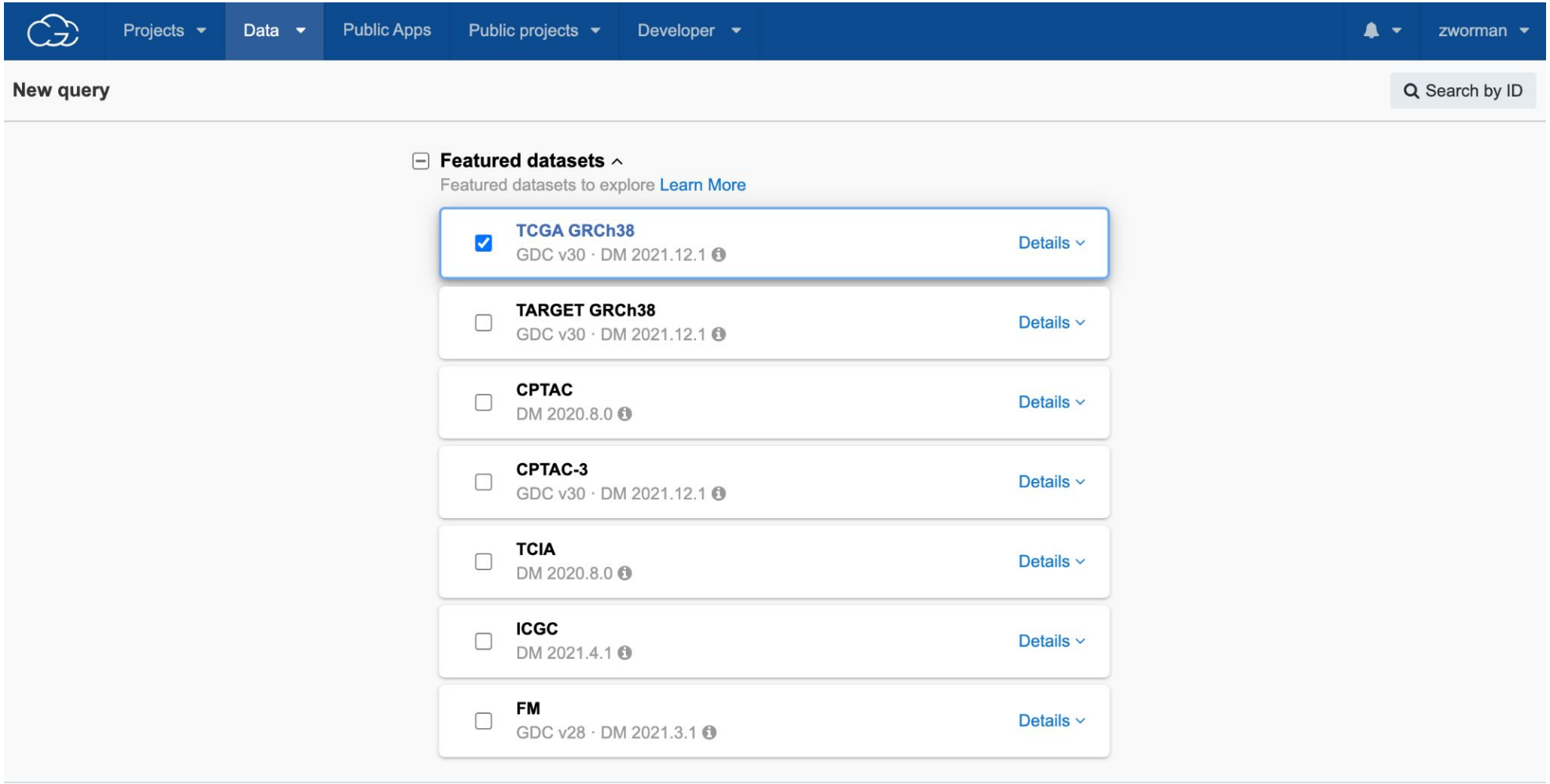
Create a Project	Select/access data	Select/create tools	Create and run analysis
Organizational unit within platform	Many ways to find and bring in data: • Data Browser • Desktop uploader • Command line uploader • Volumes	Tools, workflows, and software packages • Public Apps Gallery • Tools or workflows wrapped in CWL • R packages • Python libraries	Specify how an analysis will be run • Task page • Notebooks in RStudio or JupyterLab

Access and search huge public datasets on the CGC

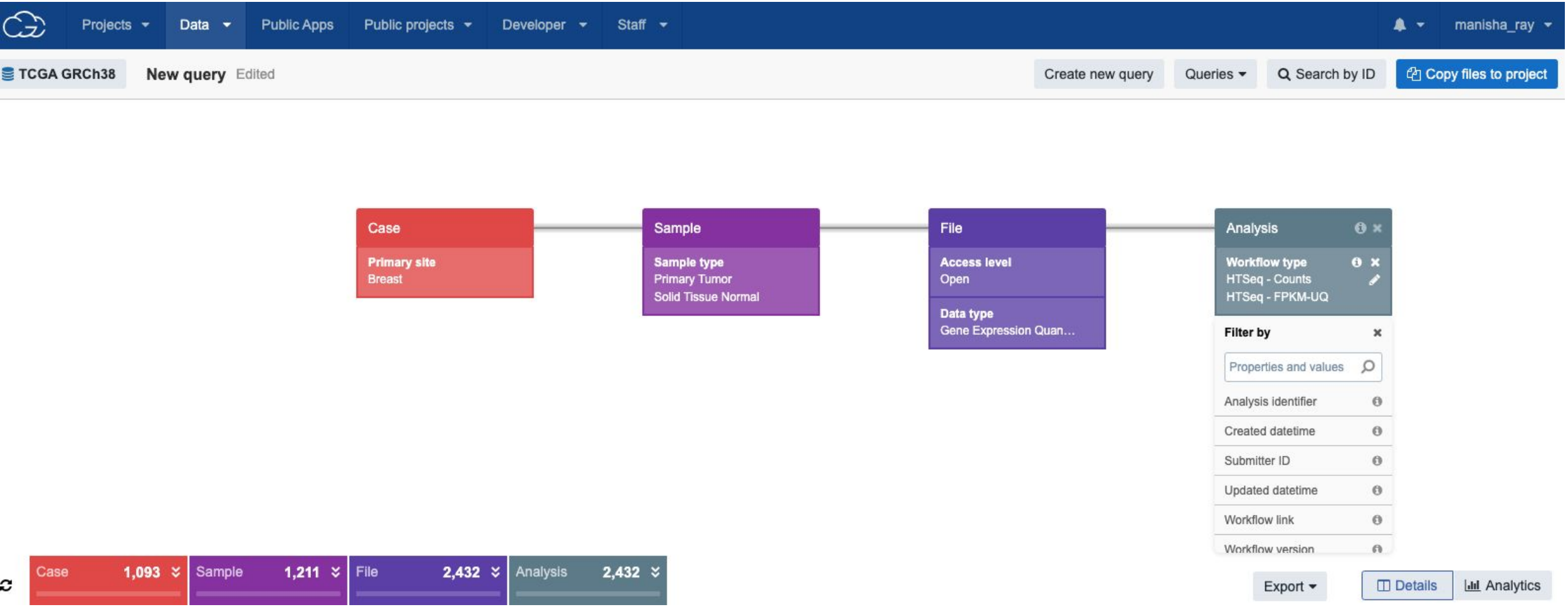
- Web and API based query tools
- Ability to search multiple datasets together

TCGA
TARGET
FOUNDATION MEDICINE®
ICPC INTERNATIONAL CANCER PROTEOGENOME CONSORTIUM
ICGC
CCLE Cancer Cell Line Encyclopedia
HUMAN CELL ATLAS
CPTAC
THE CANCER IMAGING ARCHIVE
Personal Genome Project
Children's Brain Tumor Network
Until every child is cured
SIMONS FOUNDATION

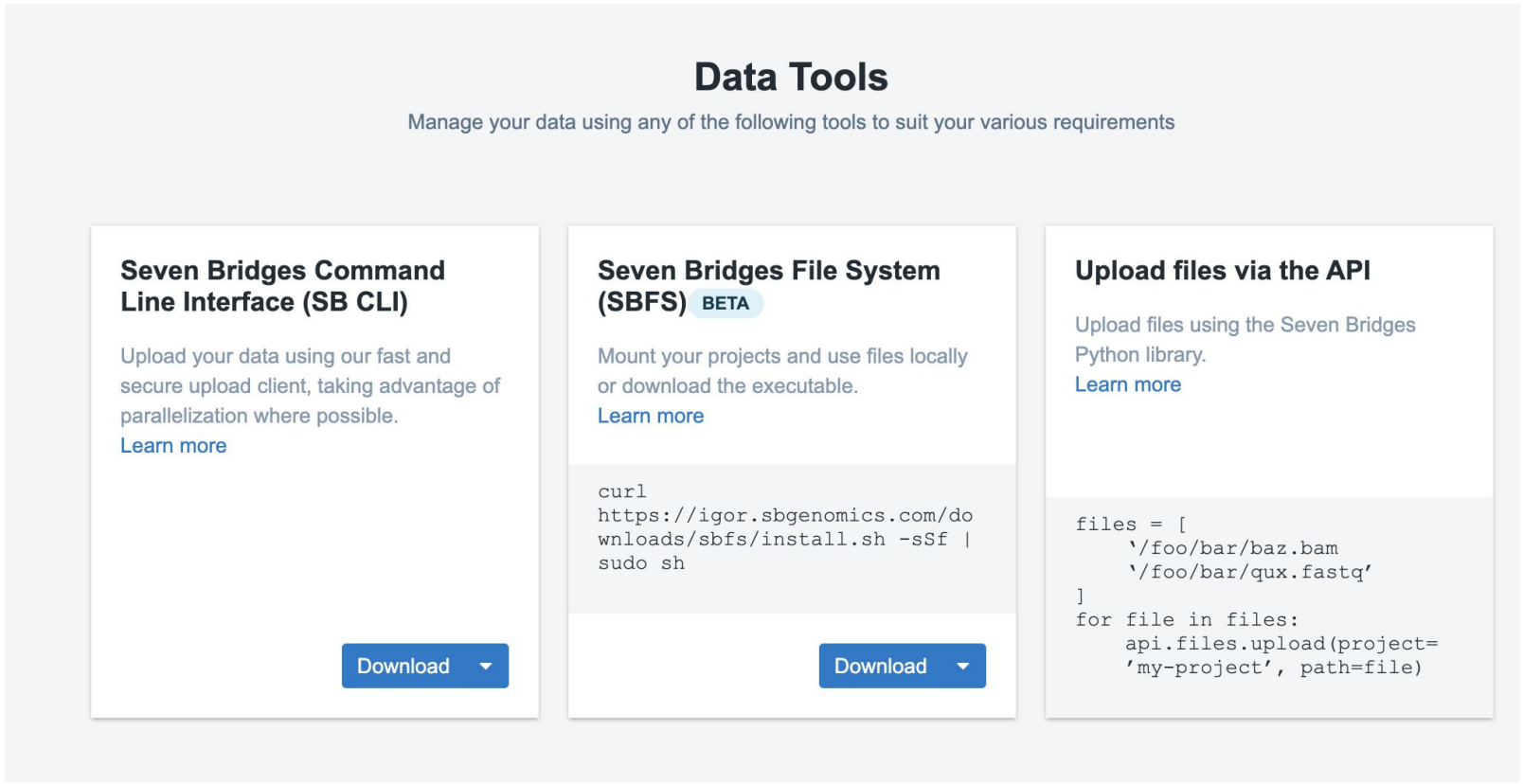
Access and search huge public datasets on the CGC

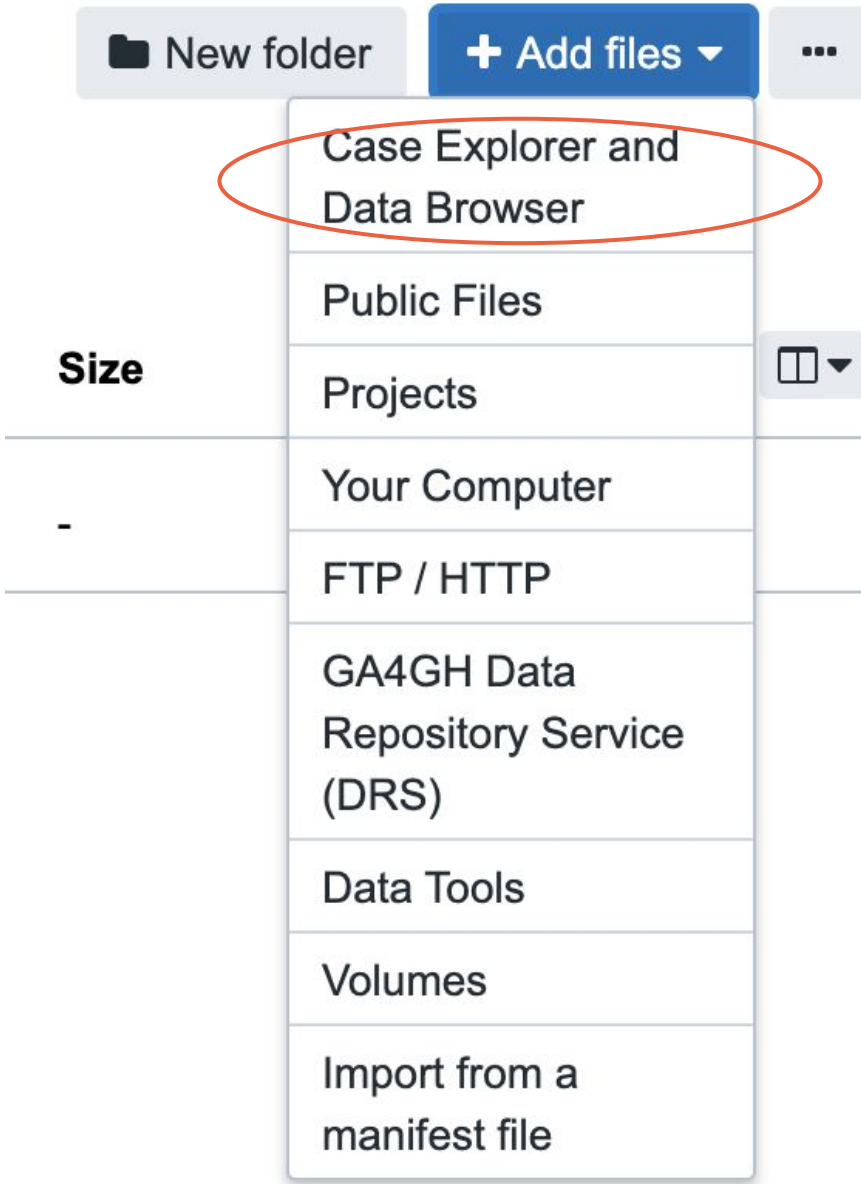


Data Browser feature for building queries and cohorts

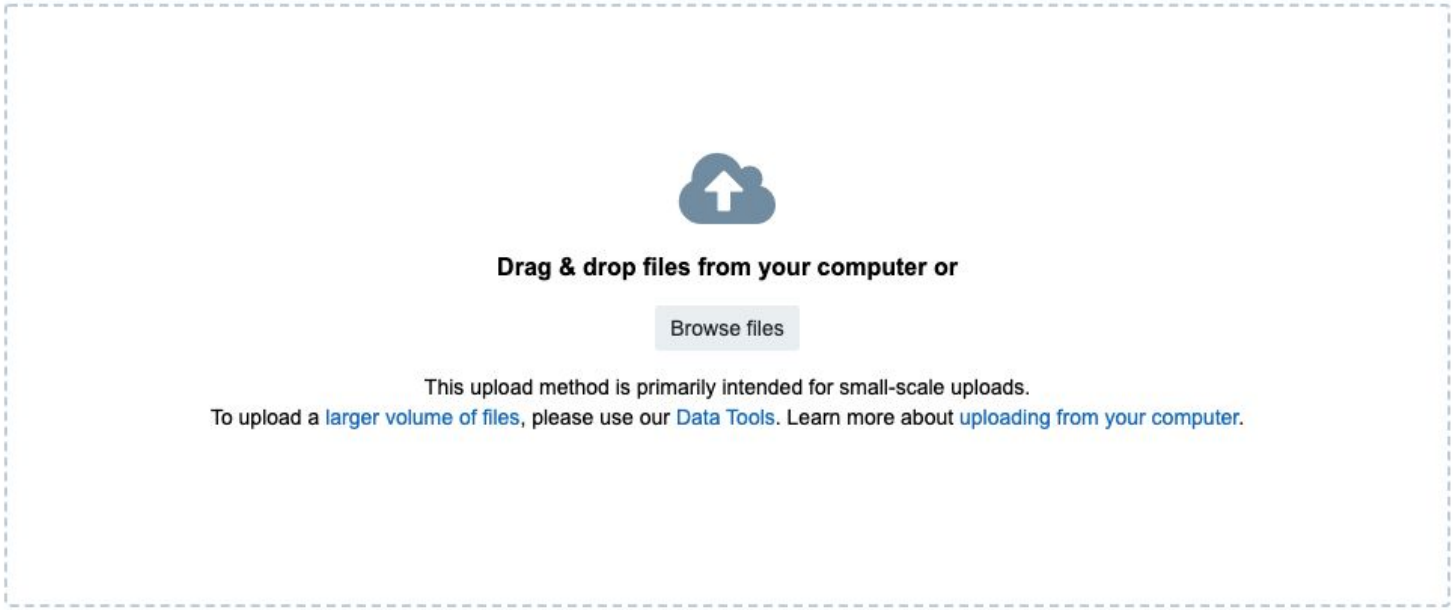


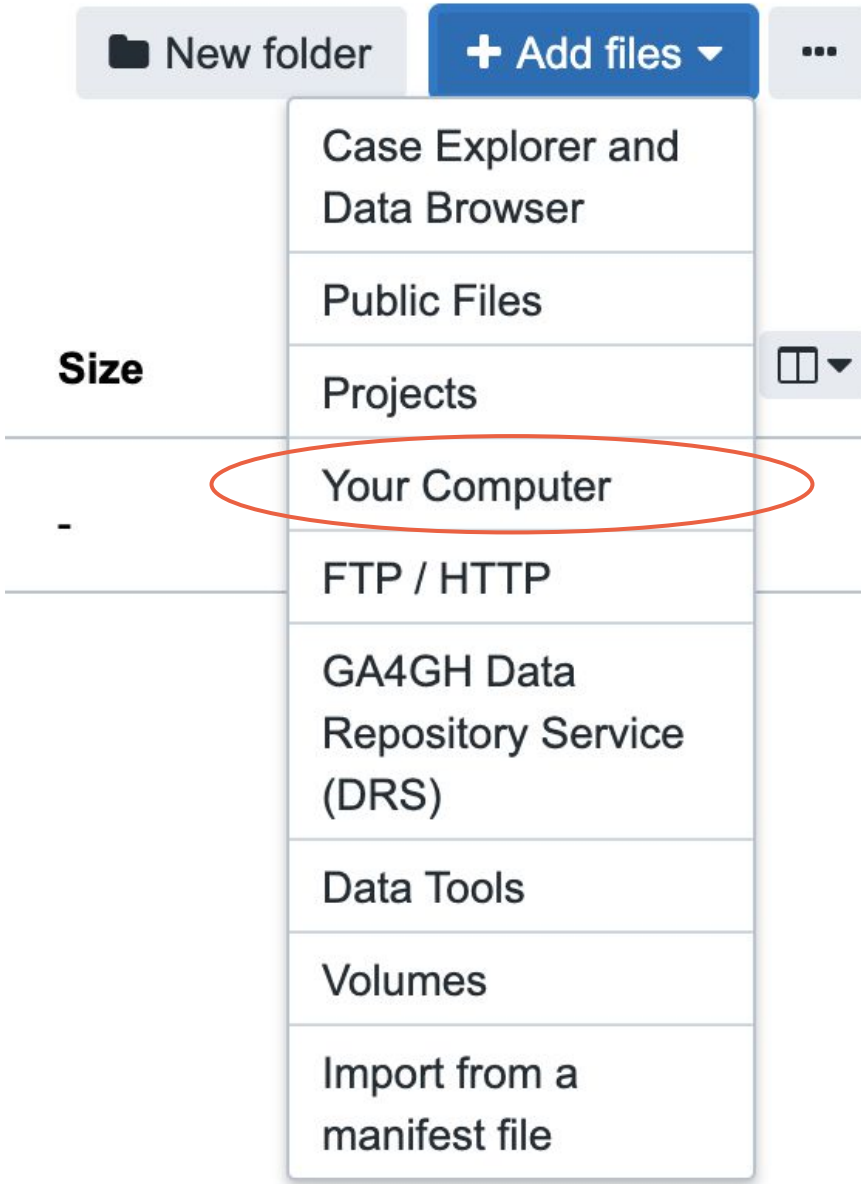
Conveniently bring in your own data





Conveniently bring in your own data





Prototypical User Flow

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Tools, workflows, and software packages

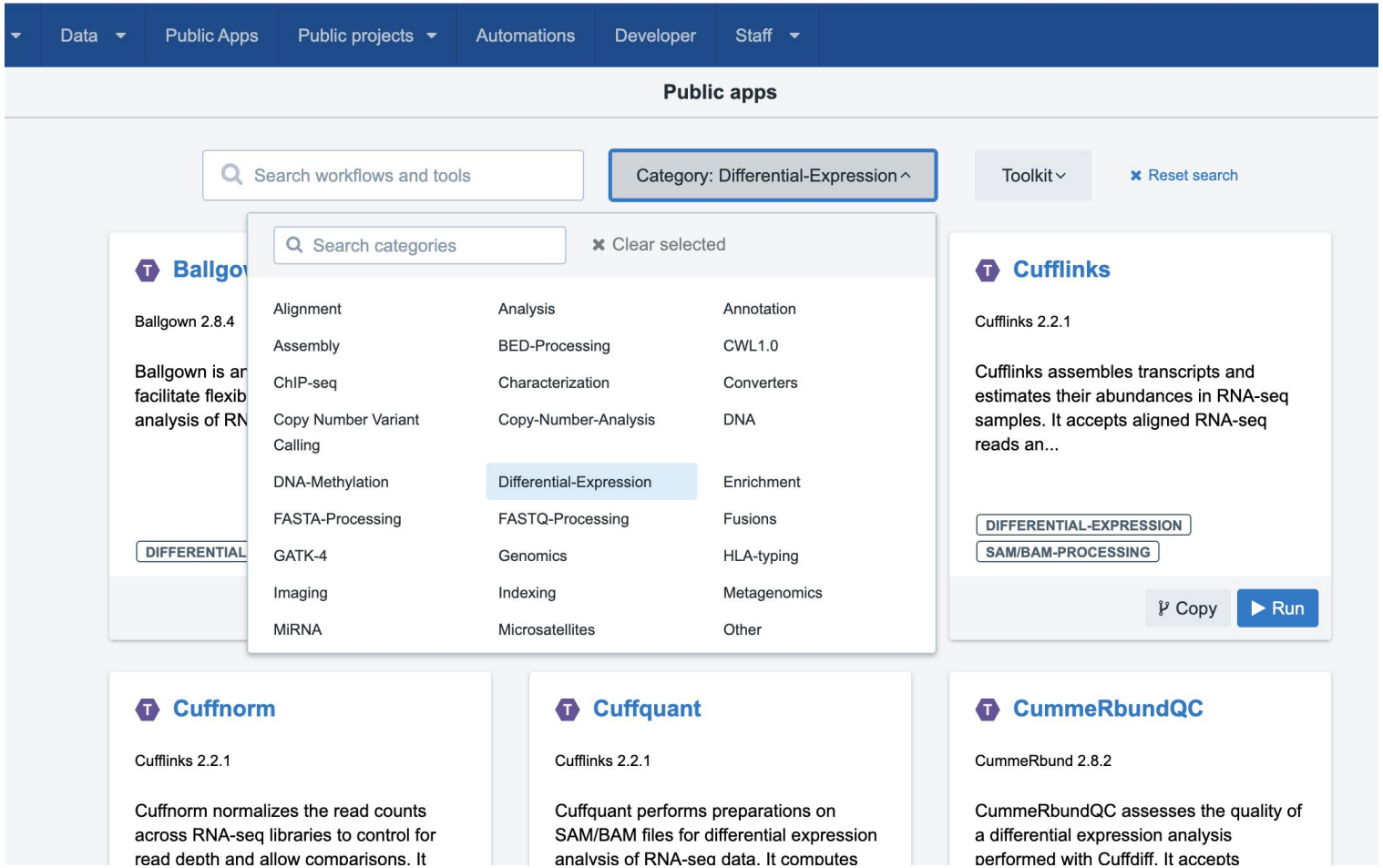
- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

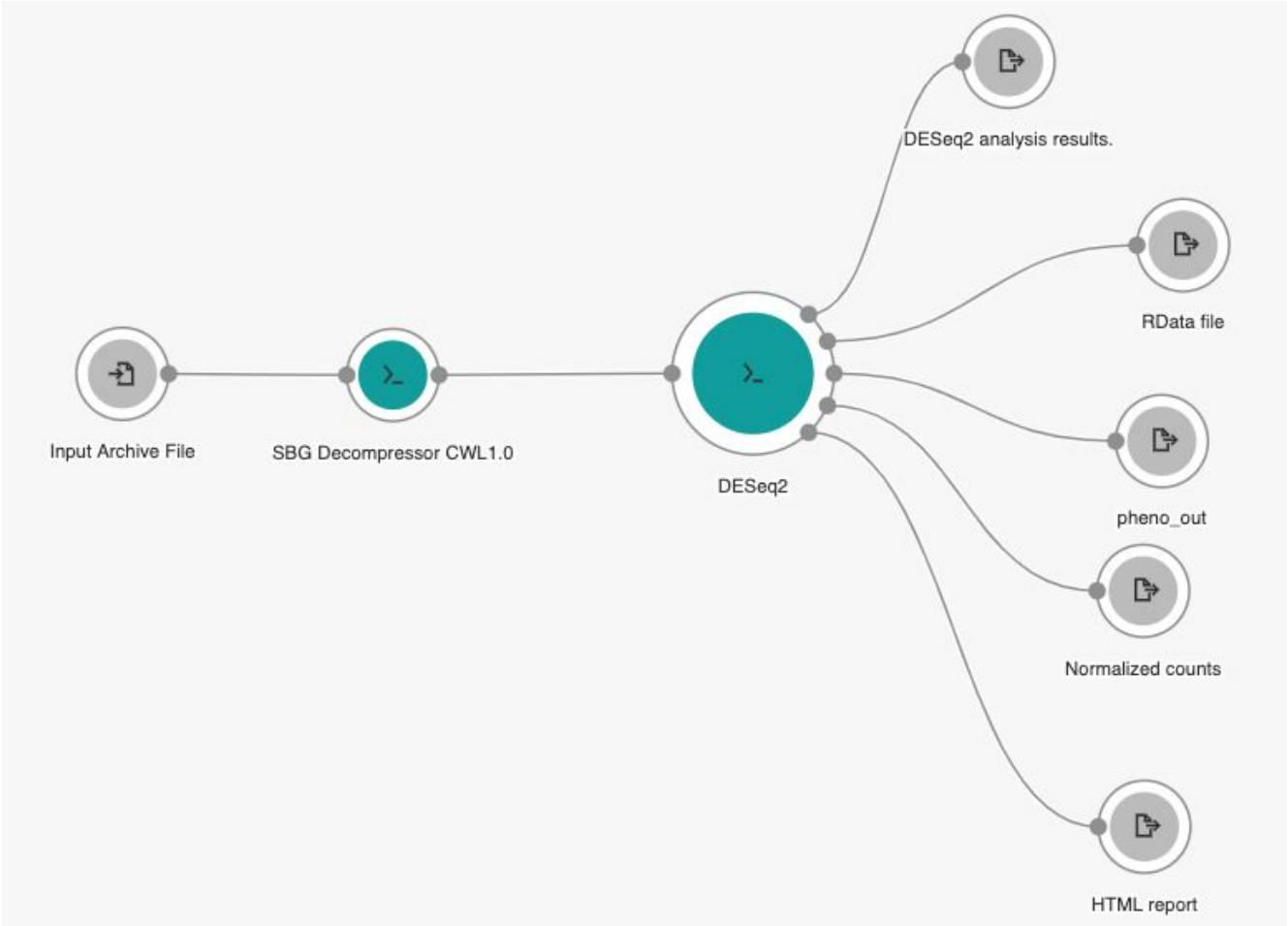
Find the tools you need in the Public Apps Gallery

- A curated collection of **600+** bioinformatics tools & workflows
  - Optimized for speed & cost in the cloud
  - Fully parameterized & customizable
  - Accessible via the GUI & API



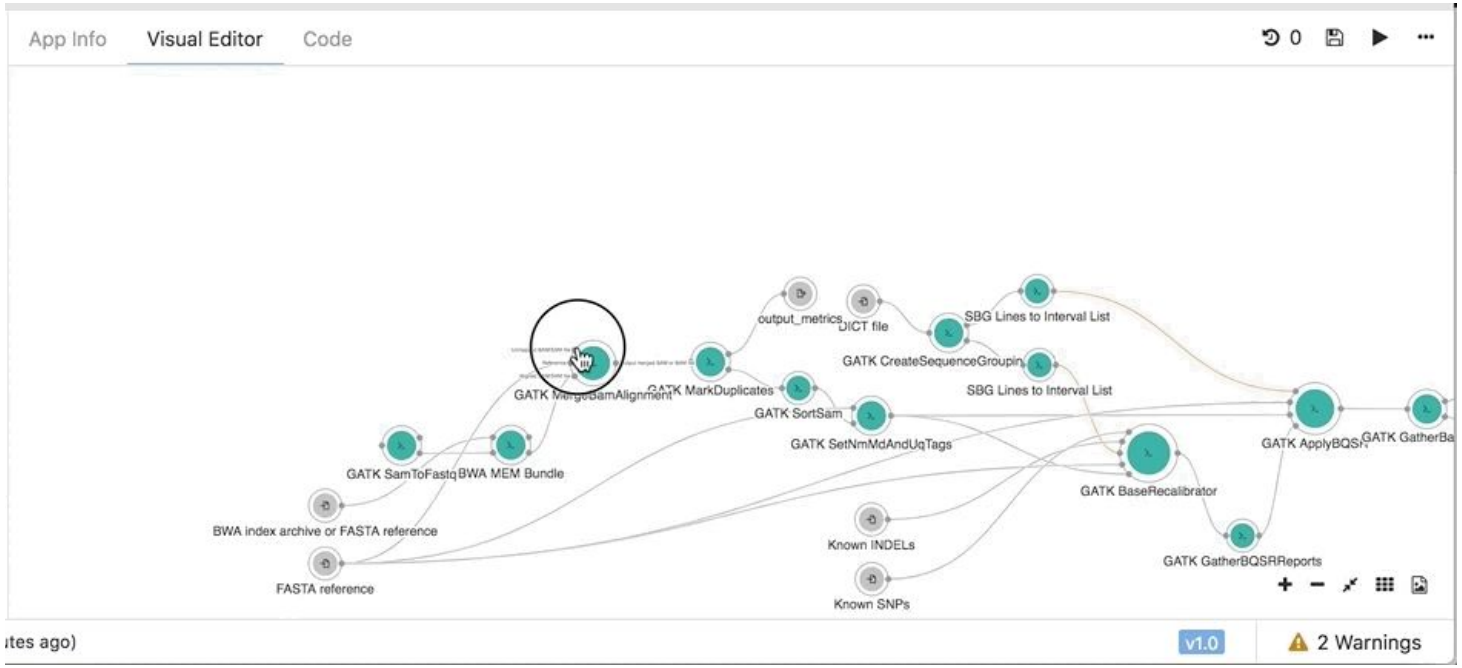
Use Web Composer to tailor a workflow

- Graphical, easy to use interface



Bring your pipelines to the platform with Web Composer

- An intuitive and flexible software development kit for developing and porting custom tools to the platform
- Conformance with community standards to ensure pipeline portability & reproducibility



docker

COMMON WORKFLOW LANGUAGE

Rabix
[Reproducible Analysis for Bioinformatics]

Prototypical User Flow

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

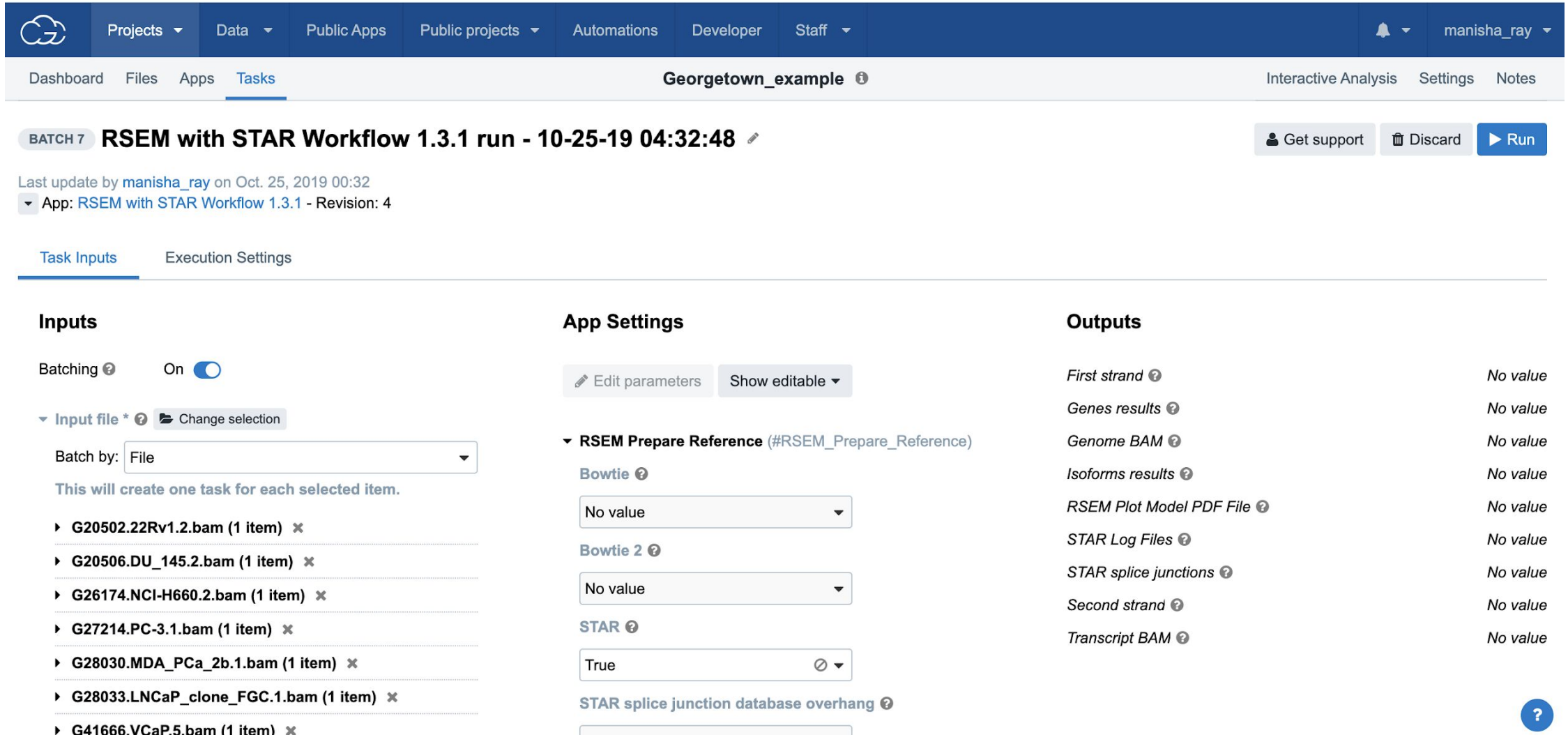
Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

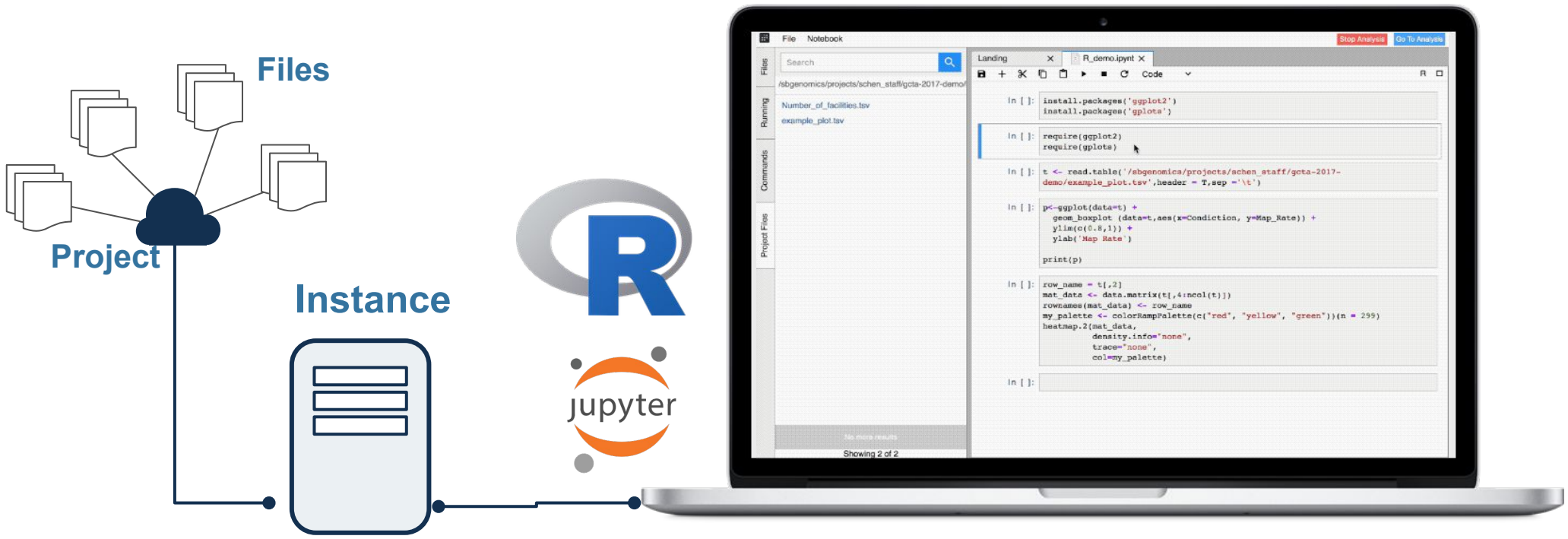
- Task page
- Notebooks in RStudio or JupyterLab

Set up tasks to analyze your data

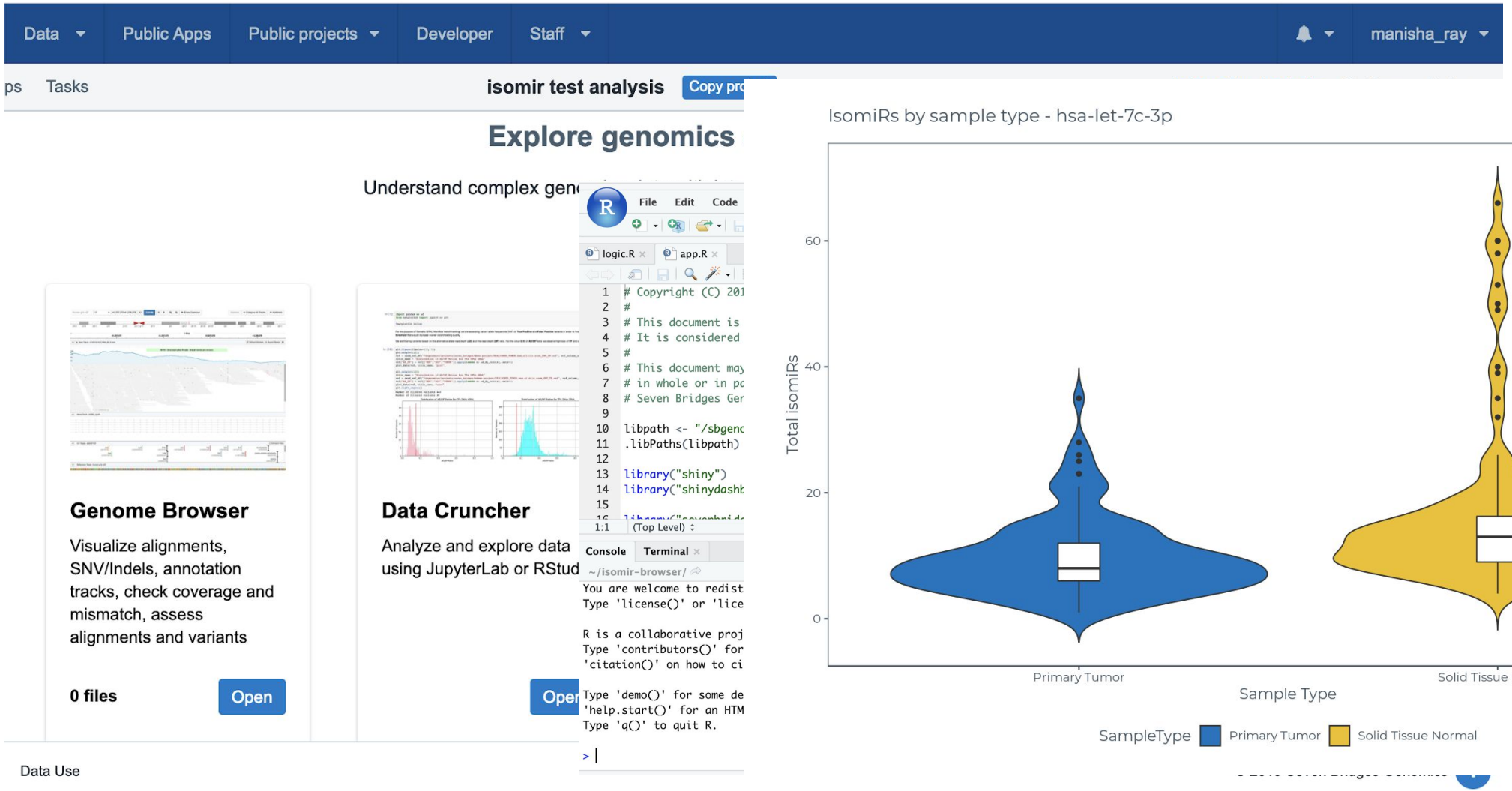


Powerful, collaborative, & reproducible interactive analysis

Users create interactive analysis sessions within a project - all files are available and over 50 instances can be used (*c3.xlarge* to *x1.32xlarge* on AWS)



Quantify and visualize expression differences in RStudio



Using the CGC for RNA sequencing analysis

What is RNAseq?

- A high-throughput sequencing-based method to quantify gene expression in a sample
- All of the RNA transcripts in a sample (transcriptome) are sequenced
    - Sequencing data is aligned to a reference genome or transcriptome
    - Allows for quantification of gene expression and differential expression analyses and characterization of alternative splicing
    - *de novo* transcriptome assembly - no genome sequence necessary!

Why do RNA sequencing?

- Unbiased - collect all of the RNA present in the sample
  - Allows for discovery of new gene targets and pathways
  - In contrast to **targeted** methods where you must choose the genes of interest in advance
- Dynamic - collect all the genes that are made at the time of sample collection
  - Monitor changes over time (ex: disease progression)
  - Monitor changes in response to treatment (ex: drug effects)

RNAseq workflow overview

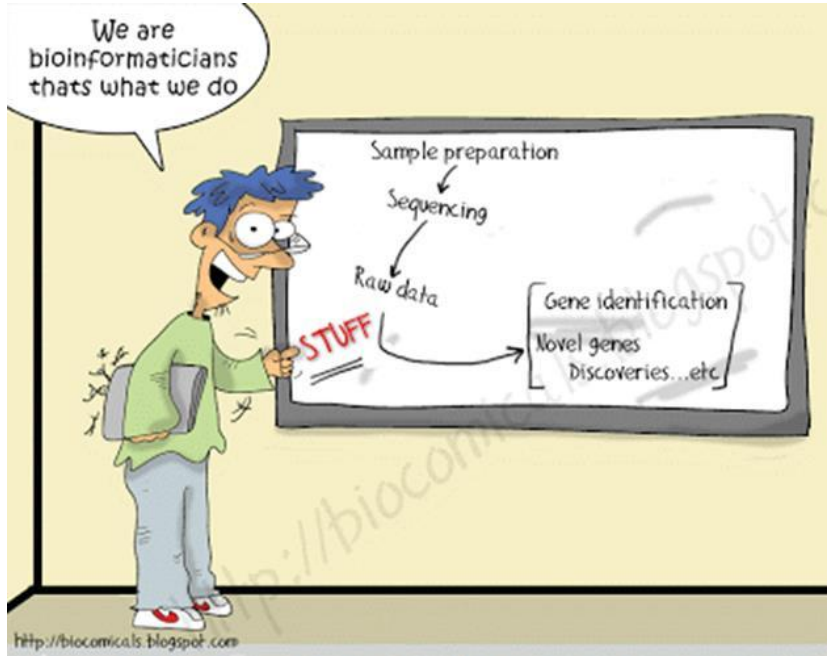
Design Experiment
- Set up the experiment to address your specific biological questions
- Meet with your bioinformatician and sequencing center!

RNA preparation
- Isolate RNA
- Purify RNA

Prepare Libraries
- Convert the RNA to cDNA
- Add sequencing adapters

Sequence
- Sequence the cDNA using a sequencing platform

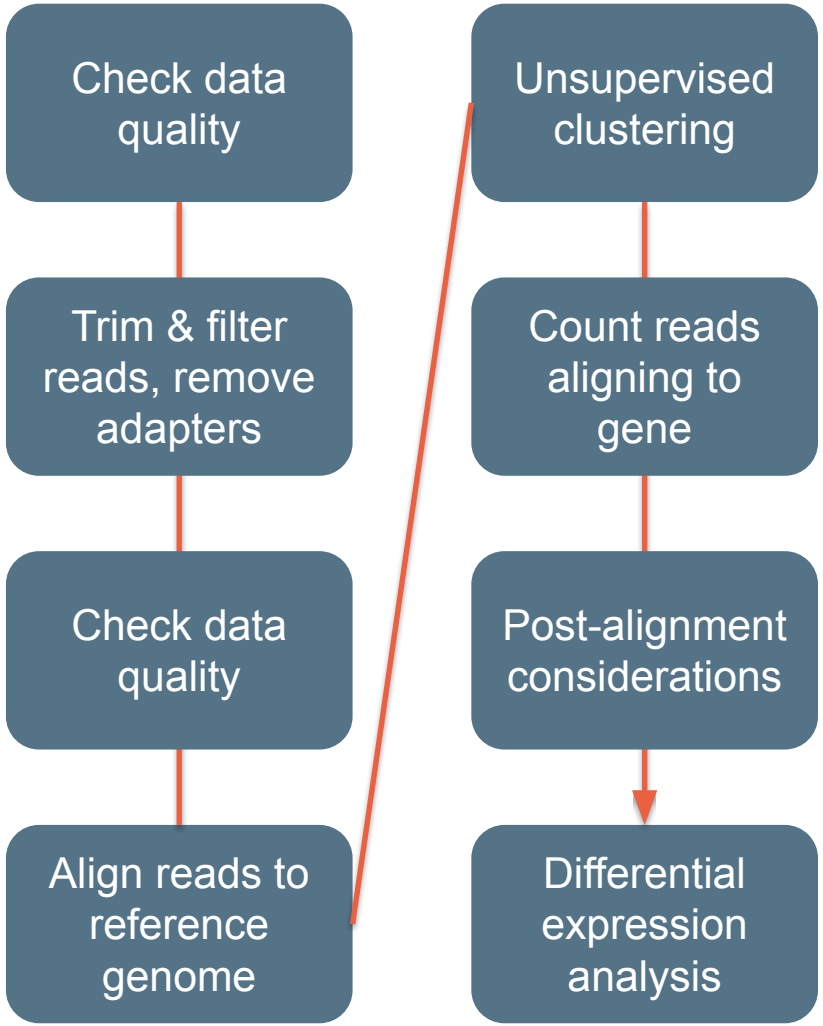
Analysis
- Quality control
- Align reads to the genome/assemble a transcriptome
- Downstream analysis based on your questions



We are here!

Slide adapted from Dr. Nadia Attalah, Purdue University

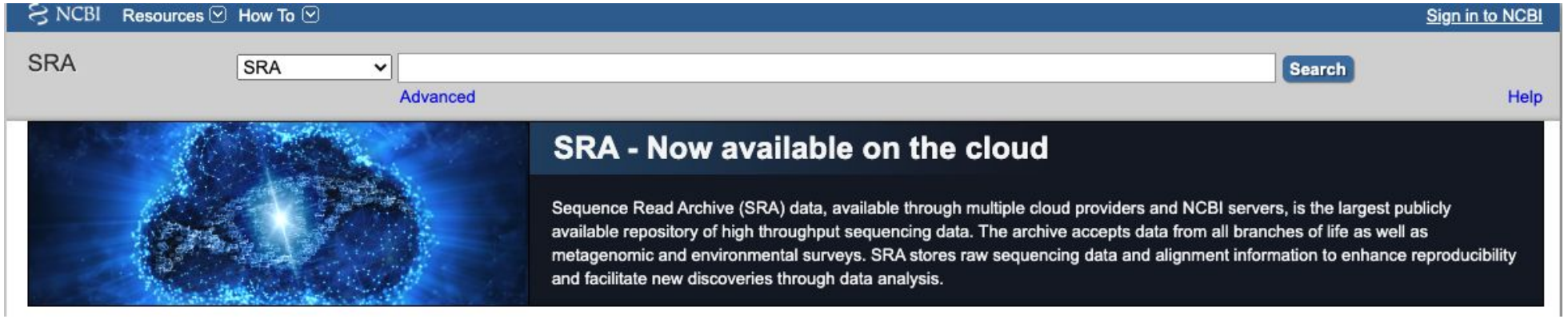
Standard Differential Expression Analysis



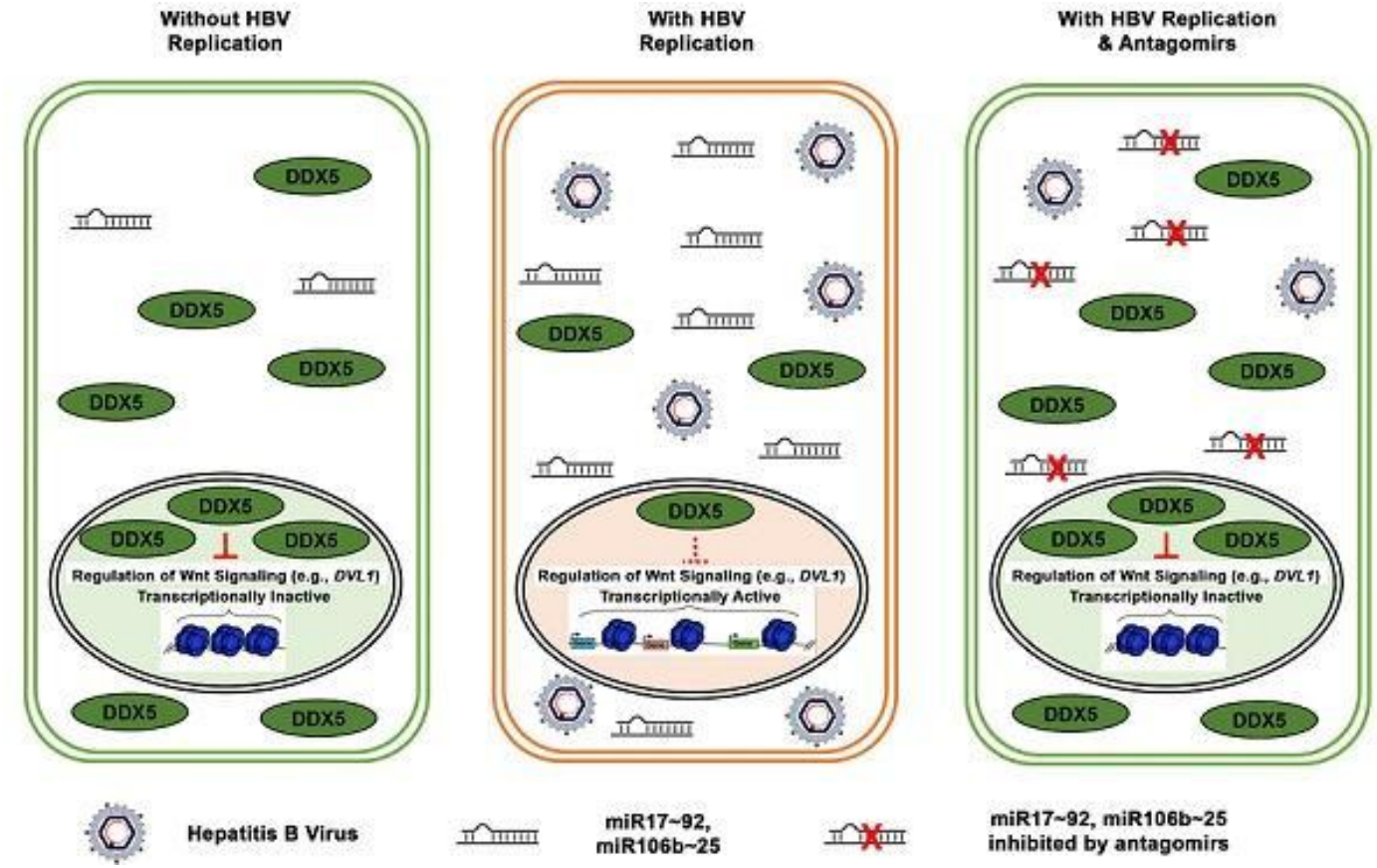
What is SRA?

- SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.
- Search by author, pubmed ID, etc.
- Download respective accession list and numbers, sequence data files using SRA Toolkit, and associated metadata associated with SRA data

https://www.ncbi.nlm.nih.gov/sra



Overview of data we will use for bulk RNA-seq analysis

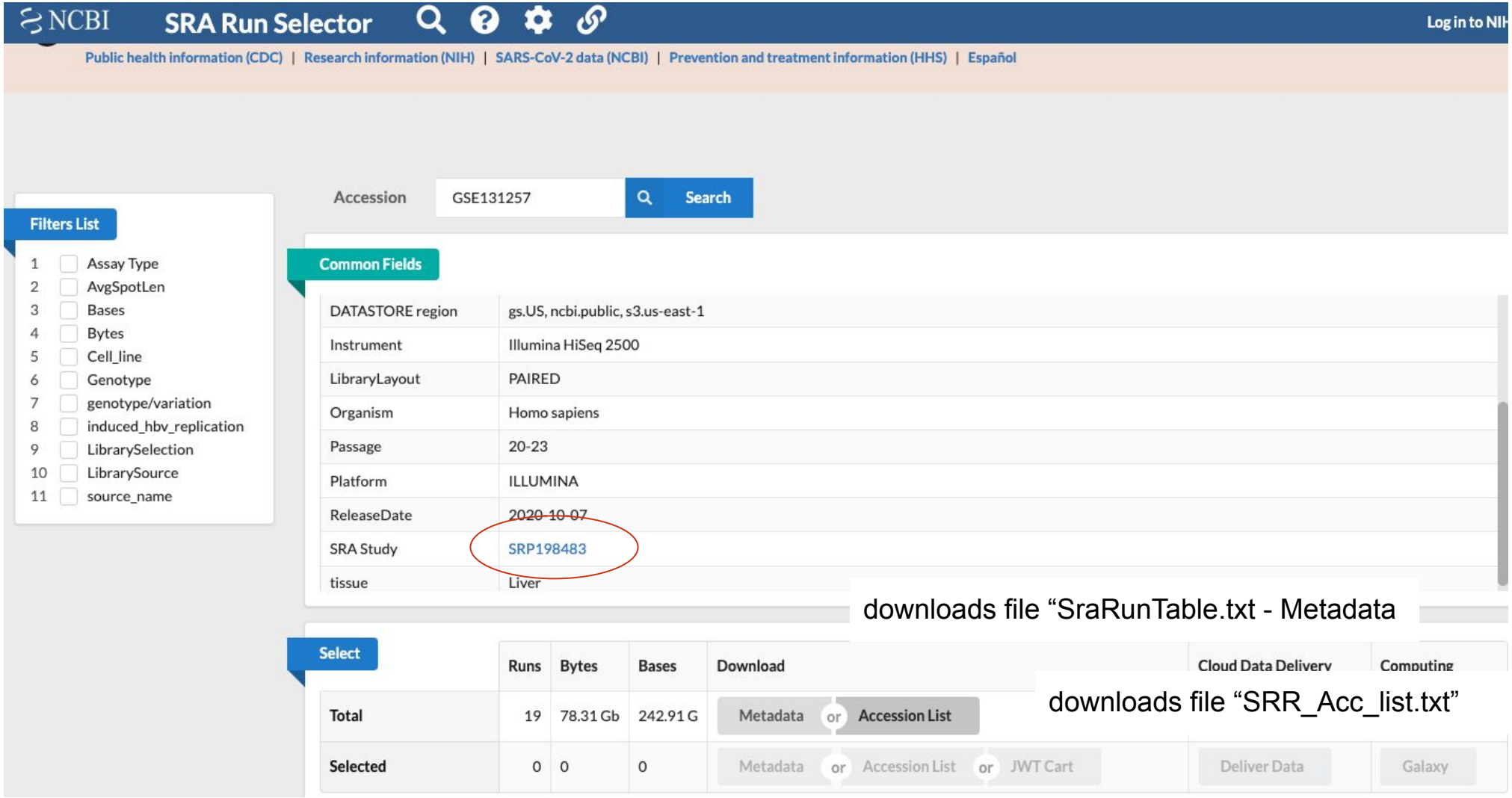


Mani SKK, et al. Restoration of RNA helicase DDX5 suppresses hepatitis B virus (HBV) biosynthesis and Wnt signaling in HBV-related hepatocellular carcinoma. *Theranostics* 2020; 10(24):10957-10972. doi:10.7150/thno.49629. https://www.thno.org/v10p10957.htm

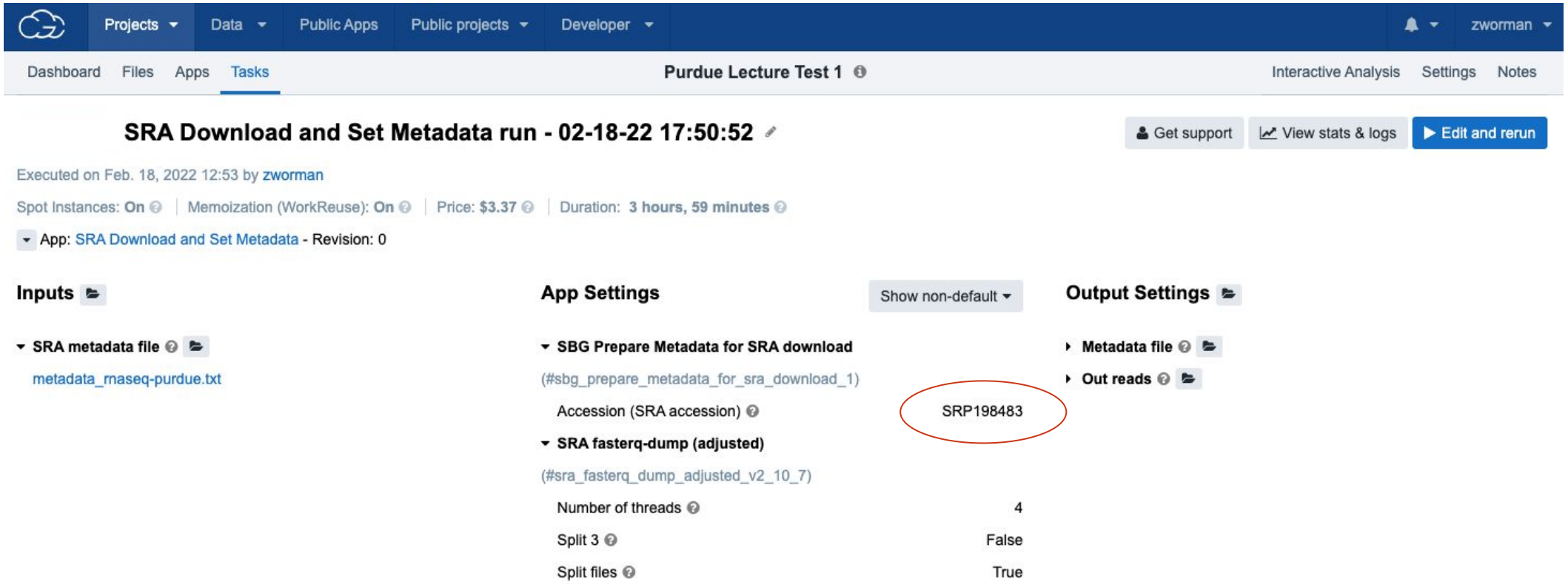
Data availability

All sequencing data are available through the NCBI Gene Expression Omnibus (GEO) database (accession number **GSE131257**).

Search SRA with GEO number!

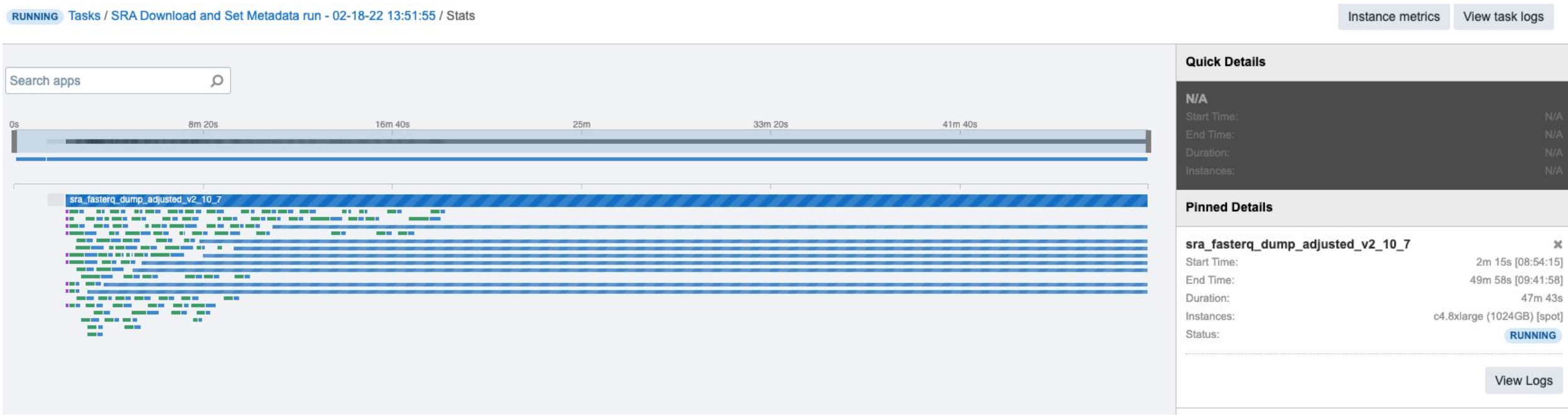


On the CGC, upload metadata, tell the tool the SRA accession…

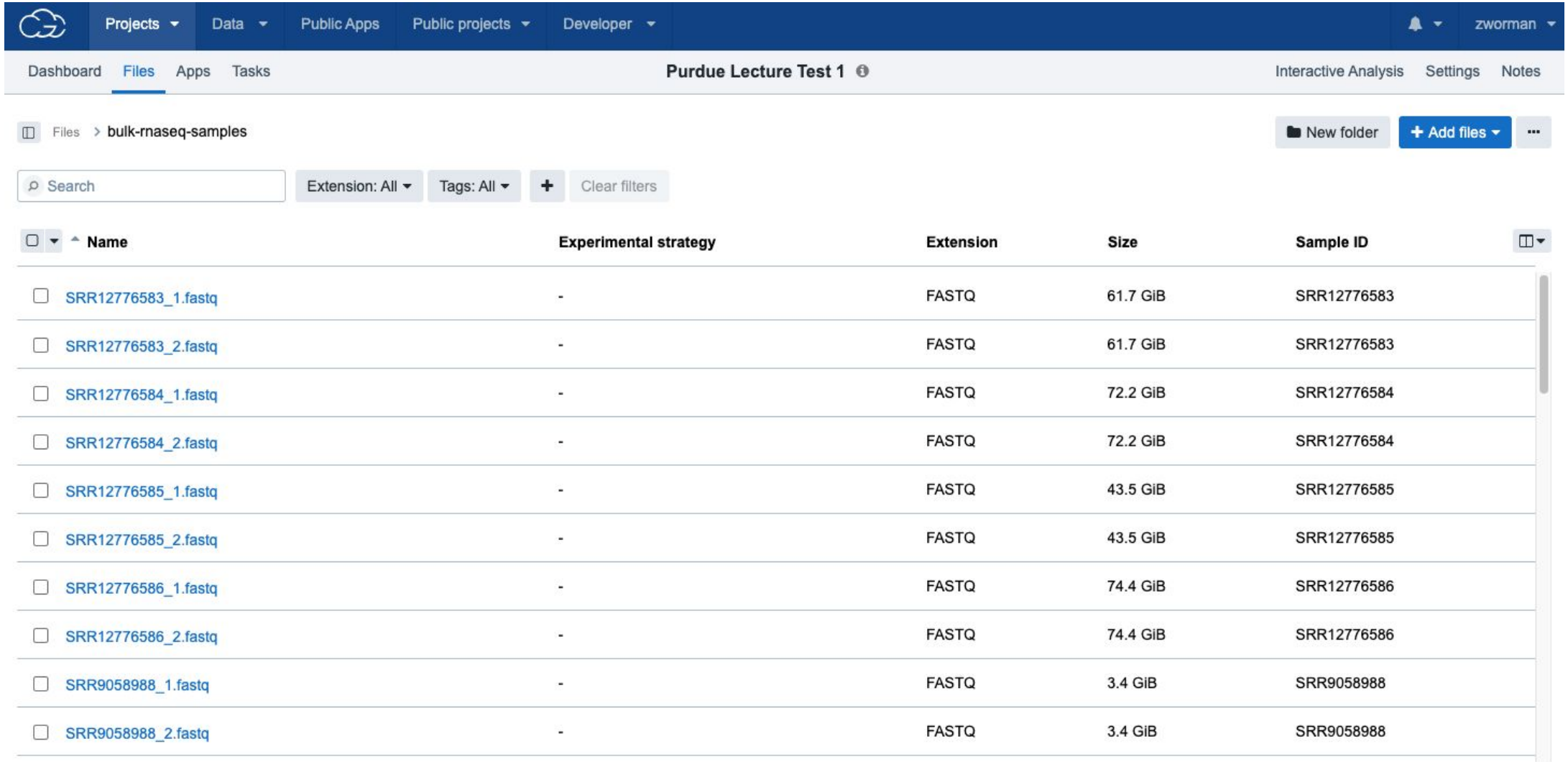


RNAseq data transferred from SRA to CGC

- Total time to transfer 3 hours, 59 minute, $3.37



RNAseq data transferred from SRA to CGC



RNA seq differential expression

Create a Project	Select/access data	Select/create tools	Create and run analysis
Organizational unit within platform	**Upload data** from a collaborator's project	Select **differential expression** tools from Public Apps Gallery Connect into a workflow	Set up task and run workflow Review outputs: **differentially expressed genes**

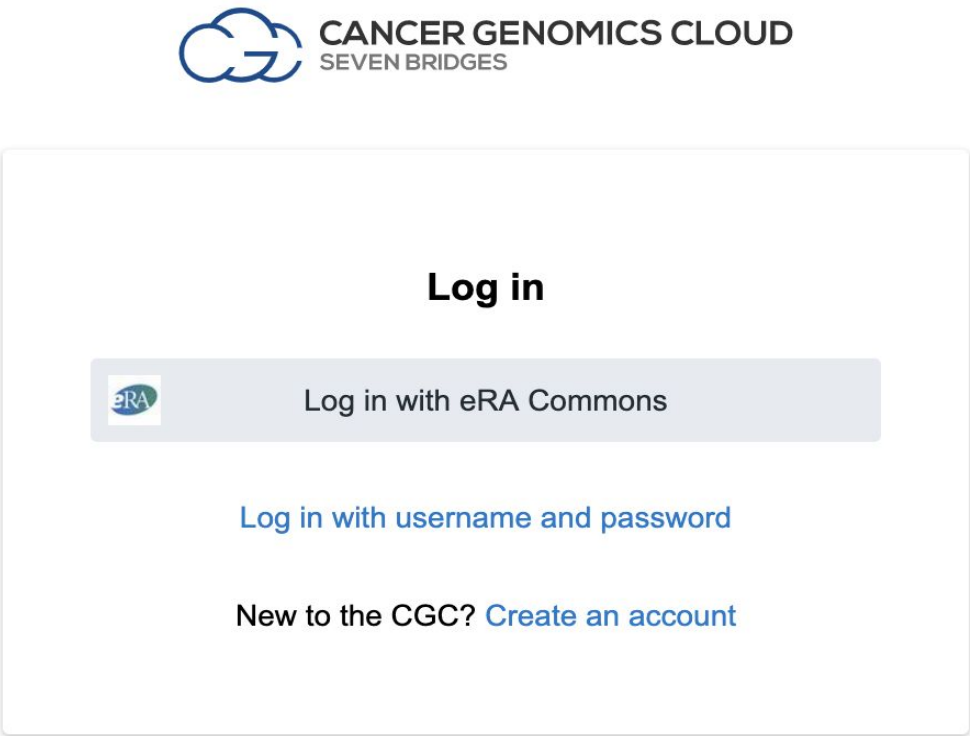
Summary: Run analyses in the cloud with the CGC

- CGC: platform for analysis, part of CRDC
- Collaborate seamlessly
- Access petabytes of public datasets
- Build workflow from new or existing tools
- Run tasks using the flexibility and scale of the cloud
- Connect to other data repositories
- Analyze data interactively using notebooks in the cloud

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Let’s get started!

Login to the platform at
cgc.sbgenomics.com



Q&A and Discussion

CANCER GENOMICS CLOUD
SEVEN BRIDGES