NIH-CFDE Cloud Workspace Partnership Pilot

Documentation Guide

For

Implementation of Bioinformatics Pipelines

For Analysis of CFDE Data

In Collaboration with

Velsera and

Children's Hospital of Philadelphia

Author:

Sangeeta Shukla Ph.D.

Children's Hospital of Philadelphia

Table of Contents

Introduction

The goal with the Kids First CFDE "Cloud Workspace Partnership Pilot" is to understand valuable data, tool and research use cases of other CFDE DCCs and collaboratively pilot integration of data and tool usage in the CAVATICA cloud workspace. At its core, the pilot is focused on demonstrating the value of a collaborative and interoperable cloud workspace for the CFDE and broader Common Fund community that supports integrated CFDE dataset analysis in the cloud and supports cross-DCC use cases that matter to investigators.

During these piloting activities, we aim to provide dedicated support to DCCs and their users, with the goal of not only successfully demonstrating solutions for specific cloud data accessibility using GA4GH DRS and other methods to enable analysis use cases from multiple DCCs, but also improving training resources and documentation to maximize ease of understanding, accessibility and reusability for the Common Fund community in future efforts.

What is CAVATICA?

CAVATICA is a data analysis and sharing platform designed to accelerate discovery in a scalable, cloud-based computing environment where data, results, and workflows are shared among the world's research community. Developed by Seven Bridges and funded in part by a grant from the National Institutes of Health (NIH) Common Fund, CAVATICA is continuously updated with new tools and datasets. Thorough documentation of available platform features is located in the CAVATICA Knowledge Center. There is a Quickstart Guide, which serves to orientate new CAVATICA users to foundational platform aspects and features, including hundreds of public apps and petabytes of public data, including genomic data on pediatric tumors. The CAVATICA platform was developed and maintained by Velsera and based on the Seven Bridges Platform for cloud storage and bioinformatics analysis.

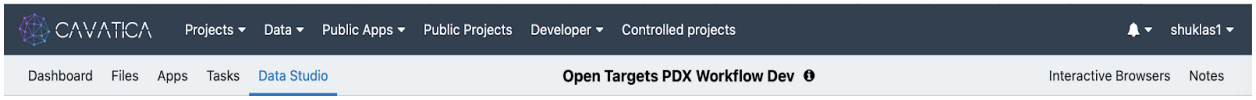
Before you start

Built on Amazon Web Services (AWS) infrastructure the storage and processing of data is presented at a cost to researchers; the costs AWS charges CAVATICA for compute time are the costs users will pay with no surcharge for CAVATICA resources and services. Implementing Spot Instances, an exclusive further economizes research on the platform. In addition, all new CFDE researchers on CAVATICA are eligible for Pilot Funds, a special billing group for new users offering funds intended for user training, exploration, and troubleshooting as they learn to use the platform and its features. Users must first sign up to use the platform, then send a short email with their username requesting Pilot Funds to support@velsera.com.

CAVATICA Quickstart

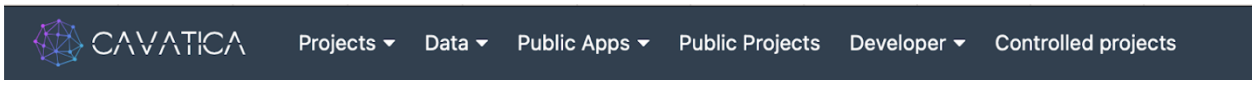
Once you are signed up on the platform and ready with the data that you are looking to process, the first step to running an analysis on CAVATICA is to create a project. A project is a development space where the user does their cloud-based research. Within a project, users can upload data, create new analysis workflows or use existing ones, run an interactive Data Studio for exploratory analysis and visualization, and view results files.

This tool can be found within your project, an example of which you can see in the screenshot below.



Users can choose to upload their data, bring data in the cloud using methods like GA4GH DRS or use publicly available data files from existing CAVATICA data sources and add them to their research project.

Users can access public files/datasets available on the platform by clicking on 'Data' on the platform homepage as you can see in the screenshot below.



Start-up Funds for CFDE Users

In order to promote more widespread use of CFDE datasets across DCCs, as part of the CFDE CWP efforts, Velsera authorizes the allocation of pilot funds for CFDE users. Once a user sets up an account on the CAVATICA platform, they must send an email to support@velsera.com with details of their association with DCC or CFDE project to which they intend to contribute. Velsera Support staff will add the CAVATICA user's project to a specific billing group. Computational resources allocated for processing the user's workflows can then use these funds.

Public Data on CAVATICA

CAVATICA hosts several datasets comprising a wide range of study interests. The Kids First database is an extensive collection of pediatric data, alongside studies such as TARGET and TCGA. These data, already on the cloud, can be accessed and added to user projects while the files remain hosted, and storage paid for, by NIH; users are only charged for compute time and storage of generated downstream analysis and results files.

Abiding by the FAIR principles, CAVATICA is also seamlessly interoperable with Velsera's sister platforms, the Cancer Genomics Cloud and BioData Catalyst Powered by Seven Bridges. Data, Apps, Results and Projects can be securely accessed and shared across platforms, promoting interdisciplinary research and collaboration.

Data Upload Methods

CAVATICA offers variety of methods to allow data upload from local, cloud, and server storage. A brief description can be found below, and a comprehensive document is available on the platform here.

- Upload from local storage by browsing and selecting directly through CAVATICA's visual interface
- Upload using the Command-Line (CLI) Uploader from your local machine or cluster when the data volume is large
- Upload via CAVATICA API as it offers more direct control over uploads
- Import from cloud storage such as AWS S3 or Google Cloud Storage without transferring it to CAVATICA storage using the Connect Cloud Storage feature
- Upload from an HTTP(S)/FTP server endpoint using the HTTP(S)/FTP upload option

File Repositories

Data files on the platform can be stored in two types of file repositories.

- Project Files – This repository Is located within the project and is specific to every project. It contains the input and output files for workflows in that project. Users can upload directly to a project or copy them from other projects and repositories.
- Public Files – This repository is maintained by the Bioinformatics team at Velsera. It contains the latest and most frequently used reference genomes and annotation files so users won't have to upload reference files every time to run a task.

CAVATICA for Bioinformatics Research

In bioinformatics research, an analytical pipeline is essentially defined as a series of software algorithms that process raw sequencing data and generate interpretations that can potentially advance the overall understanding of the biological process and its key players. A bioinformatics analysis pipeline consists of three basic steps: preprocessing of sequencing data, discovery of variants, and integrative analysis of variants/related genes.
To enable and encourage more users to take advantage of cloud-based infrastructure to store, process and analyze bioinformatics data, the Gabriella Miller Kids First Data Resource Center (Kids First DRC) and the NIH Common Fund Data Ecosystem (CFDE) have joined hands as part of the CFDE Cloud Workspace Partnership (CWP) Pilot. The goal of the CFDE CWP Pilot is to understand valuable data, tool and research use cases and collaboratively pilot integration of data and tool usage in the CAVATICA cloud workspace.

Public Apps on CAVATICA

The KF DRC in collaboration with the Center for Data Driven Discovery of Biomedicine at Children's Hospital of Philadelphia have built and deployed various bioinformatics workflows wrapped in tools on the CAVATICA platform that are available for use to the research community.

Below are some common Apps bioinformatics researchers may use for different categories of analyses. Each of these Apps has version control and can be copied into user Projects and edited for user-specific needs. New Public Apps are constantly under development, and users can build and deploy their applications.

Preprocessing: NGS Checkmate Sample QC, NGS Checkmate Preprocess
Alignment and variant calling: Alignment and GATKHaplotypeCaller Workflows, GATK HaplotypeCaller CRAM to gVCF Workflow, Germline SV Workflow, Germline Variant Workflow, Joint Genotyping Workflow, Pathogenicity Preprocessing Workflow
RNAseq analysis: HuBMAP scRNA-seq pipeline

List of available Public Apps

***Application Name*:** NGS Checkmate Preprocess
***Publisher*:** KFDRC
Contributors:

Goal/Purpose: preprocessing workflow to use bcftools to subset bams and create a bcftools-called vcf
Input File(s): BAM file, subset character list, reference fasta file, SNP_bed file
Output File(s): VCF file
Comments:
Highlights:
Gaps:
Relevant Links: https://github.com/kids-first/ngs_checkmate_wf, https://cavatica.sbgenomics.com/u/kfdrc-harmonization/kf-references

Application Name: NGS Checkmate Sample QC
Publisher: KFDRC
Contributors: brownm28
Goal/Purpose: A software pipeline for validating sample identity in NGS studies within and across data types
Input File(s): FASTQ, BAM or VCF
Output File(s): (i) a list of matched sample pairs with genotype correlation coefficients; (ii) a sample clustering dendrogram; and (iii) a graphical representation of sample clustering that can be entered into graphical visualization tools such as Cytoscape
Comments:
Highlights:
Gaps:
Relevant Links: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5499645/, https://github.com/kids-first/ngs_checkmate_wf

Application Name: Alignment and GATKHaplotypeCaller Workflows
Publisher: KFDRC
Contributors: nathanj, danmiller, brownm28, sicklera
Goal/Purpose: Using BWA, align input file(s) with reference genome version hg38, to generate a resulting BAM file. Optionally, also calculate contamination via gVCF. Additionally, workflow is also capable of performing a basic evaluation of the X and Y sex chromosomes using idxstats.
Input File(s): SAMs/BAMs/CRAMs (Alignment/Map files, or AMs), PE reads, and/or SE reads;
conditionally generate gVCF and metrics.

Output File(s): BAM file
Comments: Duplicates are flagged in a process that is connected to bwa mem. This design decision implies that duplicates are flagged only on the inputs of that are scattered into bwa. Duplicates, therefore, are not being flagged at a library level and, for large BAM and FASTQ inputs, duplicates are only being detected within a portion of the read group.
Highlights:
Gaps:
Relevant Links:

Application Name: GATK HaplotypeCaller CRAM to gVCF Workflow
Publisher: KFDRC
Contributors: danmiller, brownm28, sickera
Goal/Purpose: GATK Convert a CRAM file into a BAM file, determine contamination value, then run GATK HaplotypeCaller to generate a gVCF, gVCF calling metrics, and if no contamination value is provided, the VerifyBAMID output
Input File(s): input_cram, reference_tar, dbsnp_vcf, dbsnp_idx, contamination_sites_bed, contamination_sites_mu, contamination_sites_ud, wgs_calling_interval_list, wgs_ecaluation_interval_list
Output File(s): gvcf, gvcf_calling_metrics, verifybamid_output
Comments:
Highlights:
Gaps:
Relevant Links:
https://github.com/kids-first/kf-alignment-workflow/releases/tag/v2.8.2

Application Name: Germline SV Workflow
Publisher: KFDRC
Contributors: danmiller
Goal/Purpose: Generate SV calls from an aligned reads BAM or CRAM file, using Manta or SvABA to call variants, then annotate the variants using AnnotSV
Input File(s): KFDRC germline_reads(BAM/CRAM), indexed_reference_fasta, annotsc_annotations_dir, annotsc_genome_build, output_basename
Output File(s): KFDRC Structural variants and Small INDELS called by Manta (manta_svs_manta_indels), Structural variants and Small INDELS called by SvABA (svaba_svs, svaba_indels), Annotation results from AnnotSV

(manta_annotated_svs, manta_unannotated_svs, svaba_annotated_svs, avaba_unannotated_svs)
Comments:
Highlights:
Gaps:
Relevant Links:
https://github.com/kids-first/kf-germline-workflow/releases/tag/v0.3.0, https://cavatica.sbgenomics.com/u/kfdrc-harmonization/kf-references/, https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/, KFDRC AWS s3 bucket: s3://kids-first-seq-data/broad-references/

Application Name: HuBMAP scRNA-seq pipeline
Publisher: cavatica
Contributors: cavatica
Goal/Purpose: The HuBMAP scRNA-seq pipeline is built on Salmon, Scanpy, and scVelo, and is implemented as a CWL workflow wrapping command-line tools encapsulated in Docker containers.
Input File(s): fastq_dir
Output File(s): Salmon output, count matrices, scanpu QC results, dispersion plot, umap plot, umap density plot, scvelo annotated matrices
Comments: The app itself shows no documentation. It would be a good idea to add some lines on CAVATICA app instead of having to navigate to GitHub.
Highlights:
Gaps:
Relevant Links: https://github.com/hubmapconsortium/salmon-rnaseq

Application Name: Germline Variant Workflow
Publisher: KFDRC
Contributors: danmiller
Goal/Purpose: generate variant calls from an aligned reads BAM or CRAM file. using copy number, single nucleotide, and structural variant calling software to call variants. Annotation is performed on the single nucleotide and structural variants.
Input File(s): Long list of desired input parameters, See app references linked below

Output File(s): Long list of expected input parameters, See app references linked below
Comments: Extra markdown code seen on the app, which can be removed.
Highlights:
Gaps:
Relevant Links: s3://kids-first-seq-data/broad-references/, https://cavatica.sbgenomics.com/u/kfdrc-harmonization/kf-references/, https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/

Application Name: Joint Genotyping Workflow
Publisher: KFDRC
Contributors: danmiller
Goal/Purpose: Cohort sample variant calling and genotype refinement
Input File(s): Long list of desired input parameters, See app references linked below
Output File(s): Long list of expected input parameters, See app references linked below
Comments:
Highlights:
Gaps:
Relevant Links:
https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/, s3://kids-first-seq-data/broad-references/, https://cavatica.sbgenomics.com/u/kfdrc-harmonization/kf-references/, https://github.com/d3b-center/bixtools

Application Name: Pathogenicity Preprocessing Workflow
Publisher: KFDRC
Contributors: brownm28
Goal/Purpose: This tool performs an automatic classification for PVS1 interpretation of null variants
Input File(s): Note - first run the Kids First Germline Annotation Workflow first.
Vep_vcf, annovar_db, intervar_db, autpvs1_db
Please refer to the app for to understand individual input files.

Output File(s): intervar_classification, autopvs1_tsv, annovar_vcfoutput, annovar_txt
Comments: Refer to the app for links to find additional documentation for InterVar Classification workflow and AutoPVS1 for pathogenicity scoring.
Highlights:
Gaps:
Relevant Links: https://github.com/d3b-center/D3b-Pathogenicity-Preprocessing

CFDE Portal - CAVATICA Compatibility

The goal of this CFDE-CWP Pilot is to encourage and enable CFDE users to access and use NIH-CFDE data to its full potential and offer advanced computing capabilities to allow bioinformatics researchers without having to go through the unnecessary effort of creating custom scripts for different stages of their analytical pipelines, especially when parts of the overall algorithm are the same and only the data file(s) differ(s).

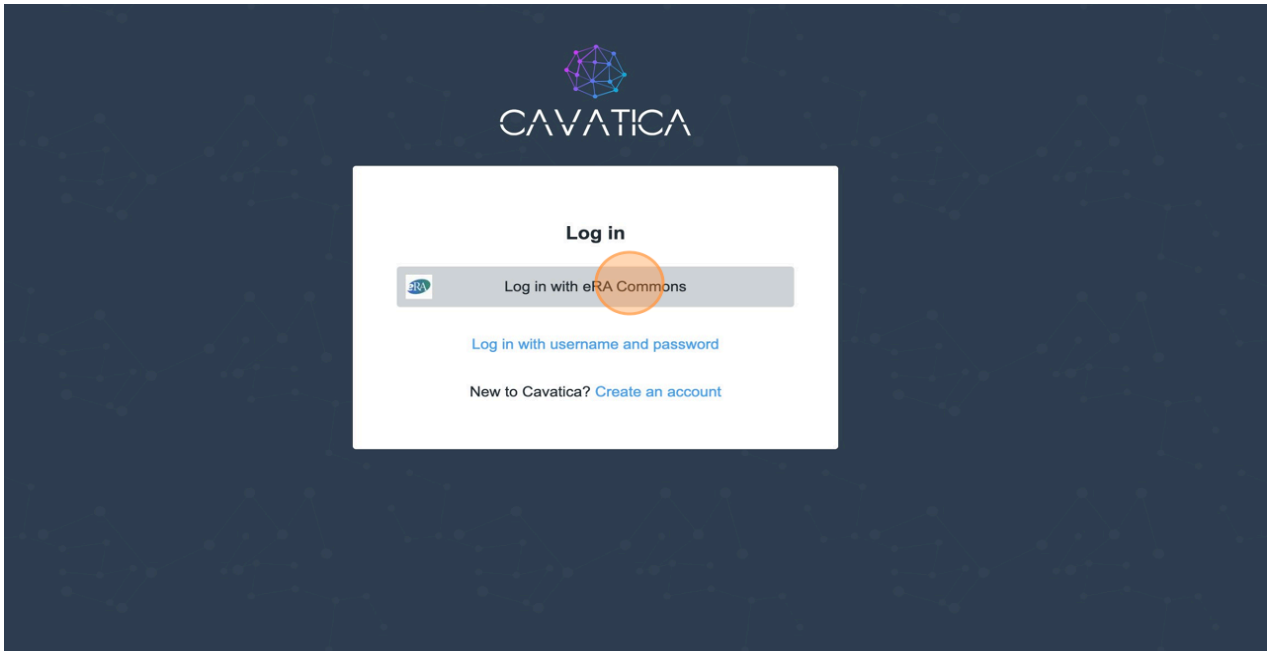
To this end, CAVATICA platform offers additional compatibility to import data files directly from the CFDE Data portal using Persistent IDs. A section of this document will discuss this method for import. In case of datasets that do not have a Persistent ID starting with 'drs', there is an ongoing effort to engineer such IDs.

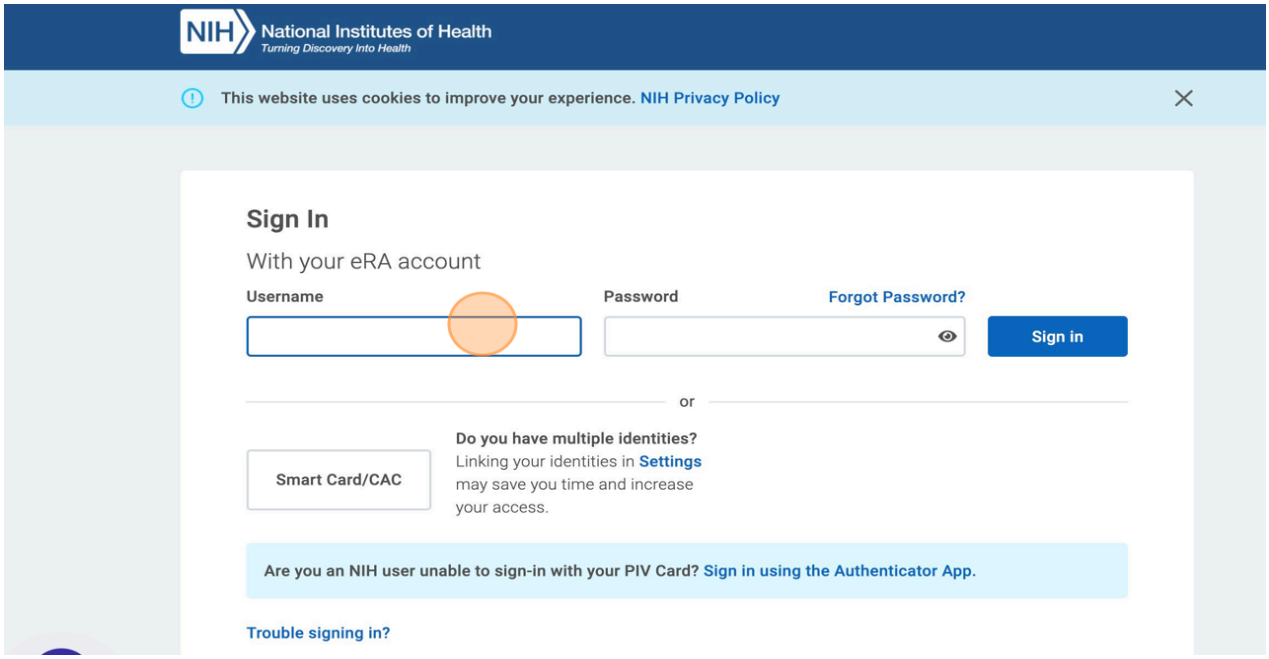
How-To Guide

This detailed step-by-step guide on how users can access data from the CFDE data portal, bring the files over to the CAVATICA platform and implement a bioinformatics analysis workflow is designed with the intent to get you started quickly. Detailed instructions for both the CFDE Portal and CAVATICA are available, along with user support at support@cfde.atlassian.net and support@velsera.com.

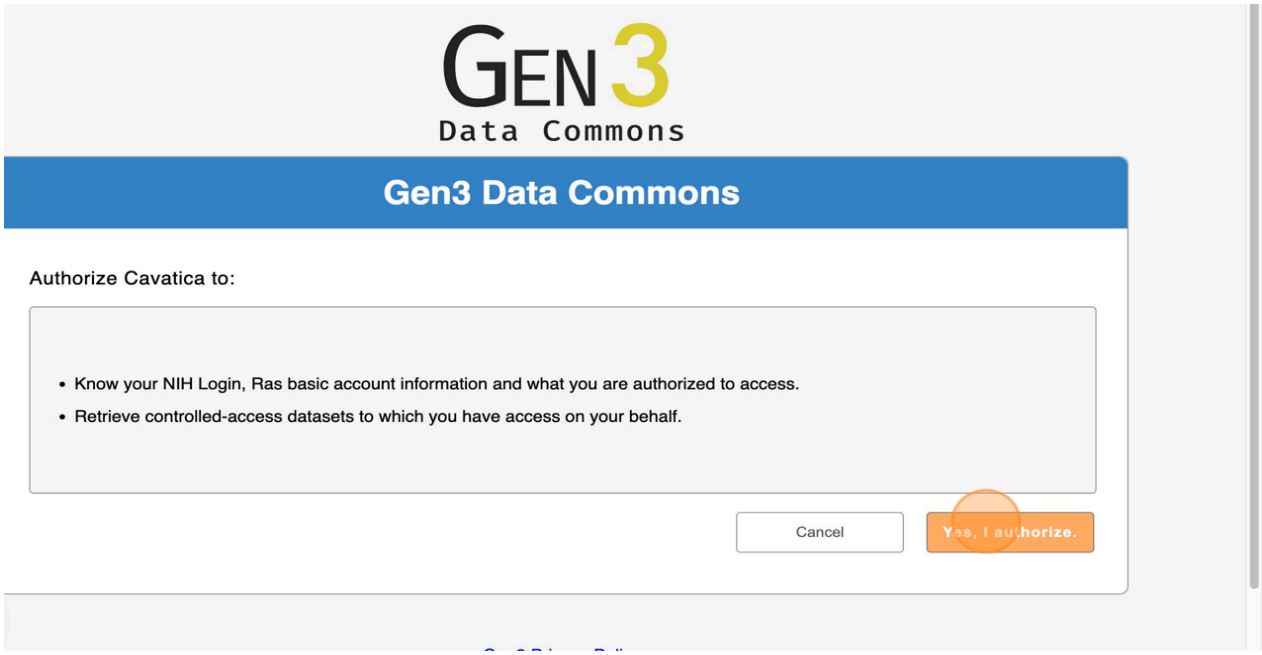
CAVATICA Access and Login

1) Navigate to cavatica.sbgenomics.com and log in using your eRA Commons ID. Linking your CAVATICA account to eRA commons allows access to numerous public datasets and eases access to controlled data to which a user is granted access.





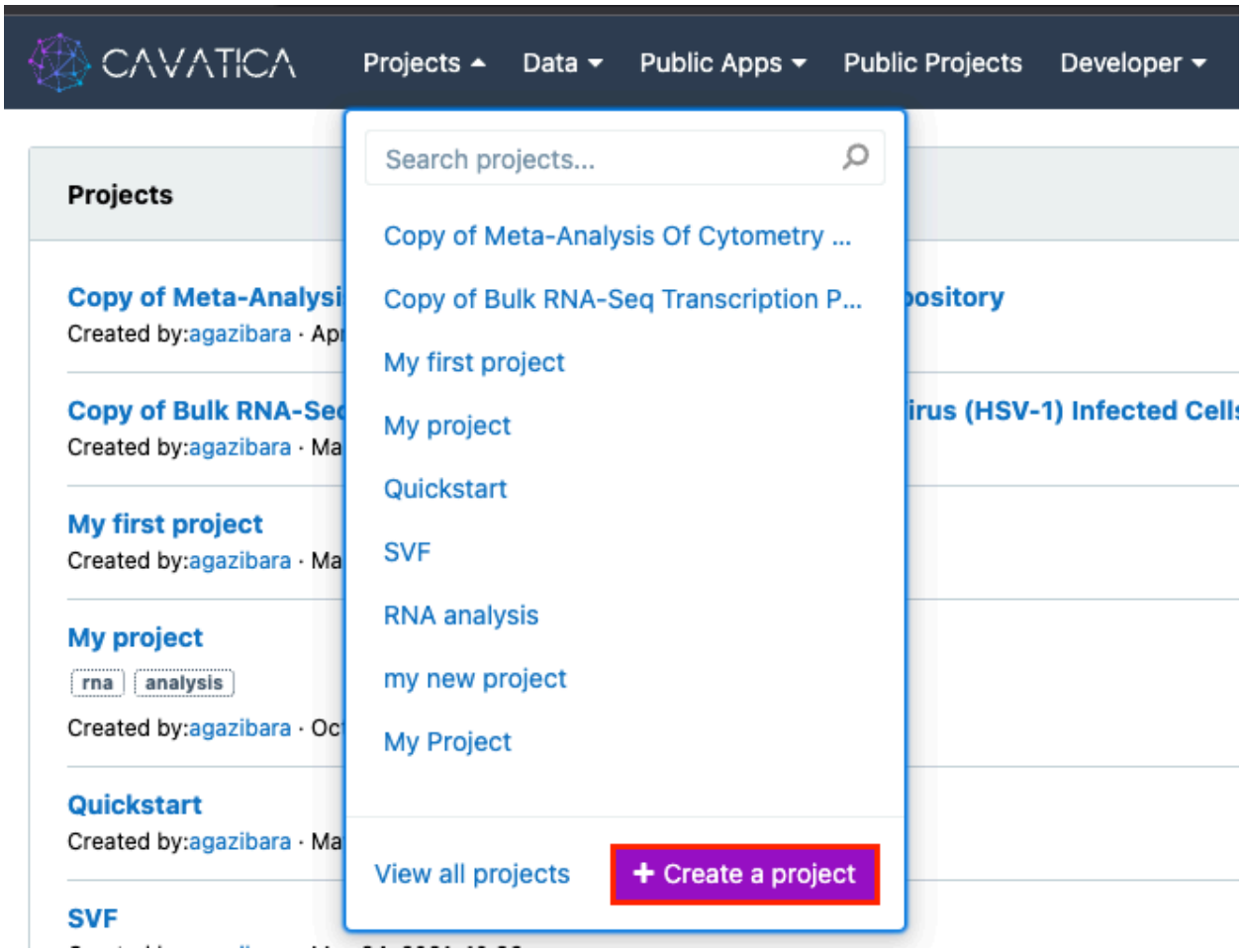
Logging in will require an additional authorization step with Gen3 Data Commons.

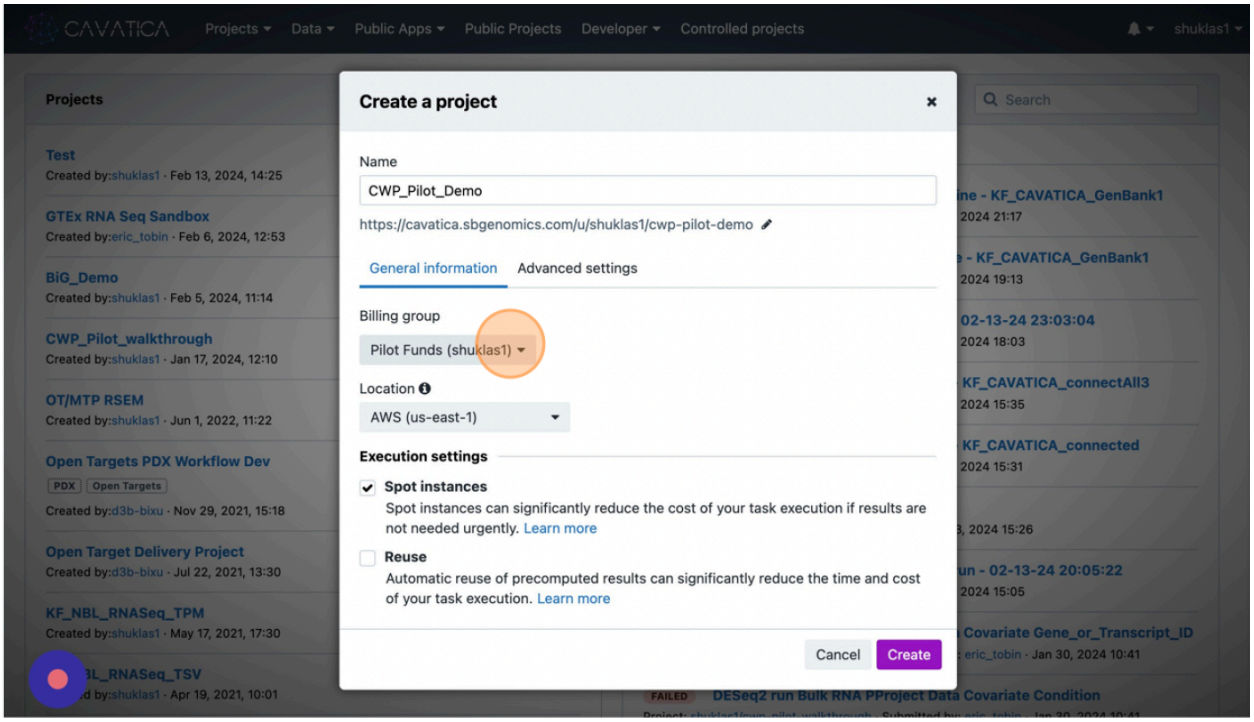


Create a Project

The CAVATICA dashboard is the landing page for your research on the platform. You will see two sections, Projects and Analysis/Data Studio. For this tutorial, we will focus on the Projects section and show you how to create a new project to explore and then house the CFDE data you wish to export to the platform for analysis.

Click on the "Project"' menu and then the magenta "Create Project" button.


CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Projects
Search projects...
Copy of Meta-Analysis Of Cytometry ...
Copy of Bulk RNA-Seq Transcription P...
My first project
My project
Quickstart
SVF
RNA analysis
my new project
My Project
View all projects
+ Create a project
Copy of Meta-Analysi
Created by:agazibara · Ap
Copy of Bulk RNA-Sec
Created by:agazibara · Ma
My first project
Created by:agazibara · Ma
My project
rna
analysis
Created by:agazibara · Oc
Quickstart
Created by:agazibara · Ma
SVF
pository
irus (HSV-1) Infected Cell



Also, be sure to review the Billing group with enough funds to be able to run the analysis. New academic researchers to the platform can apply for [Pilot funds]() which serve as introductory credits to explore the platform and train new cloud computing skills. Email [suppot@sbgenomics.com]() with your platform username and a request for funds.

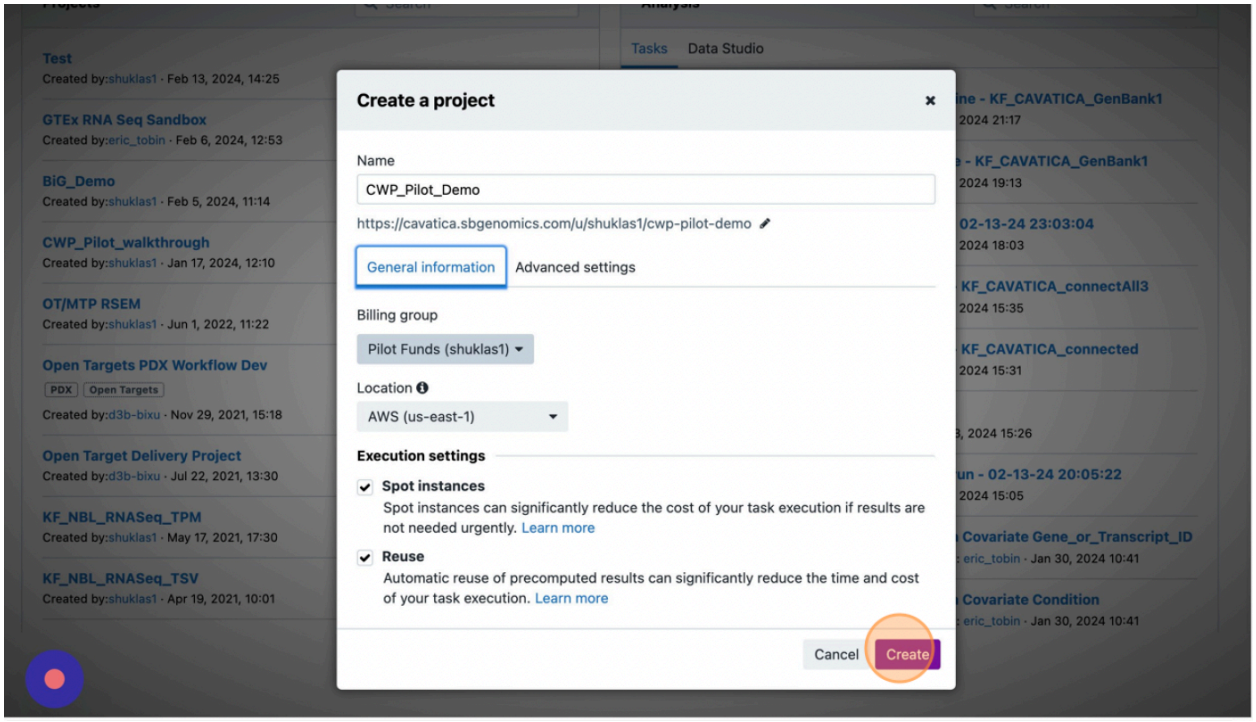
Name the project for your analysis, set your billing group, decide on spot instances and work reuse, and under “Advanced Setting” make sure to “Enable Network Access” for the project. All but the URL for the project are able to be modified later.

General information Advanced settings

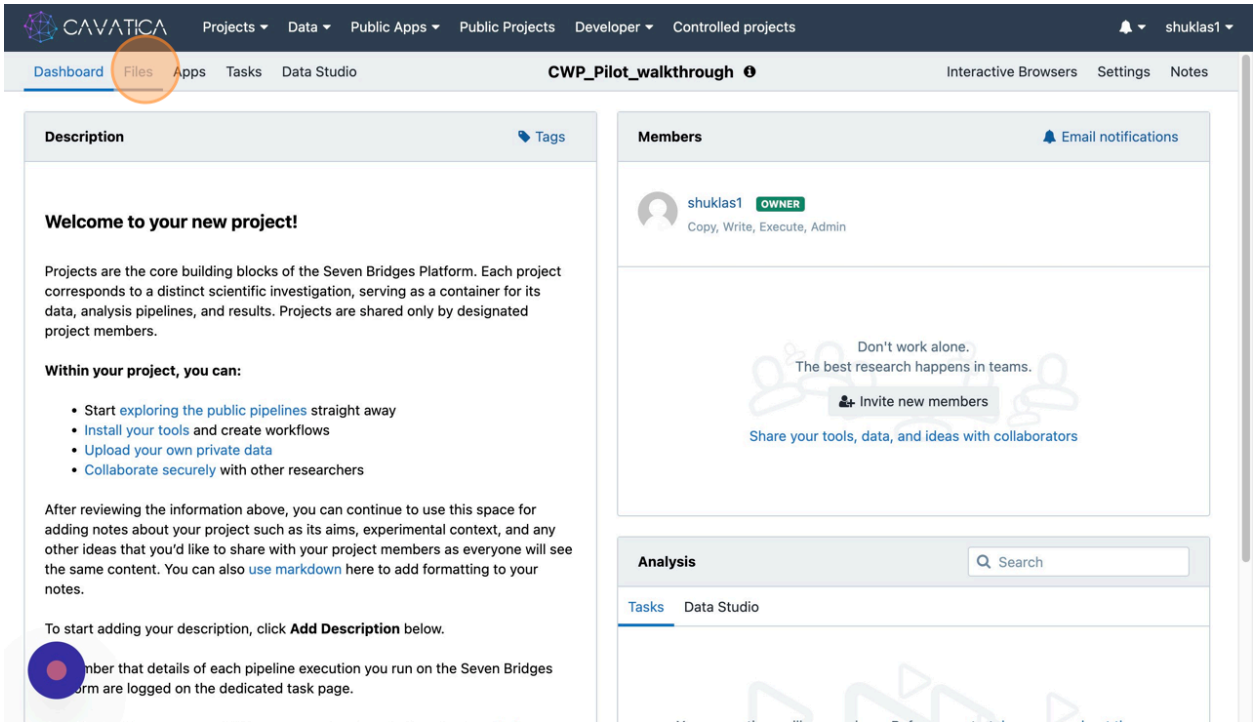
Network Access settings

- ○ **Block network access**
  Execution within the project won't have network access.
- ◉ **Allow network access**
  Execution will have unrestricted network access.

Cancel Create



When all the necessary fields are populated, and options toggled, click the “Create” button to finish setting up your project and be taken to the project dashboard.

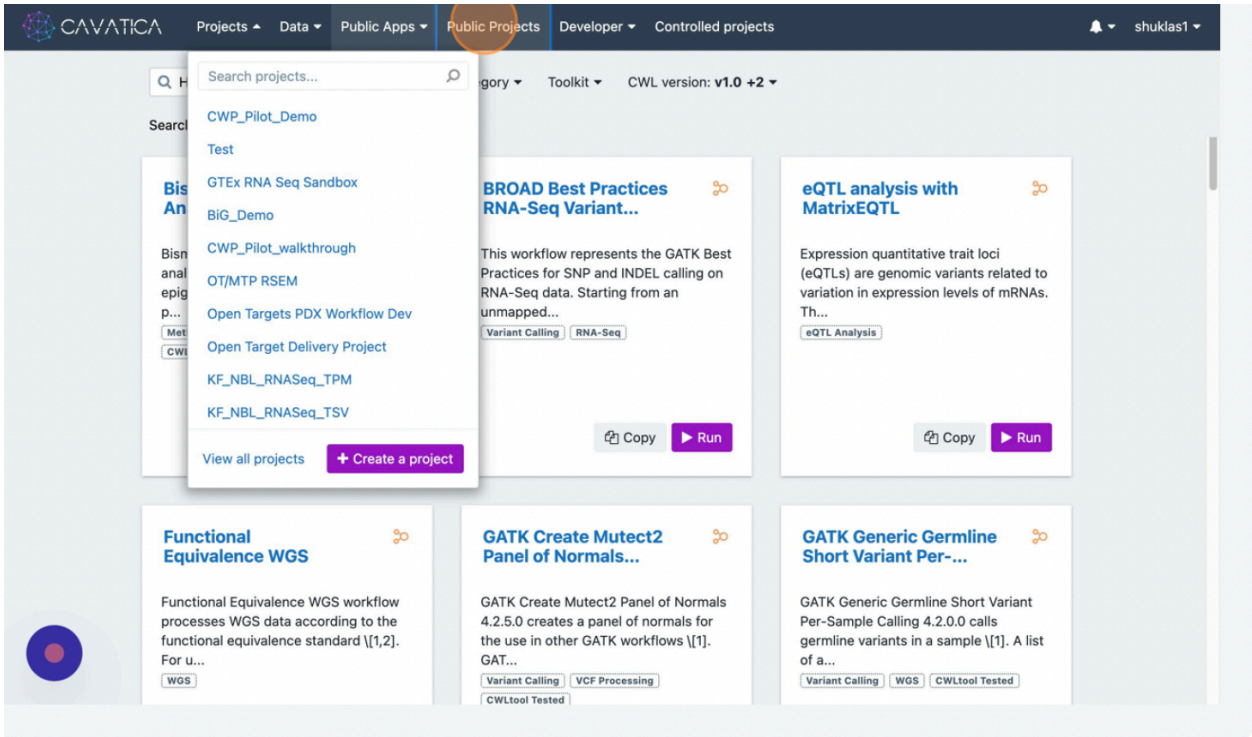


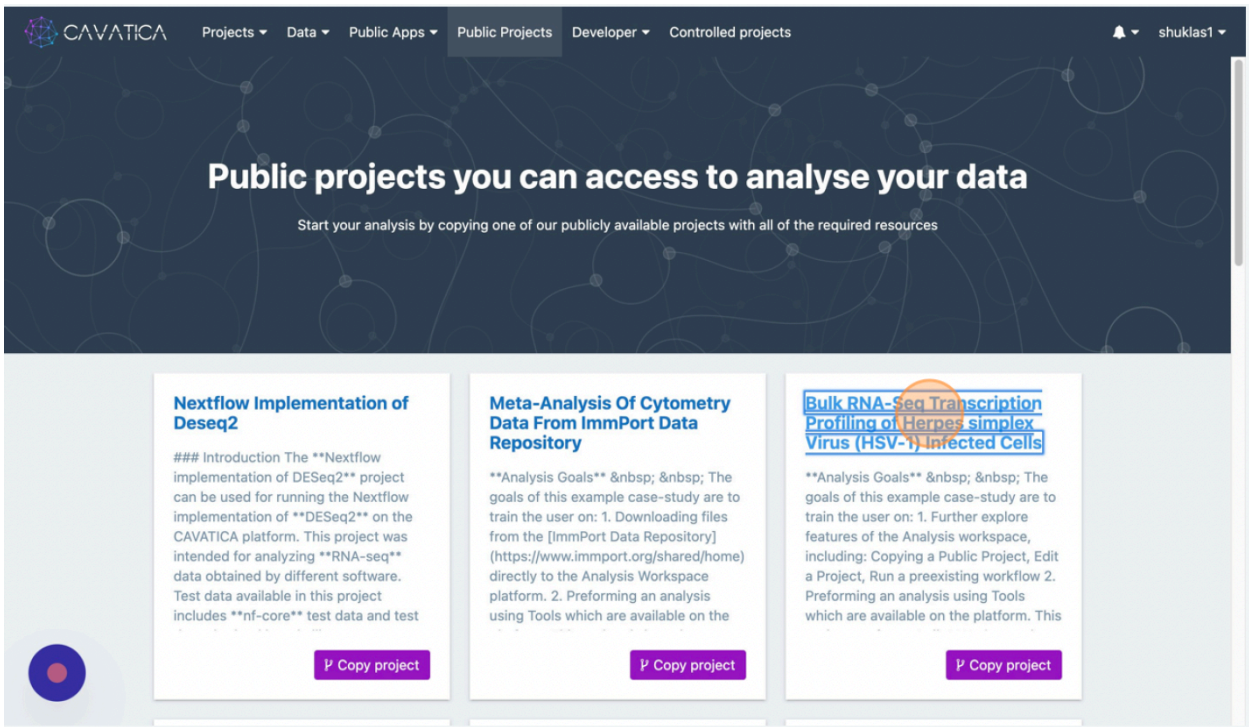
Now that you have created a new project, you may want to import appropriate workflow for analysis and data into the project. Once you run your analytical pipeline, you can also visualize and interpret results.

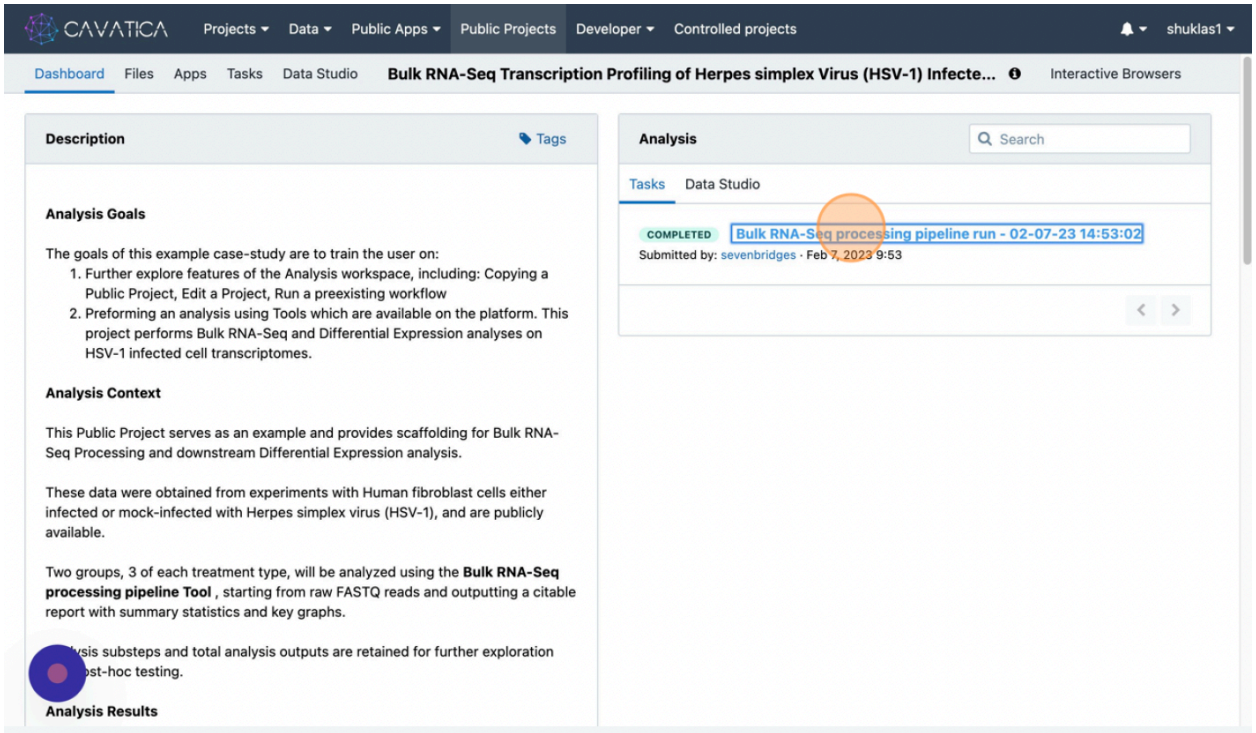
Create an App

For this document, we will copy/import an analysis from an existing public project that uses Fastq files, performs quantification, and implements differential expression analysis.

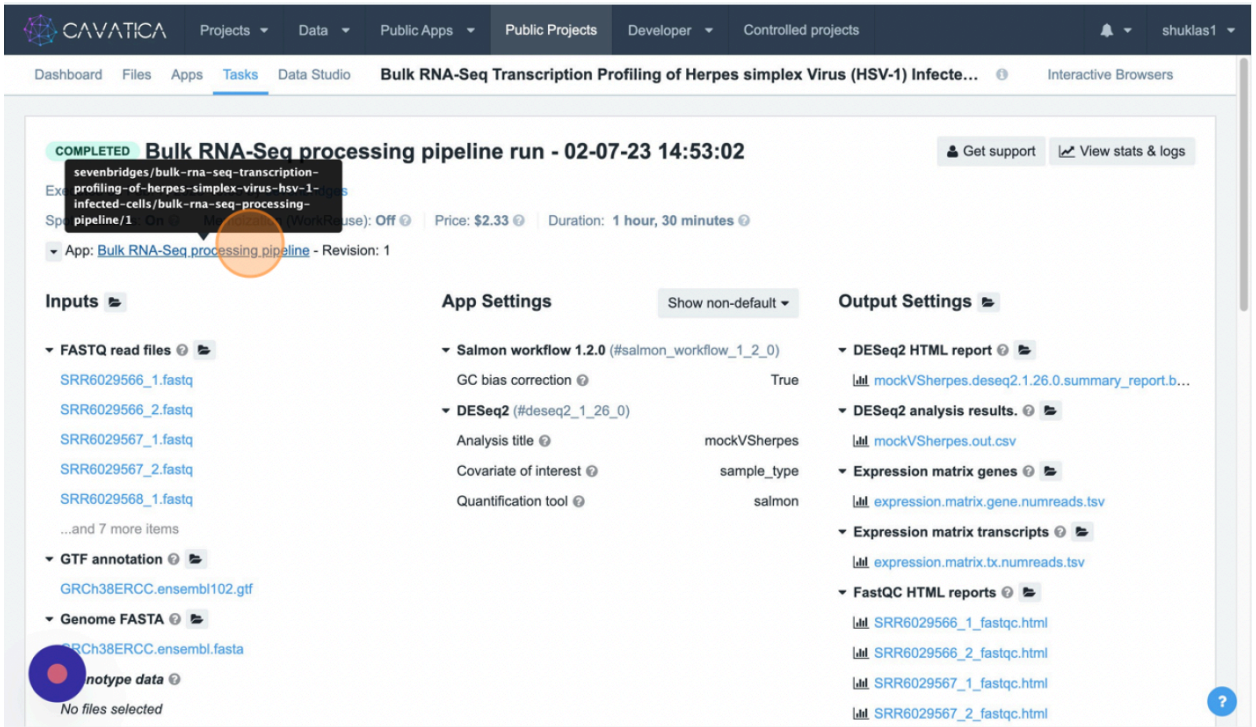
Follow the steps below to search the project and copy the embedded application into our newly created project.


CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Search projects...
CWP_Pilot_Demo
Test
GTEx RNA Seq Sandbox
BiG_Demo
CWP_Pilot_walkthrough
OT/MTP RSEM
Open Targets PDX Workflow Dev
Open Target Delivery Project
KF_NBL_RNASeq_TPM
KF_NBL_RNASeq_TSV
View all projects
Create a project
Toolkit
CWL version: v1.0 +2
BROAD Best Practices RNA-Seq Variant...
This workflow represents the GATK Best Practices for SNP and INDEL calling on RNA-Seq data. Starting from an unmapped...
Variant Calling
RNA-Seq
Copy
Run
eQTL analysis with MatrixEQTL
Expression quantitative trait loci (eQTLs) are genomic variants related to variation in expression levels of mRNAs. Th...
eQTL Analysis
Copy
Run
Functional Equivalence WGS
Functional Equivalence WGS workflow processes WGS data according to the functional equivalence standard \[1,2]. For u...
WGS
GATK Create Mutect2 Panel of Normals...
GATK Create Mutect2 Panel of Normals 4.2.5.0 creates a panel of normals for the use in other GATK workflows \[1]. GAT...
Variant Calling
VCF Processing
CWLtool Tested
GATK Generic Germline Short Variant Per-...
GATK Generic Germline Short Variant Per-Sample Calling 4.2.0.0 calls germline variants in a sample \[1]. A list of a...
Variant Calling
WGS
CWLtool Tested

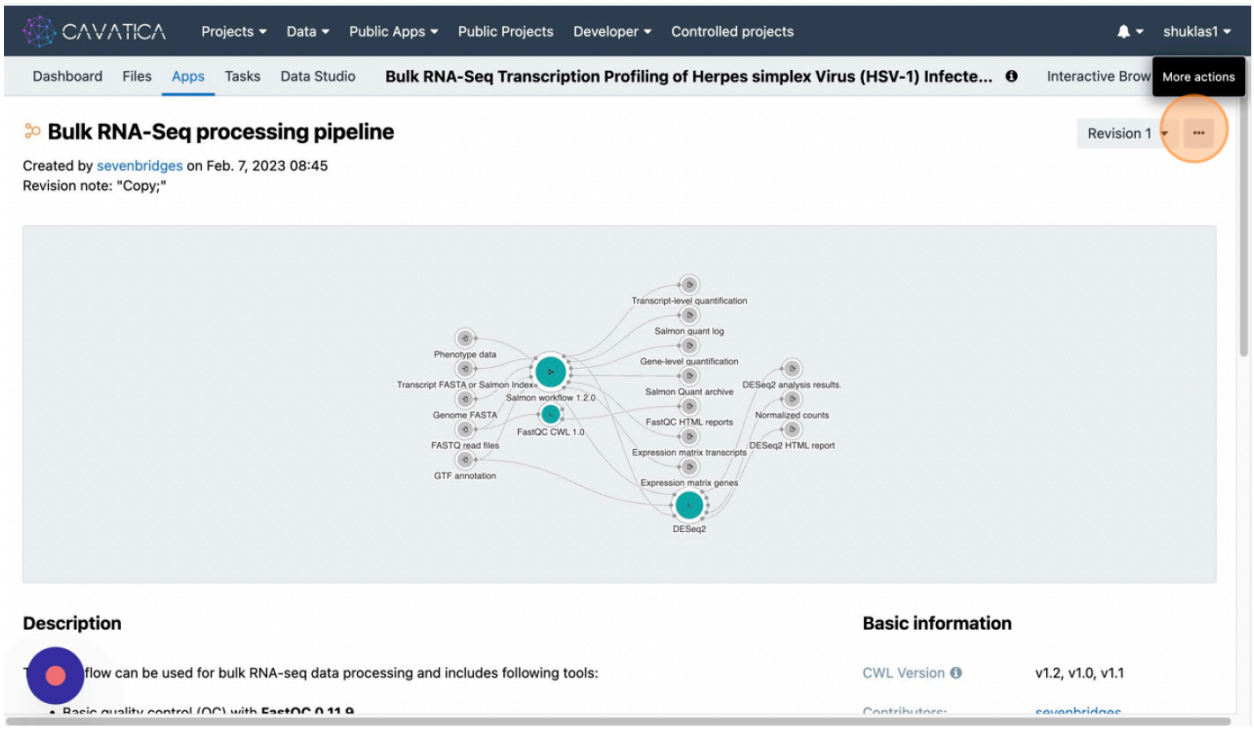

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Public projects you can access to analyse your data
Start your analysis by copying one of our publicly available projects with all of the required resources
Nextflow Implementation of Deseq2
Introduction The **Nextflow implementation of DESeq2** project can be used for running the Nextflow implementation of **DESeq2** on the CAVATICA platform. This project was intended for analyzing **RNA-seq** data obtained by different software. Test data available in this project includes **nf-core** test data and test
Copy project
Meta-Analysis Of Cytometry Data From ImmPort Data Repository
Analysis Goals   The goals of this example case-study are to train the user on: 1. Downloading files from the [ImmPort Data Repository] (https://www.immport.org/shared/home) directly to the Analysis Workspace platform. 2. Preforming an analysis using Tools which are available on the
Copy project
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (HSV-1) Infected Cells
Analysis Goals   The goals of this example case-study are to train the user on: 1. Further explore features of the Analysis workspace, including: Copying a Public Project, Edit a Project, Run a preexisting workflow 2. Preforming an analysis using Tools which are available on the platform. This
Copy project

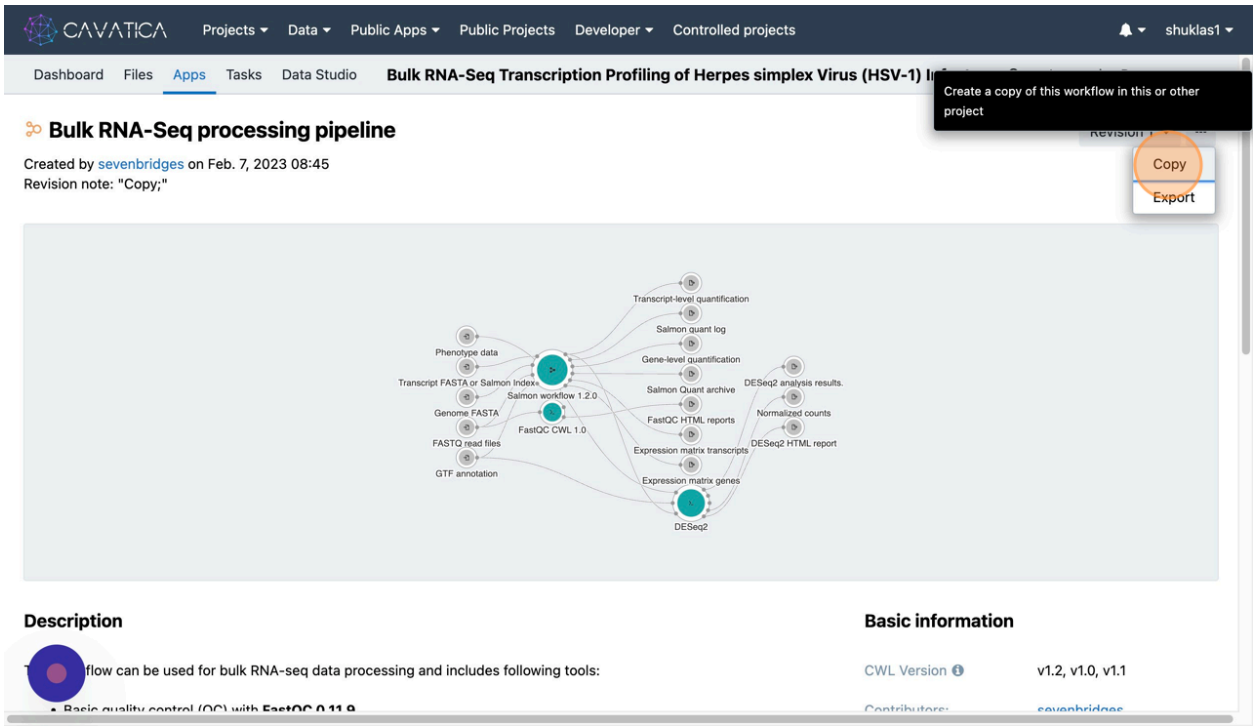


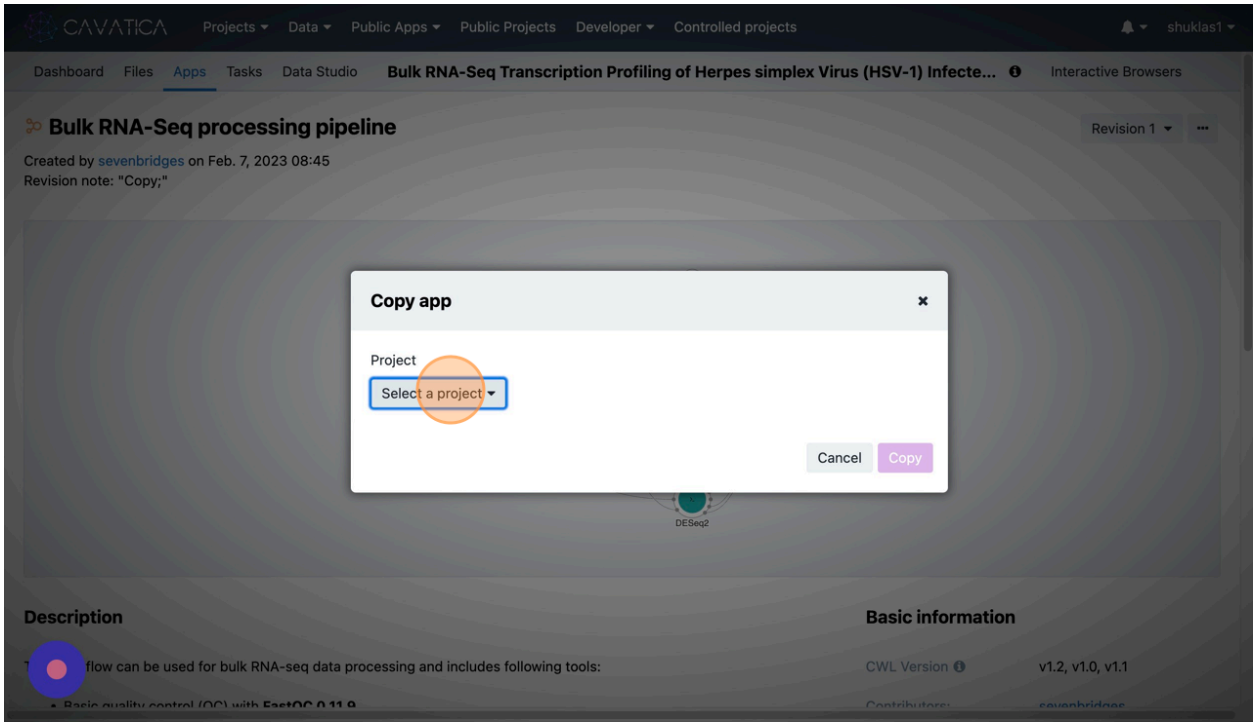
From the Dashboard of the project, navigate to the successfully completed Analysis task, to find the embedded application workflow.

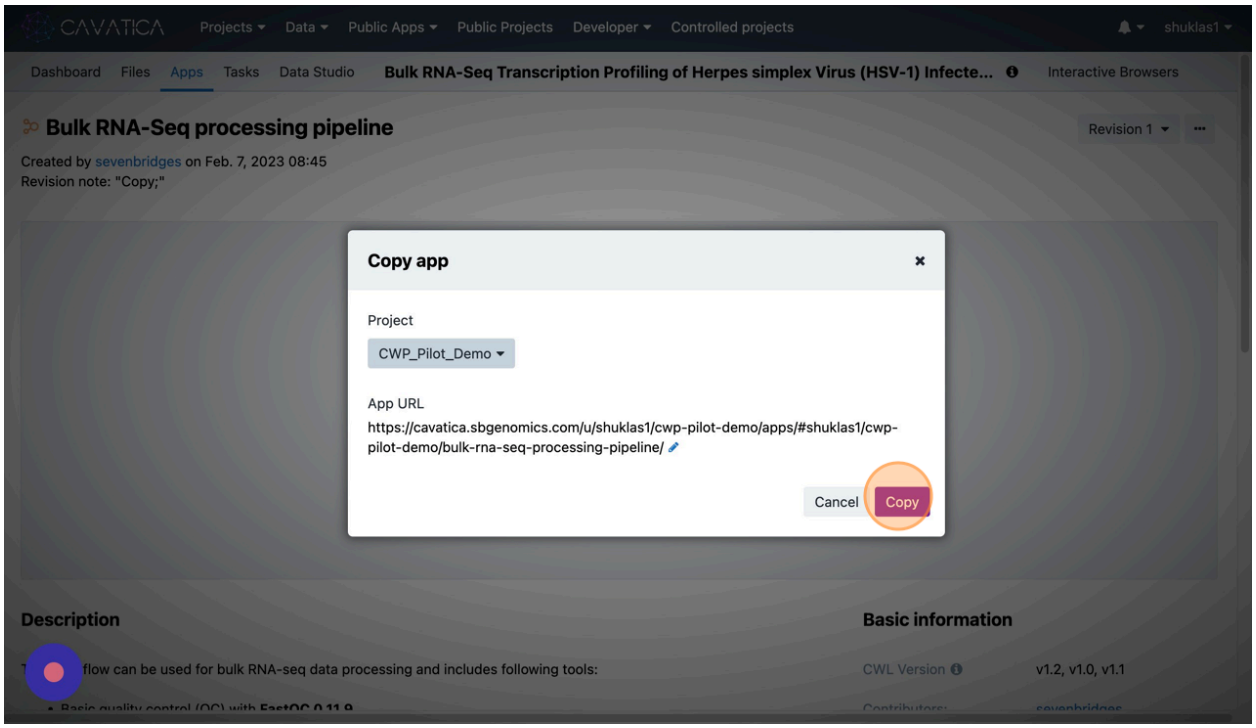


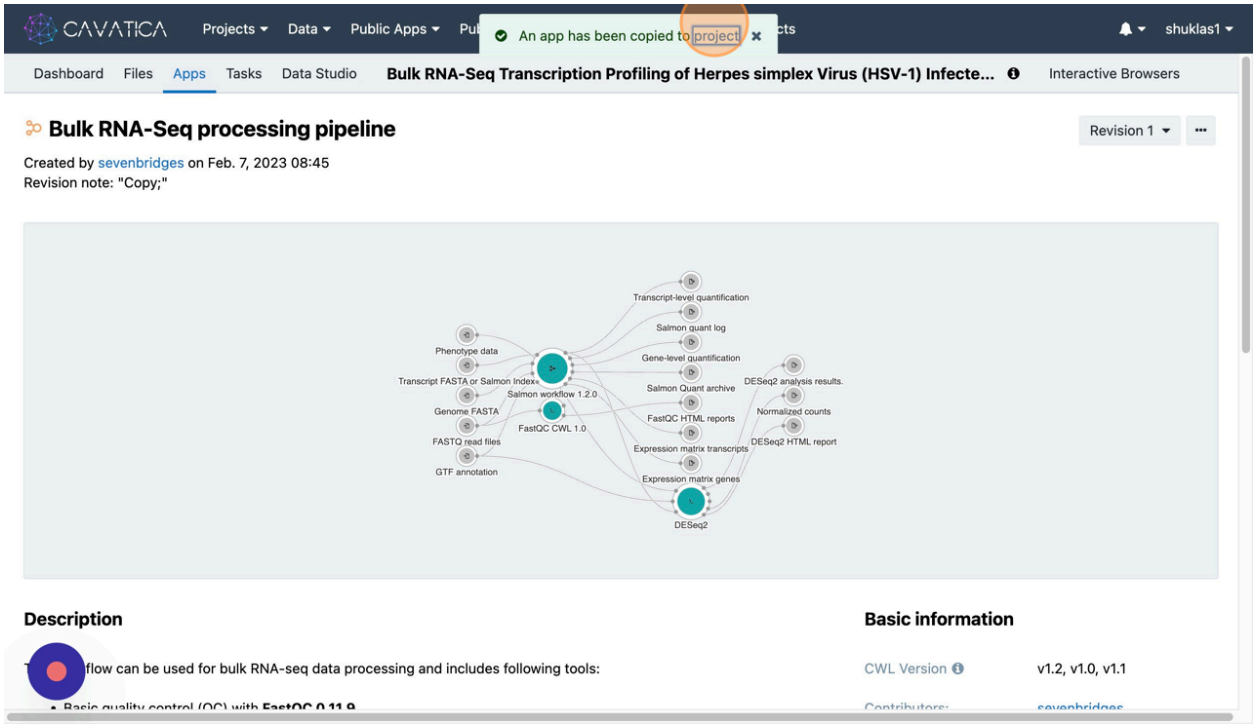
Using the ellipsis symbol, copy the app into our newly created project following the steps in screenshots below.





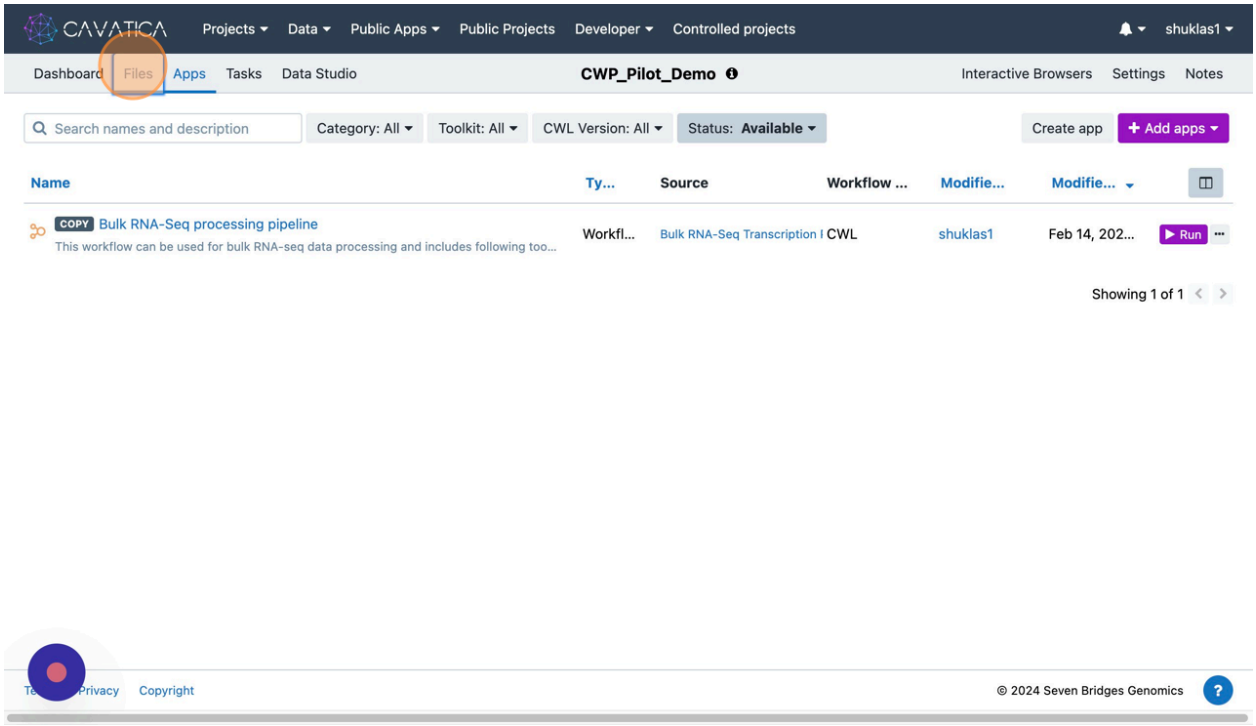

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (HSV-1) Infecte...
Interactive Browsers
Bulk RNA-Seq processing pipeline
Revision 1
Created by sevenbridges on Feb. 7, 2023 08:45
Revision note: "Copy;"
Copy app
Project
Select a project
Cancel
Copy
DESeq2
Description
Basic information
flow can be used for bulk RNA-seq data processing and includes following tools:
CWL Version
v1.2, v1.0, v1.1


CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (HSV-1) Infecte...
Interactive Browsers
Bulk RNA-Seq processing pipeline
Revision 1
Created by sevenbridges on Feb. 7, 2023 08:45
Revision note: "Copy;"
Copy app
Project
CWP_Pilot_Demo
App URL
https://cavatica.sbgenomics.com/u/shuklas1/cwp-pilot-demo/apps/#shuklas1/cwp-pilot-demo/bulk-rna-seq-processing-pipeline/
Cancel
Copy
Description
Basic information
flow can be used for bulk RNA-seq data processing and includes following tools:
CWL Version
v1.2, v1.0, v1.1

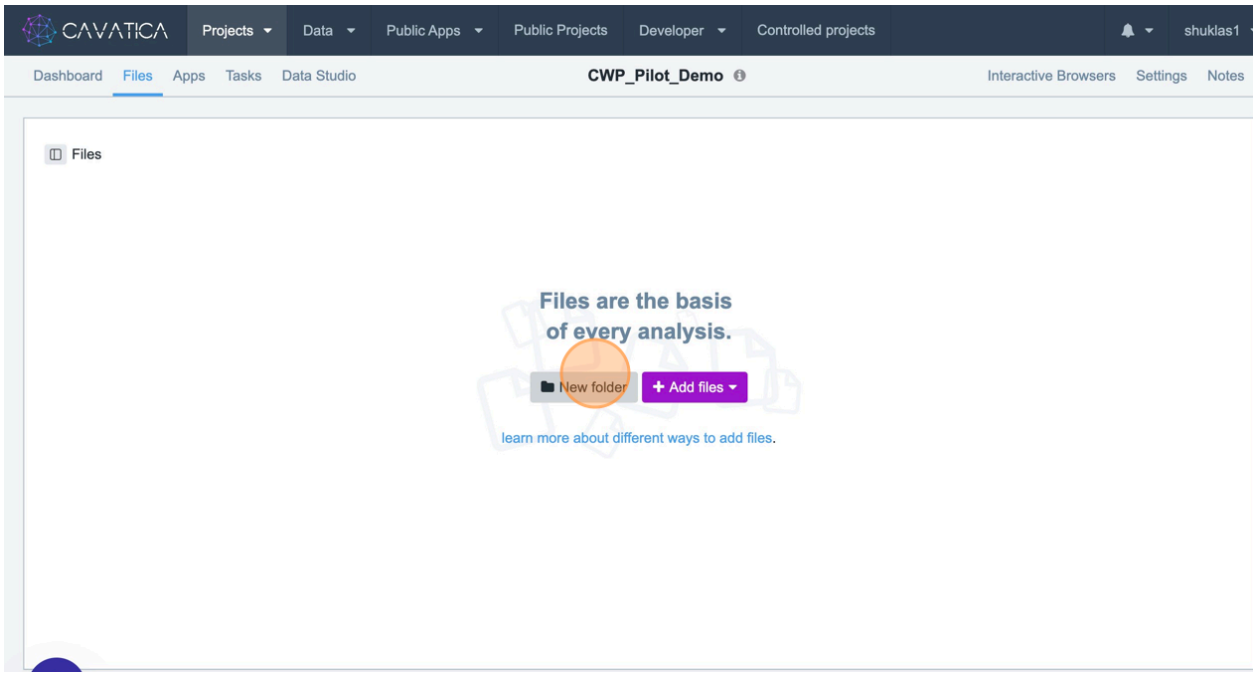


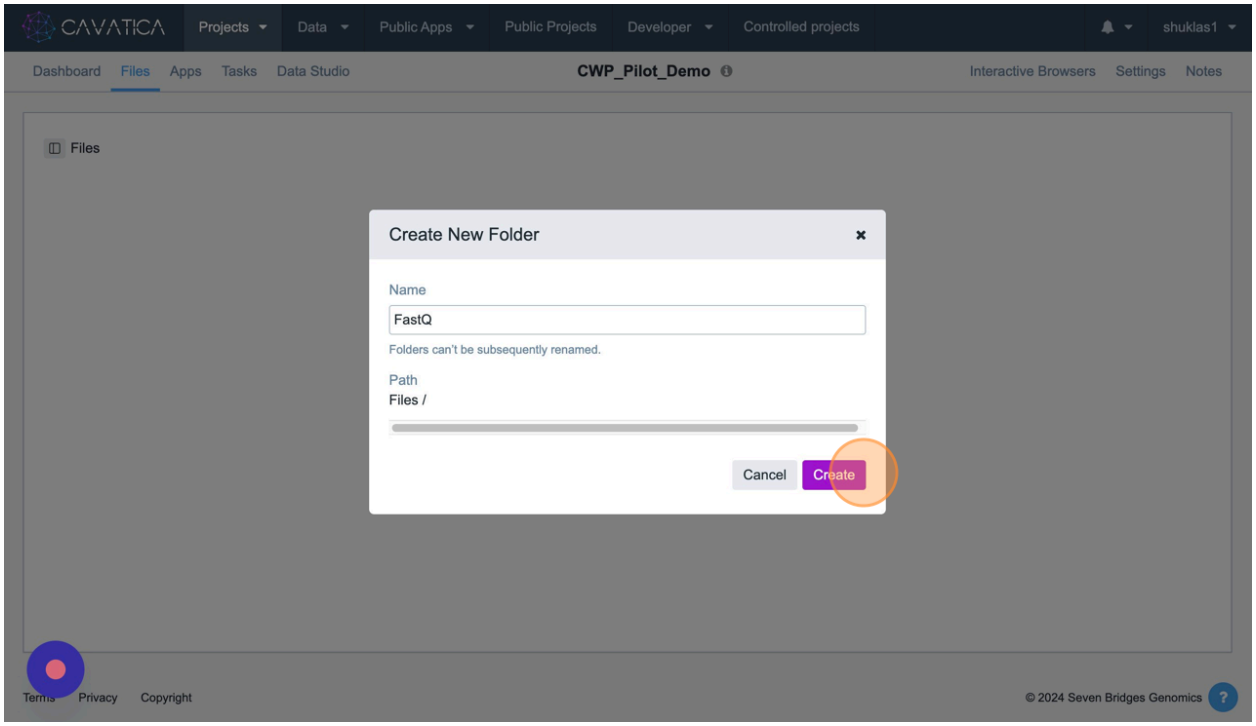
CFDE Portal Search and Data Export

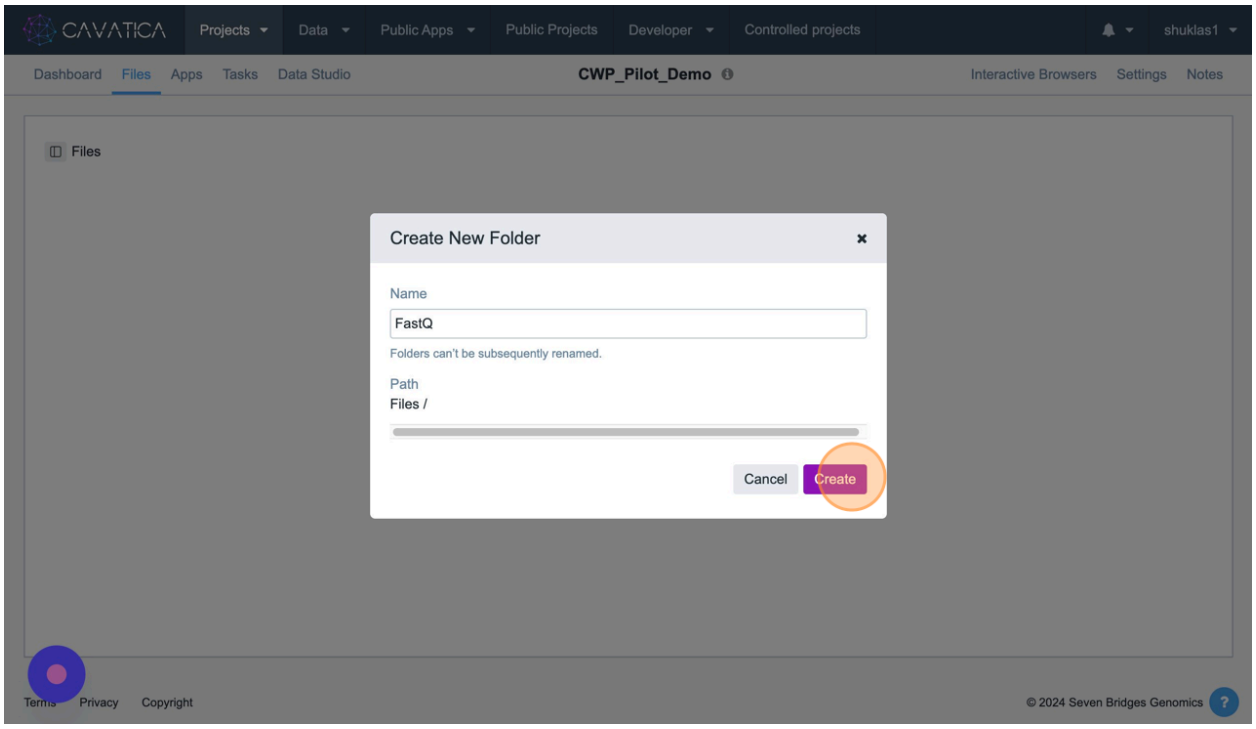
As you navigate back to the project, you will now see the app embedded.

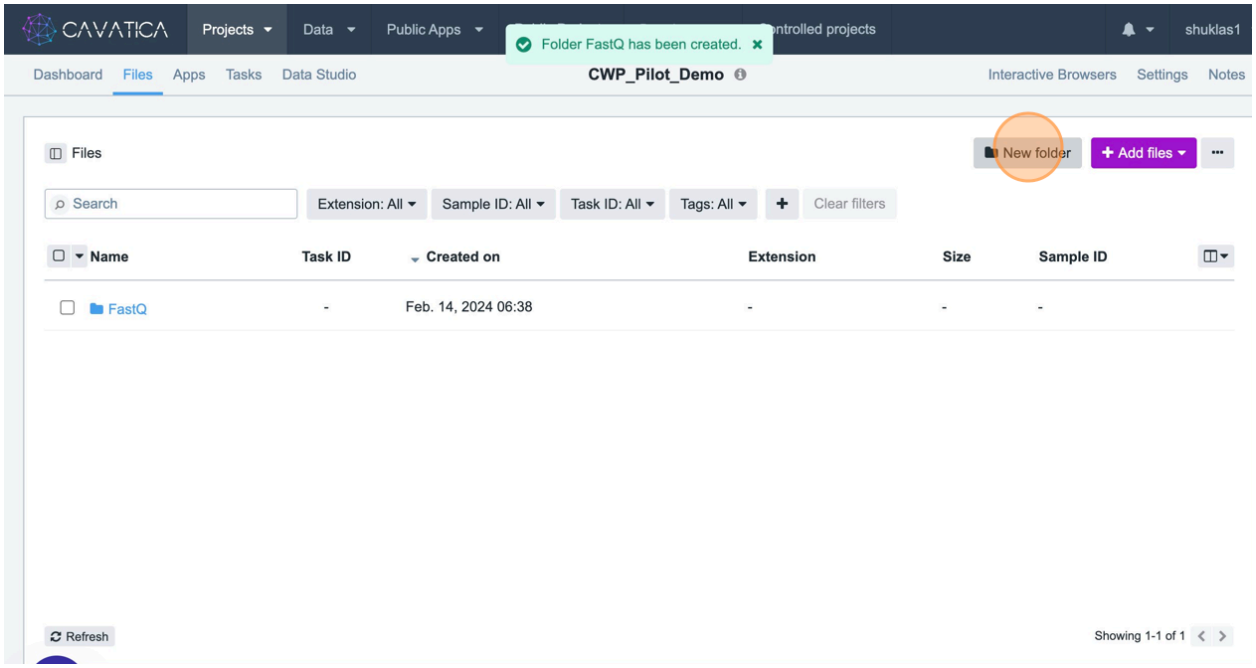


Now it is time to create appropriate folder structure and bring in data files important for the analysis. This demo requires some fastq files and reference genome files. To hold them, below screenshots show how the folders were created.



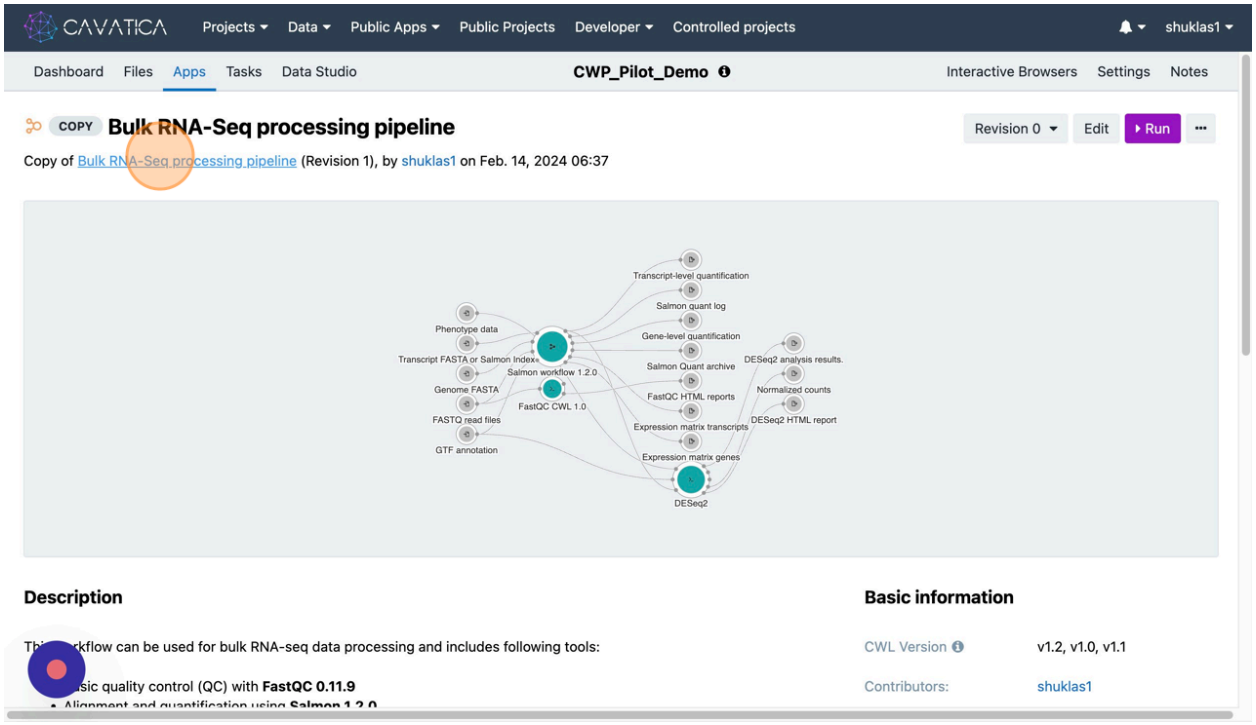

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
CWP_Pilot_Demo
Interactive Browsers
Settings
Notes
Files
Create New Folder
Name
FastQ
Folders can't be subsequently renamed.
Path
Files /
Cancel
Create
Terms
Privacy
Copyright
© 2024 Seven Bridges Genomics

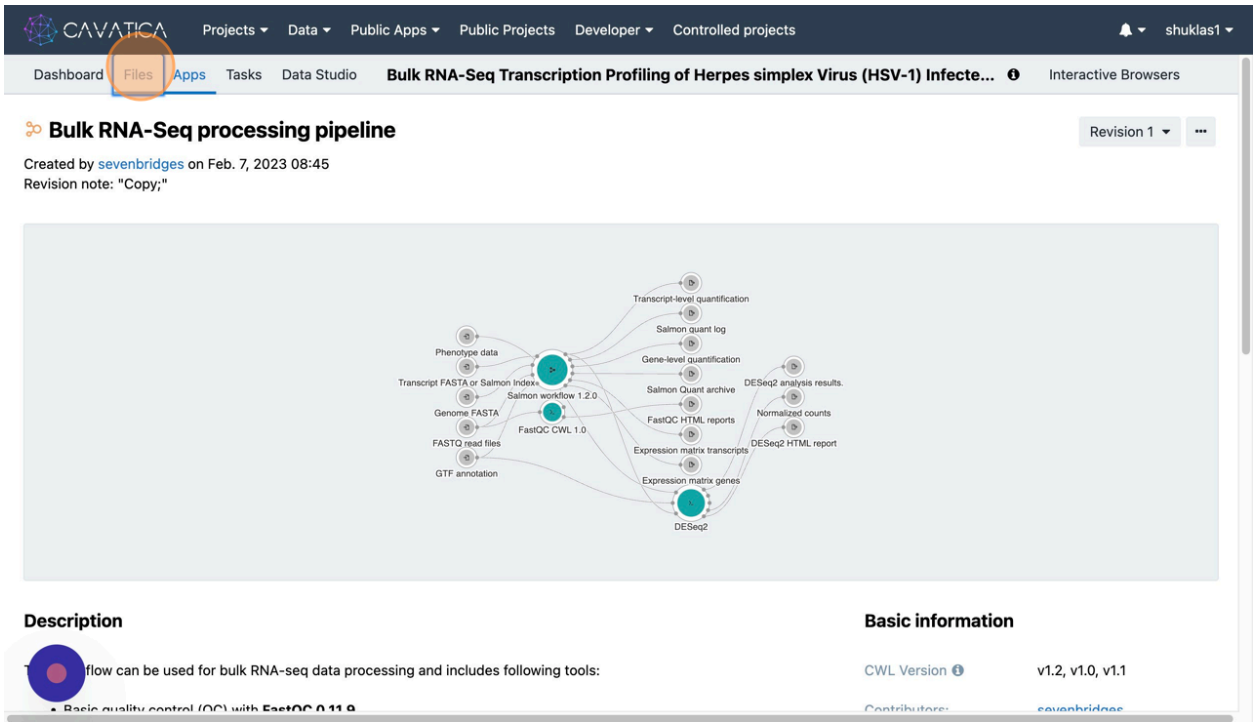

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
CWP_Pilot_Demo
Interactive Browsers
Settings
Notes
Files
Create New Folder
Name
FastQ
Folders can't be subsequently renamed.
Path
Files /
Cancel
Create
Terms
Privacy
Copyright
© 2024 Seven Bridges Genomics



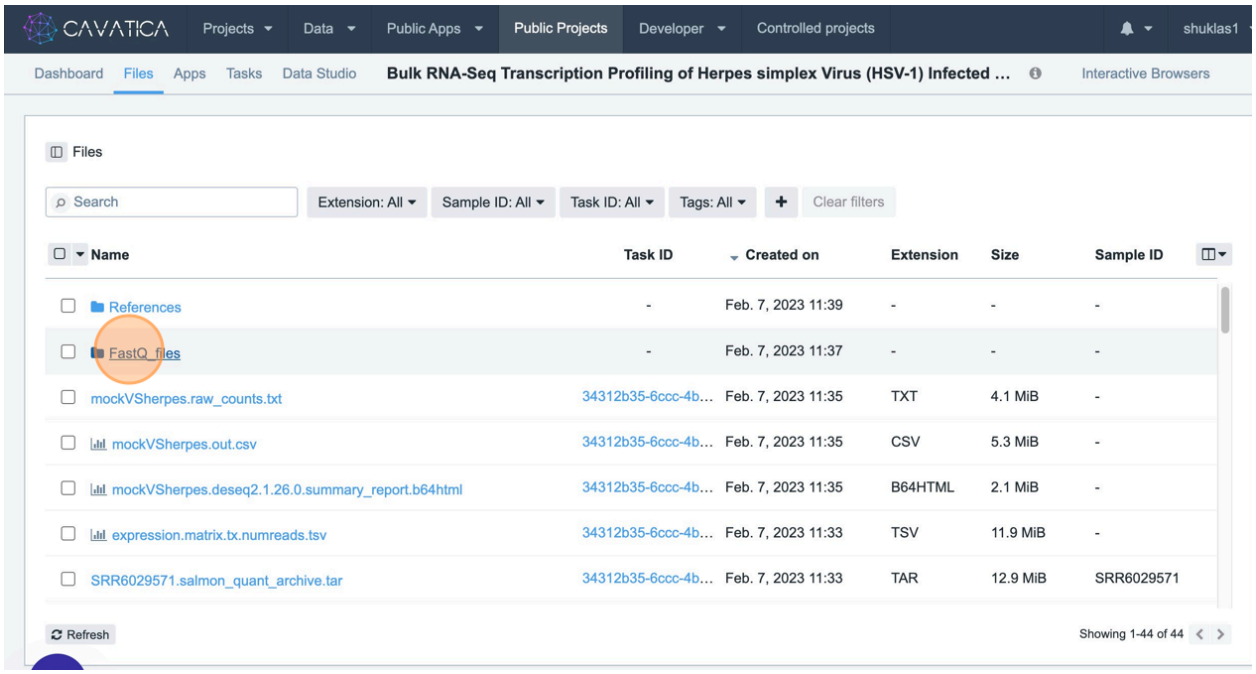
It is a simple process that can be repeated to create as many folders as needed. While the goal of this document is to highlight precise steps to import data from CFDE portal, for the purpose of this document, I will include the data files from the public project, since they are smaller in size and easier to run a quick analysis. Of course, screenshots below will also demonstrate data import from CFDE portal.

Let's start with copying the files from the CAVATICA project.
Navigate to the app again, and click on Files tab.

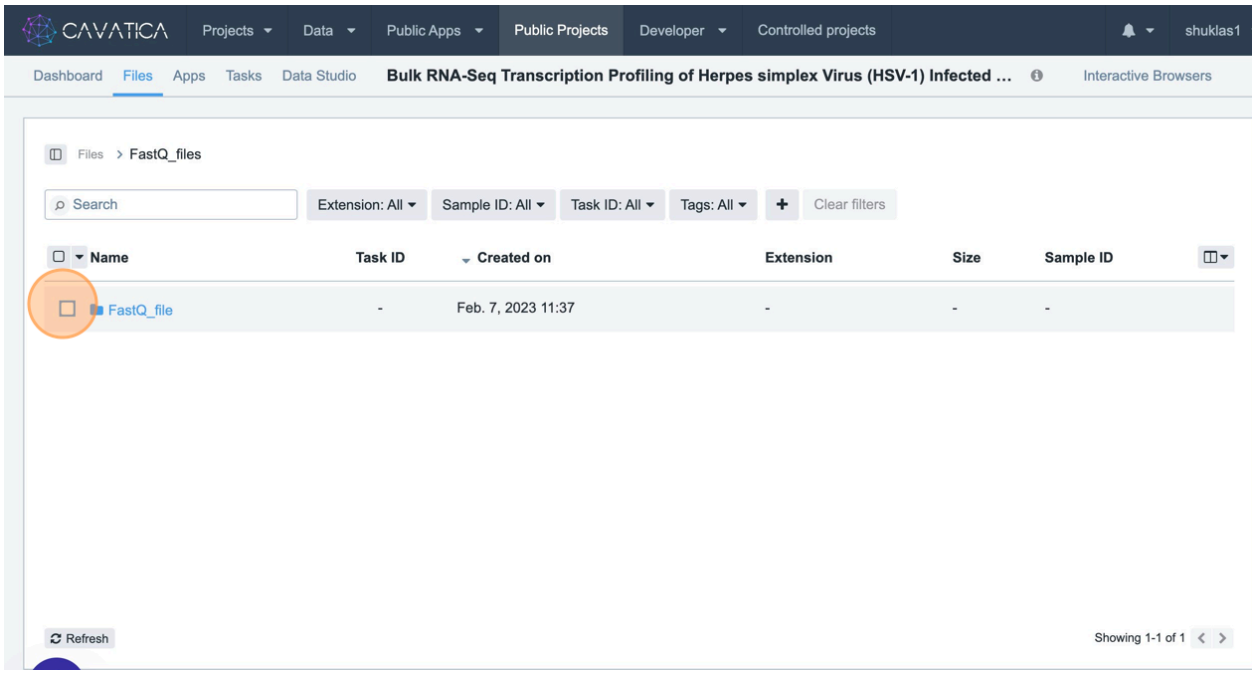


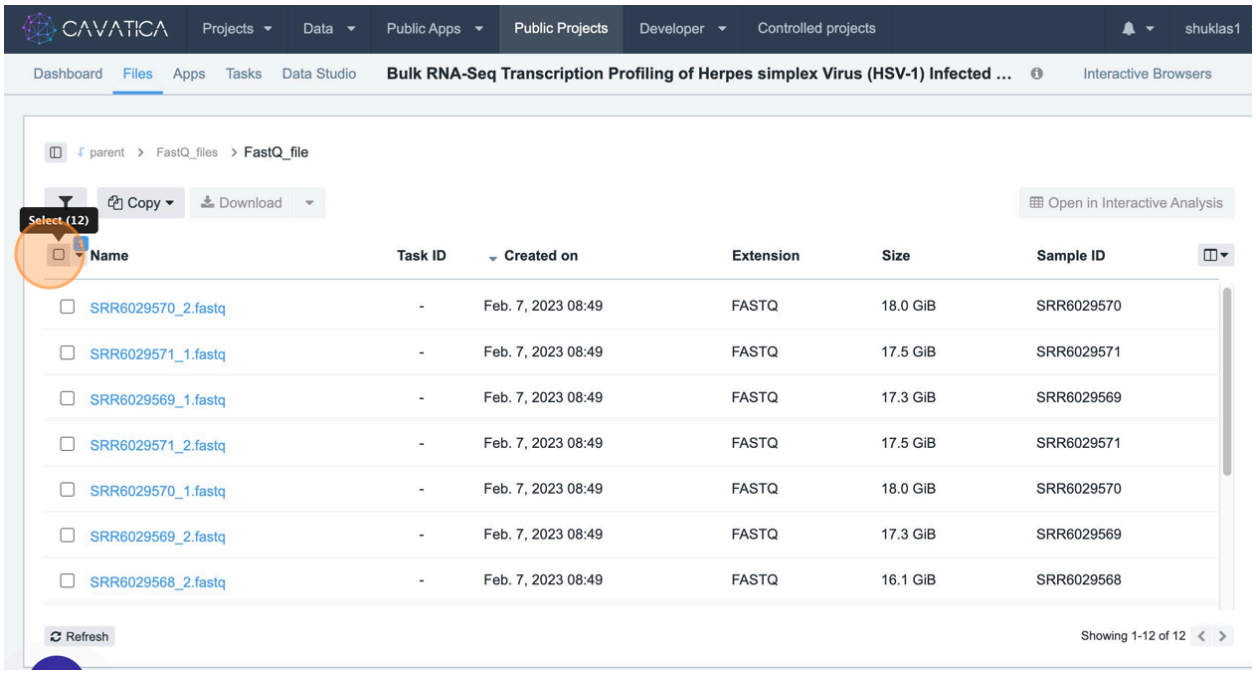


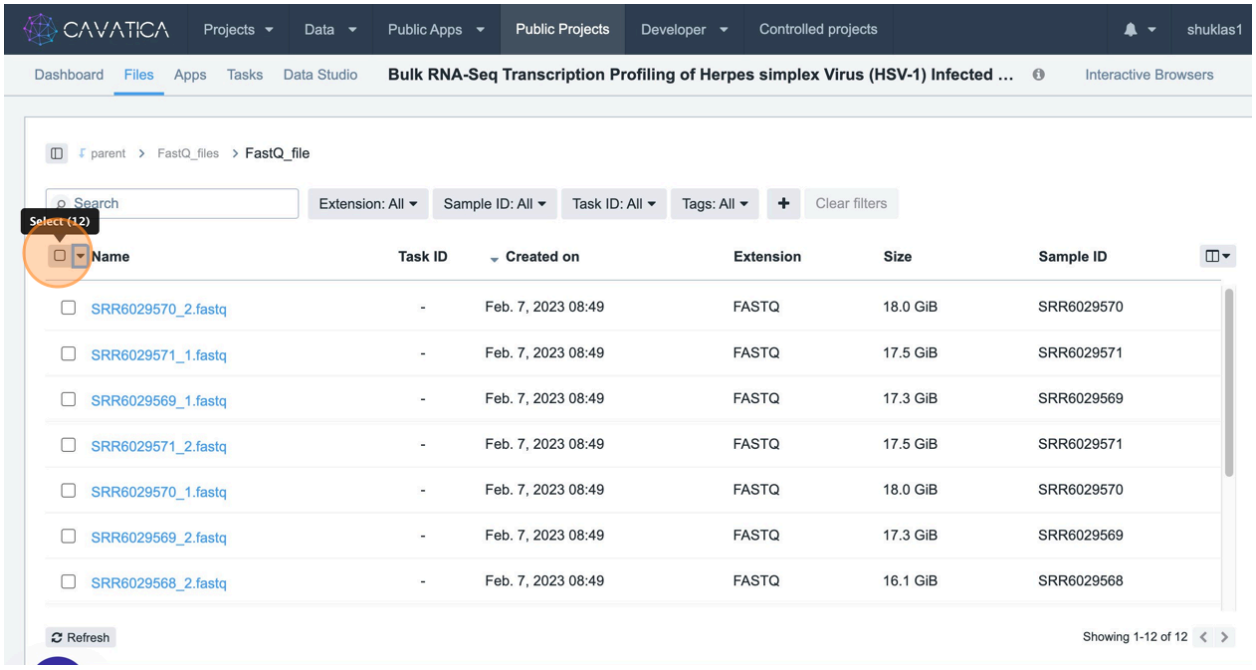
Those input files are split into two folders based on their data types. We will follow the same pattern for copying and storing files in the newly created project.

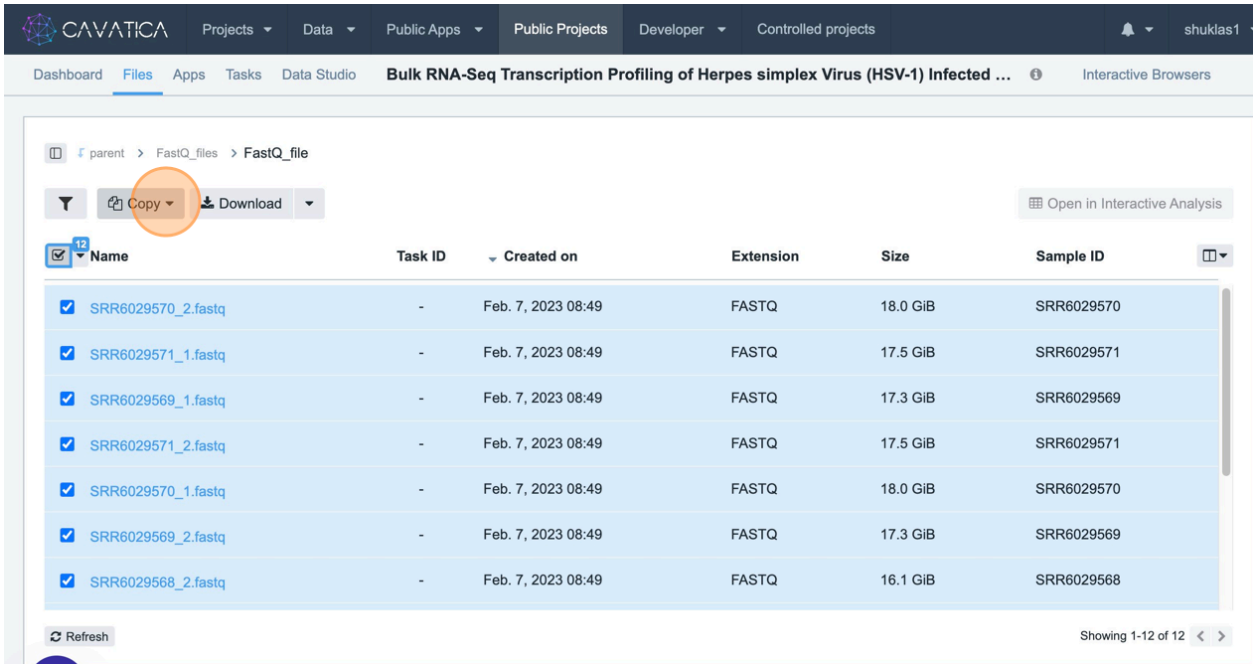


Navigate to the folder, select the file(s), click “Copy” and select the project and folder into which the files can be stored.



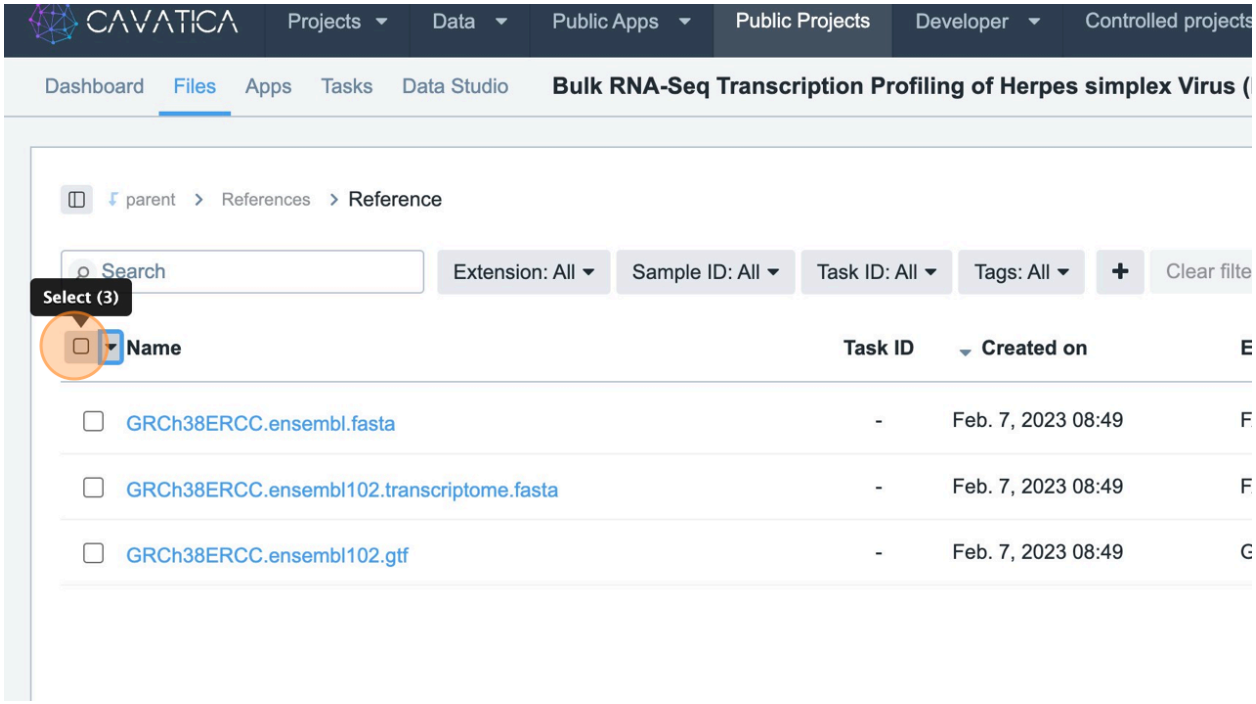

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (HSV-1) Infected ...
Interactive Browsers
parent > FastQ_files > FastQ_file
Copy
Download
Open in Interactive Analysis
Select (12)
Name
Task ID
Created on
Extension
Size
Sample ID
SRR6029570_2.fastq - Feb. 7, 2023 08:49 FASTQ 18.0 GiB SRR6029570
SRR6029571_1.fastq - Feb. 7, 2023 08:49 FASTQ 17.5 GiB SRR6029571
SRR6029569_1.fastq - Feb. 7, 2023 08:49 FASTQ 17.3 GiB SRR6029569
SRR6029571_2.fastq - Feb. 7, 2023 08:49 FASTQ 17.5 GiB SRR6029571
SRR6029570_1.fastq - Feb. 7, 2023 08:49 FASTQ 18.0 GiB SRR6029570
SRR6029569_2.fastq - Feb. 7, 2023 08:49 FASTQ 17.3 GiB SRR6029569
SRR6029568_2.fastq - Feb. 7, 2023 08:49 FASTQ 16.1 GiB SRR6029568
Refresh
Showing 1-12 of 12

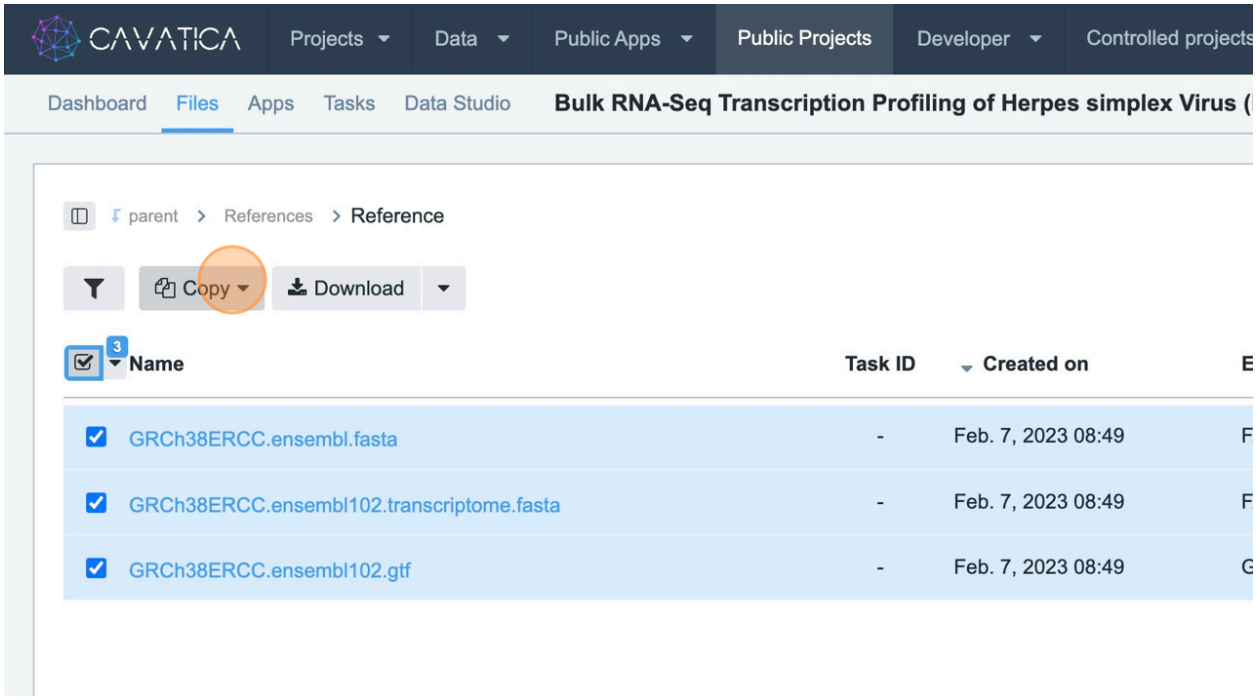

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (HSV-1) Infected ...
Interactive Browsers
parent > FastQ_files > FastQ_file
Search
Extension: All
Sample ID: All
Task ID: All
Tags: All
Clear filters
Select (12)
Name
Task ID
Created on
Extension
Size
Sample ID
SRR6029570_2.fastq - Feb. 7, 2023 08:49 FASTQ 18.0 GiB SRR6029570
SRR6029571_1.fastq - Feb. 7, 2023 08:49 FASTQ 17.5 GiB SRR6029571
SRR6029569_1.fastq - Feb. 7, 2023 08:49 FASTQ 17.3 GiB SRR6029569
SRR6029571_2.fastq - Feb. 7, 2023 08:49 FASTQ 17.5 GiB SRR6029571
SRR6029570_1.fastq - Feb. 7, 2023 08:49 FASTQ 18.0 GiB SRR6029570
SRR6029569_2.fastq - Feb. 7, 2023 08:49 FASTQ 17.3 GiB SRR6029569
SRR6029568_2.fastq - Feb. 7, 2023 08:49 FASTQ 16.1 GiB SRR6029568
Refresh
Showing 1-12 of 12

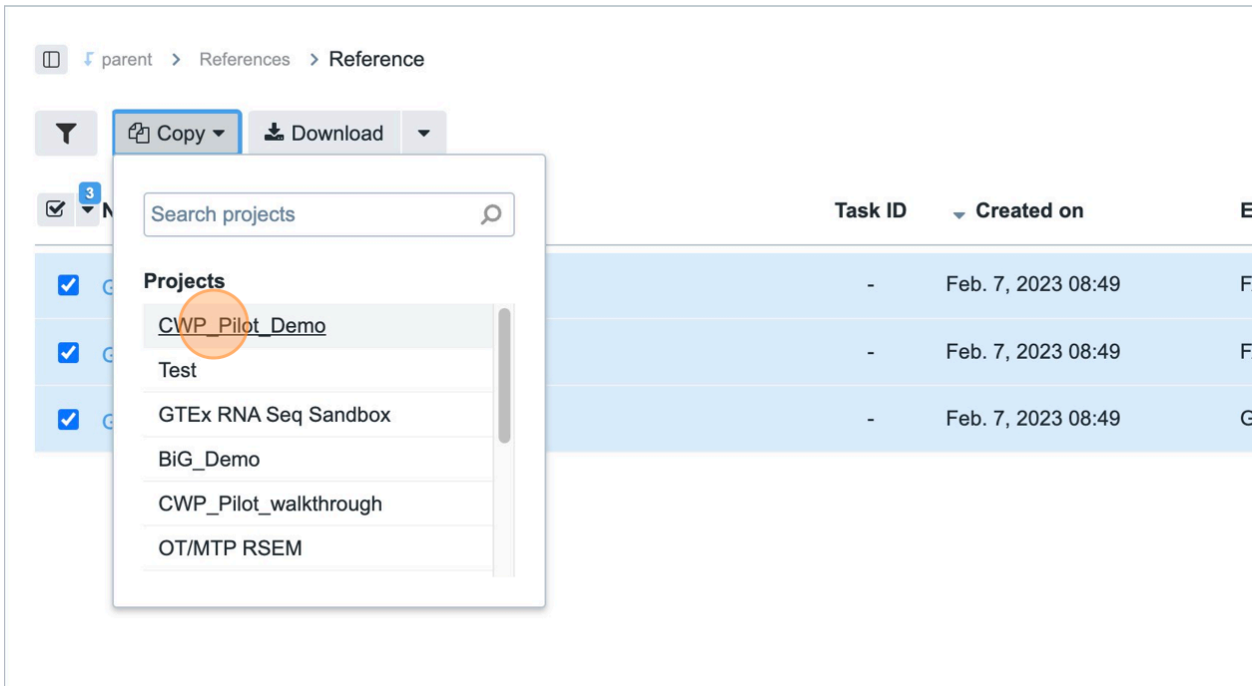


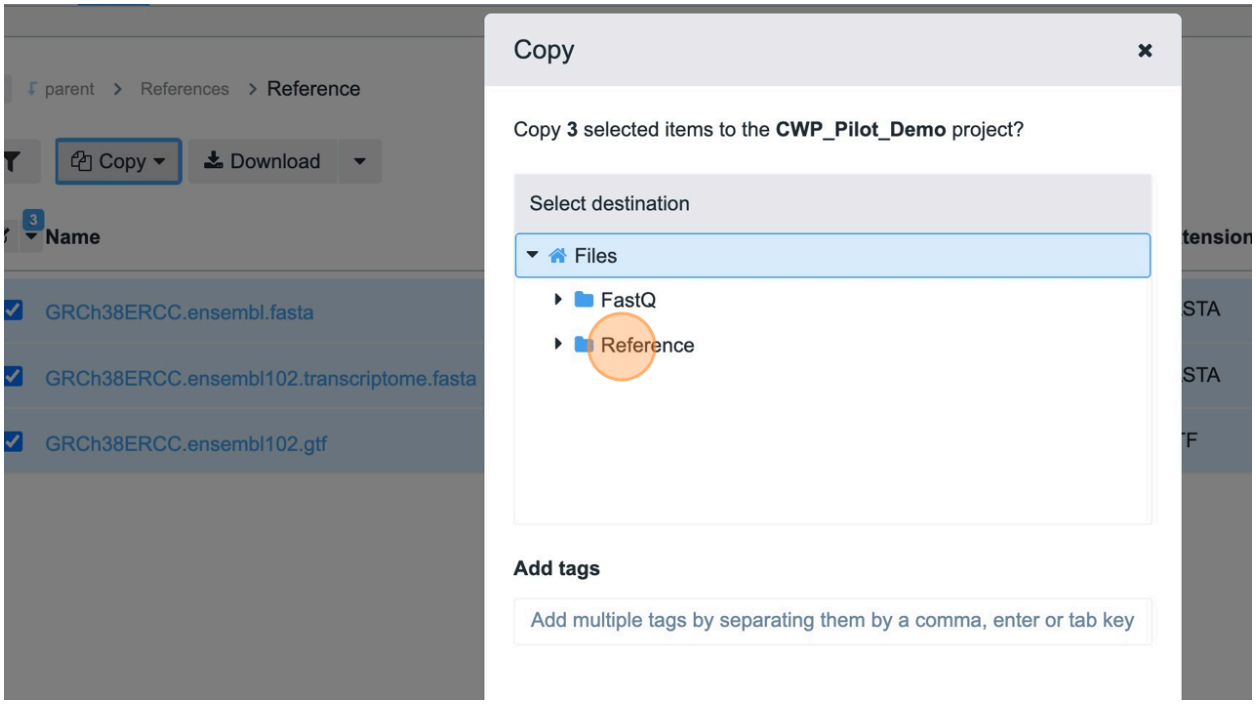
Repeat the same steps for copying as many files into as many project folders as needed.

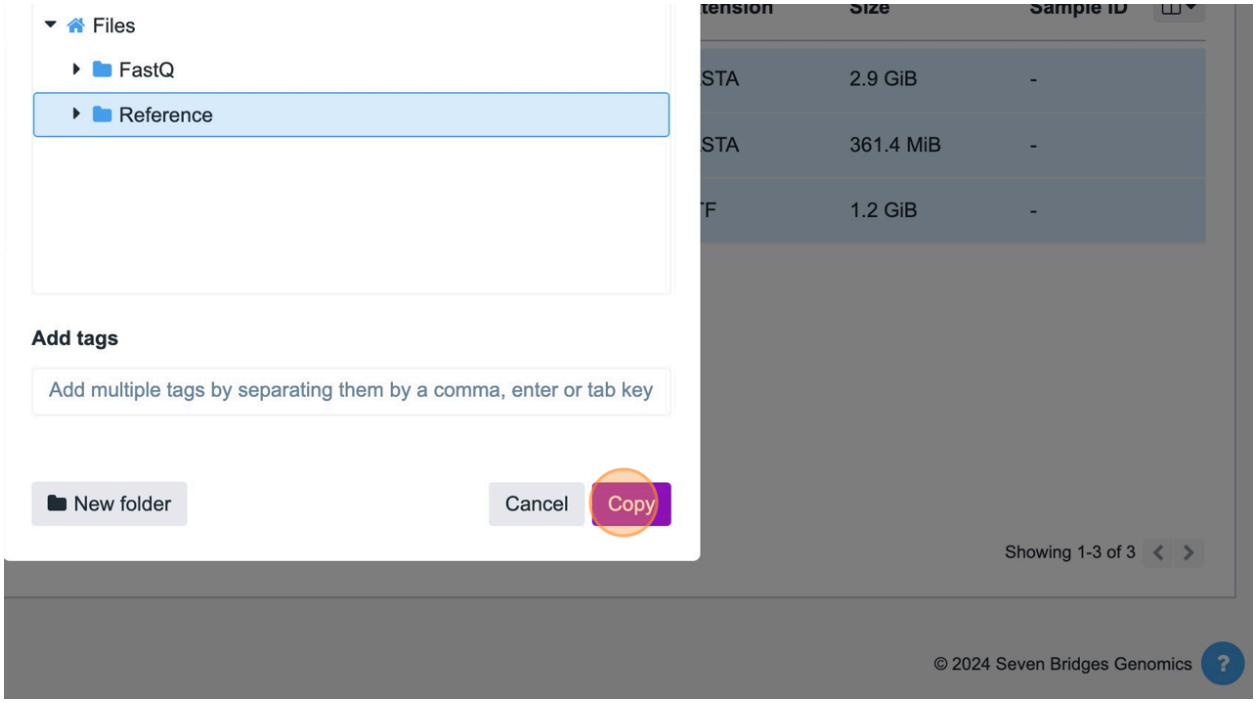
Below are screen shots for copying the reference files into the new project.



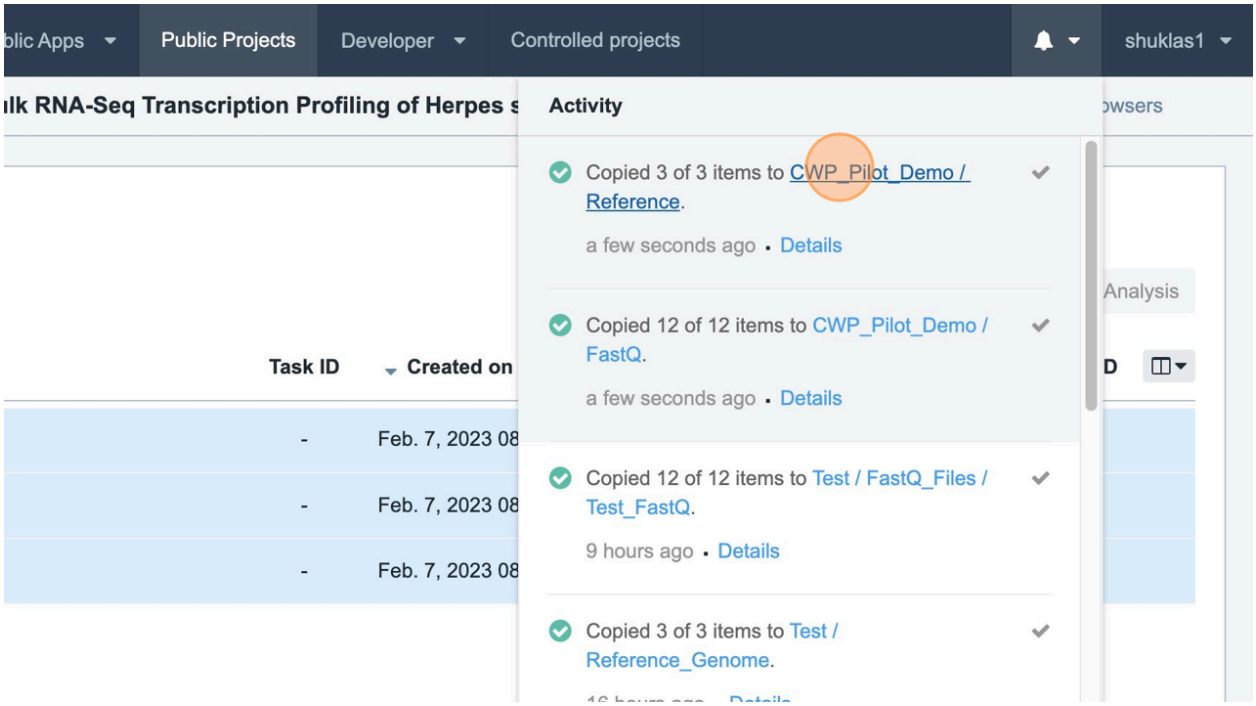

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
Dashboard
Files
Apps
Tasks
Data Studio
Bulk RNA-Seq Transcription Profiling of Herpes simplex Virus (
parent > References > Reference
Copy
Download
3
Name
Task ID
Created on
GRCh38ERCC.ensembl.fasta
-
Feb. 7, 2023 08:49
GRCh38ERCC.ensembl102.transcriptome.fasta
-
Feb. 7, 2023 08:49
GRCh38ERCC.ensembl102.gtf
-
Feb. 7, 2023 08:49


parent > References > Reference
Copy
Download
3
Search projects
Projects
CWP_Pilot_Demo
Test
GTEx RNA Seq Sandbox
BiG_Demo
CWP_Pilot_walkthrough
OT/MTP RSEM
Task ID
Created on
-
Feb. 7, 2023 08:49
-
Feb. 7, 2023 08:49
-
Feb. 7, 2023 08:49

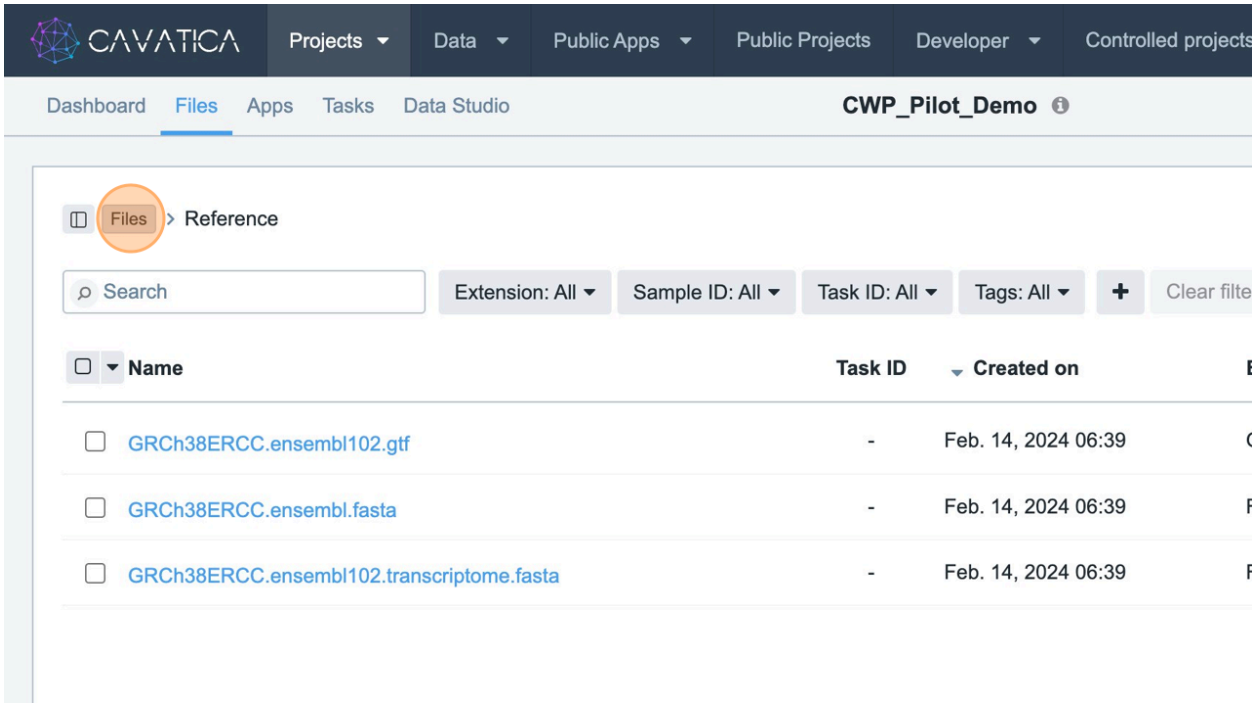




You will see a success notification for files copied as below.



And newly copied files will show in the project as below.

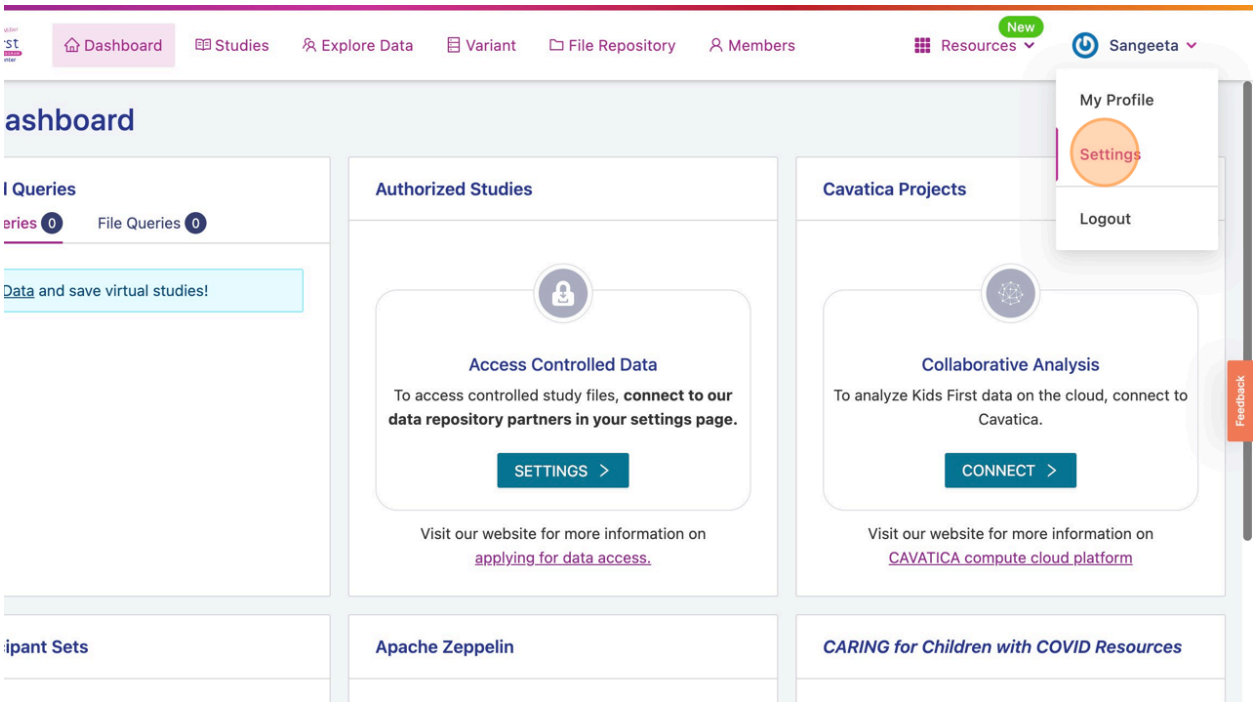


You can also choose to create an “Output” folder to store the resulting output files from the analysis. However, you will need to ensure that the paths specified within the analytical workflow are correctly directed.

Since I will be importing some files from the Kids First project, I will also need to connect my Kids First account to the CAVATICA portal shown as below.

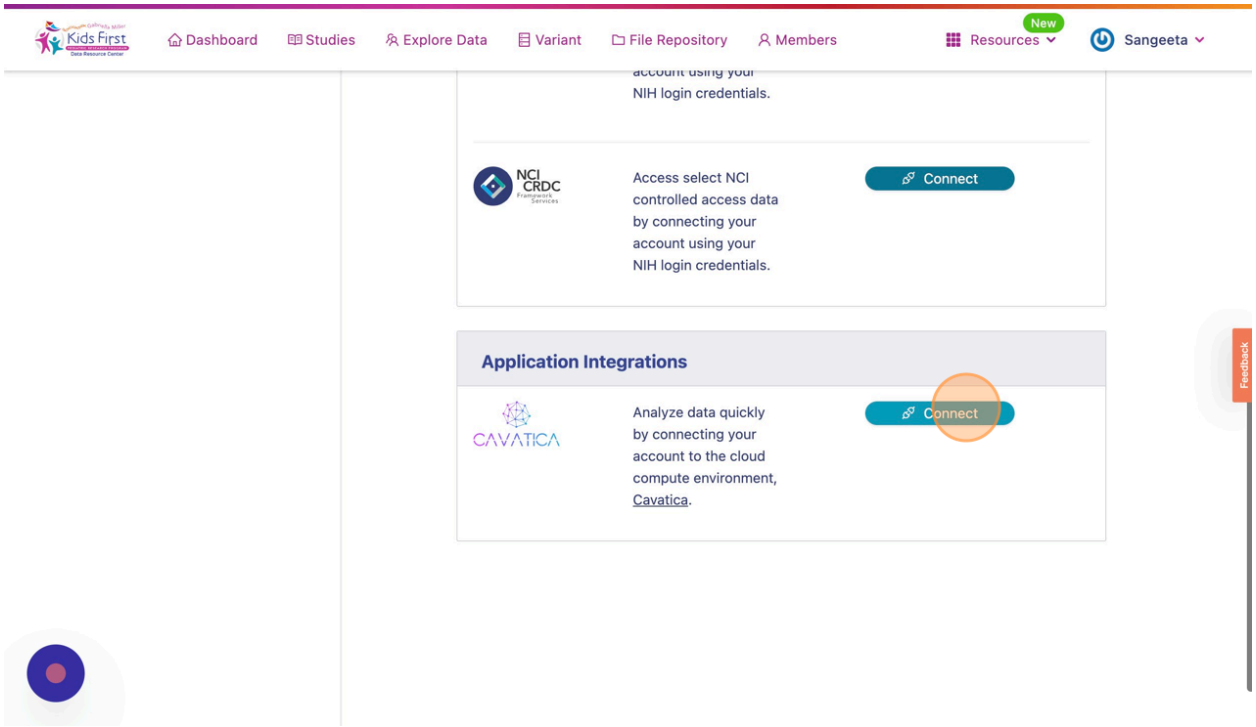
To do that, navigate and login to https://portal.kidsfirstdrc.org/dashboard

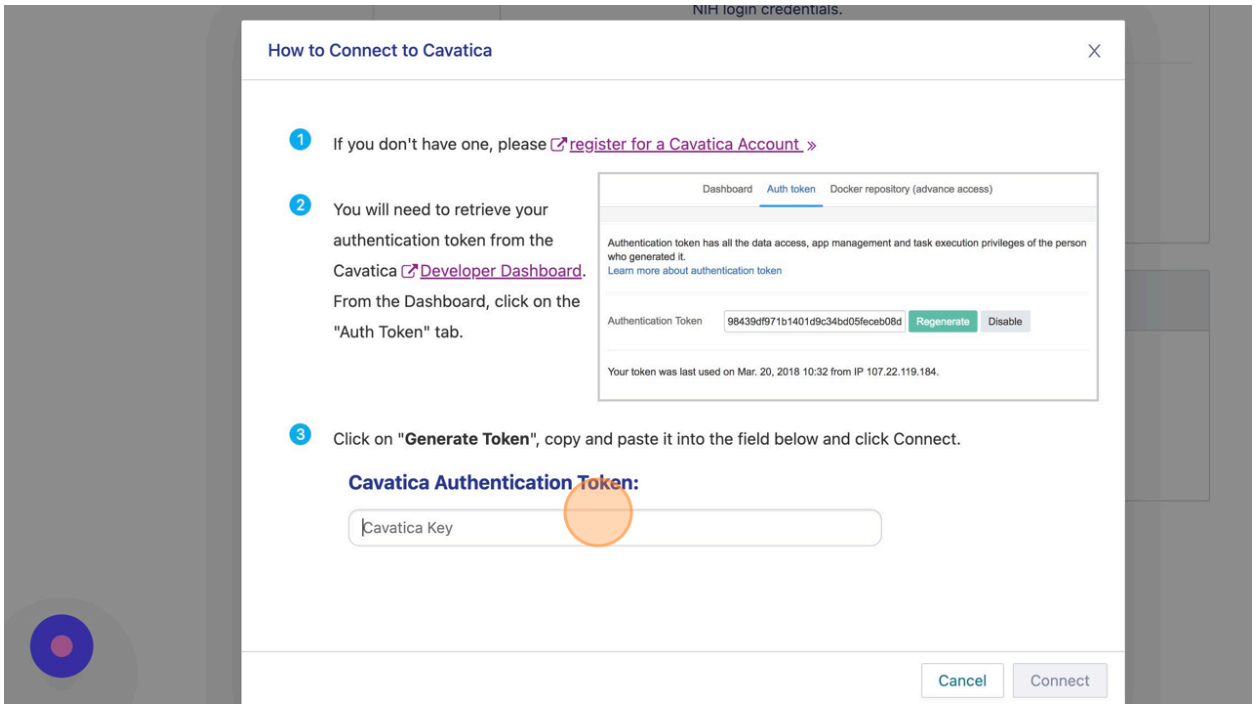
Then, go to Settings.



Scroll down to the Applications Integration section and click on ‘Connect’ to connect to CAVATICA.

You will need to get your ‘CAVATICA Authentication Token’ from the CAVATICA platform’s Developer tab, and paste it within the KidsFirst portal widget as prompted.



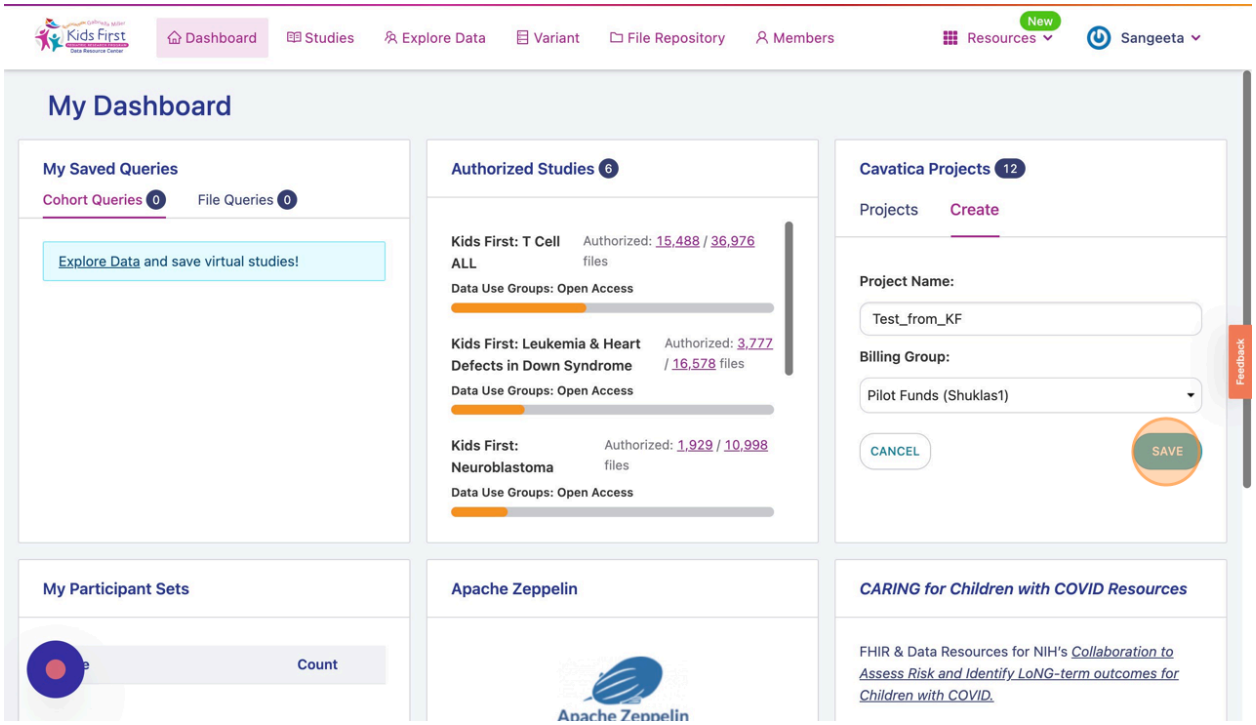


Then, click ‘Connect’.

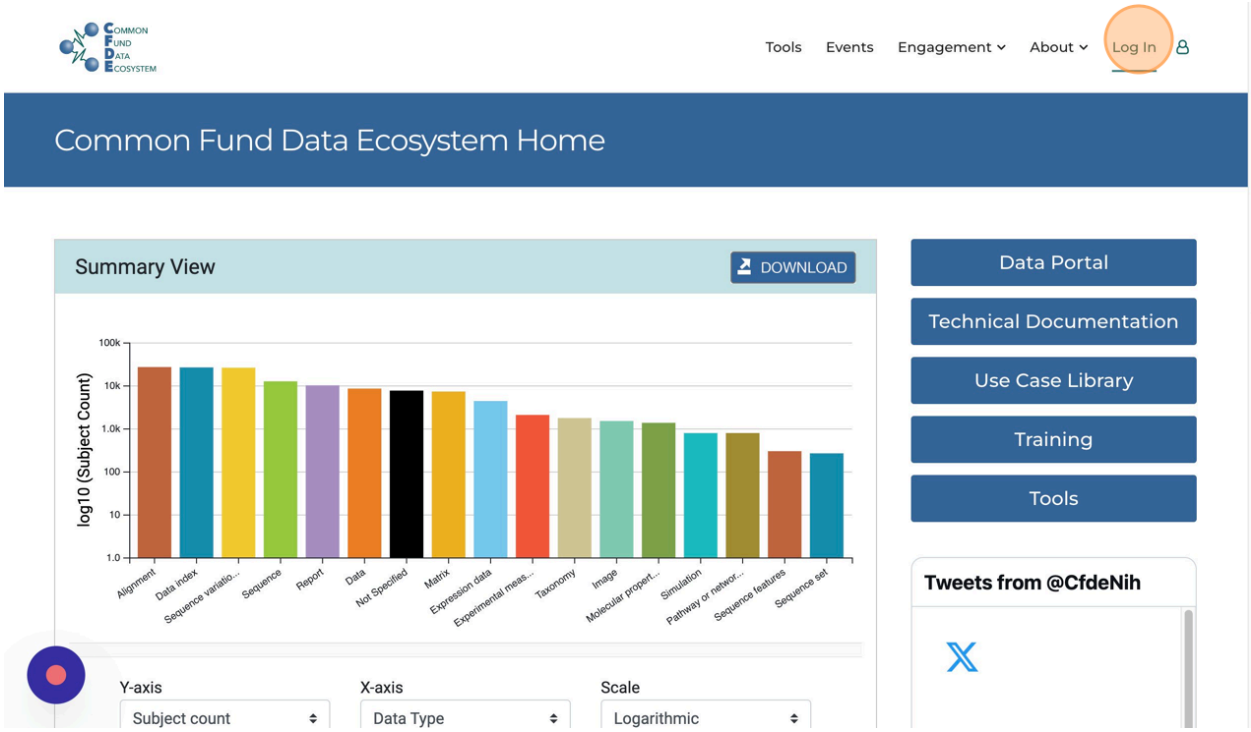
You must also connect from the KidsFirst portal to the other ‘Framework Services’ and ‘NCI CRDC’ to allow seamless connections as you import data collaboratively

across platforms. Be sure to stay logged in to your eRA Commons account while to attempt to connect to these frameworks.

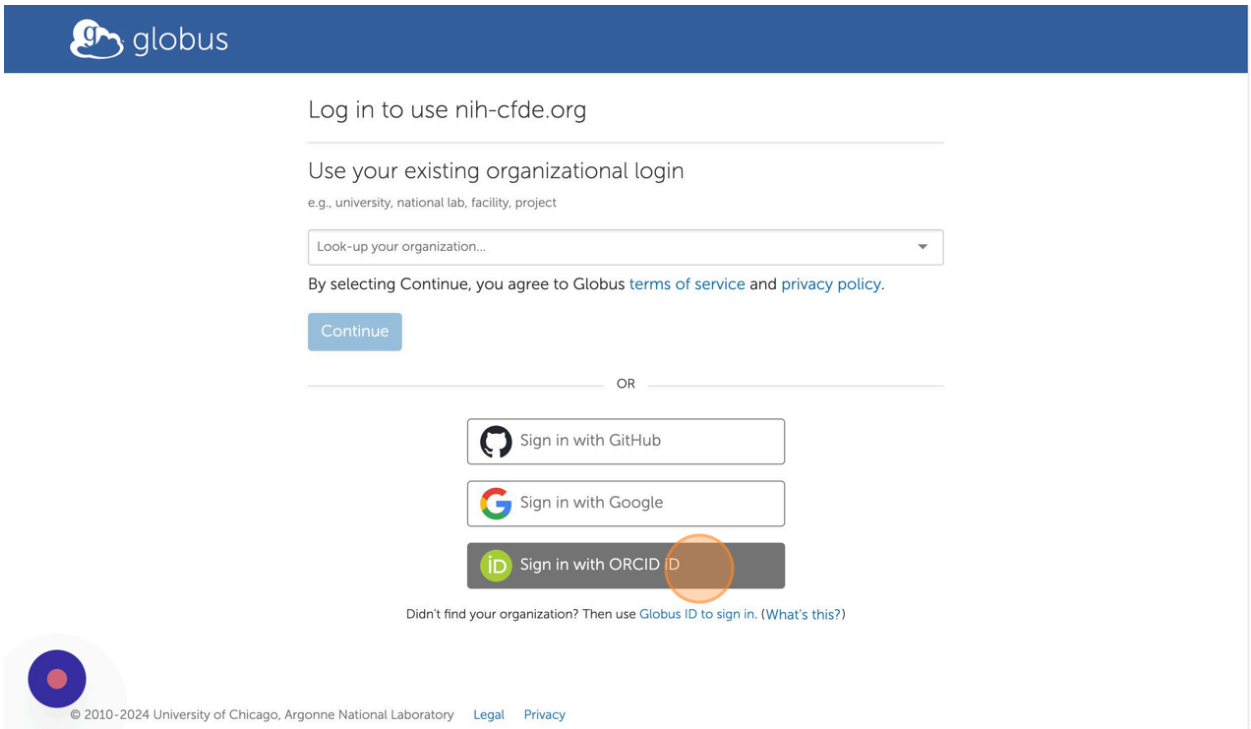
Once you are successfully connected, you can navigate back to KidsFirst portal Dashbord, where you will also see an option to create new CAVATICA project dynamically.

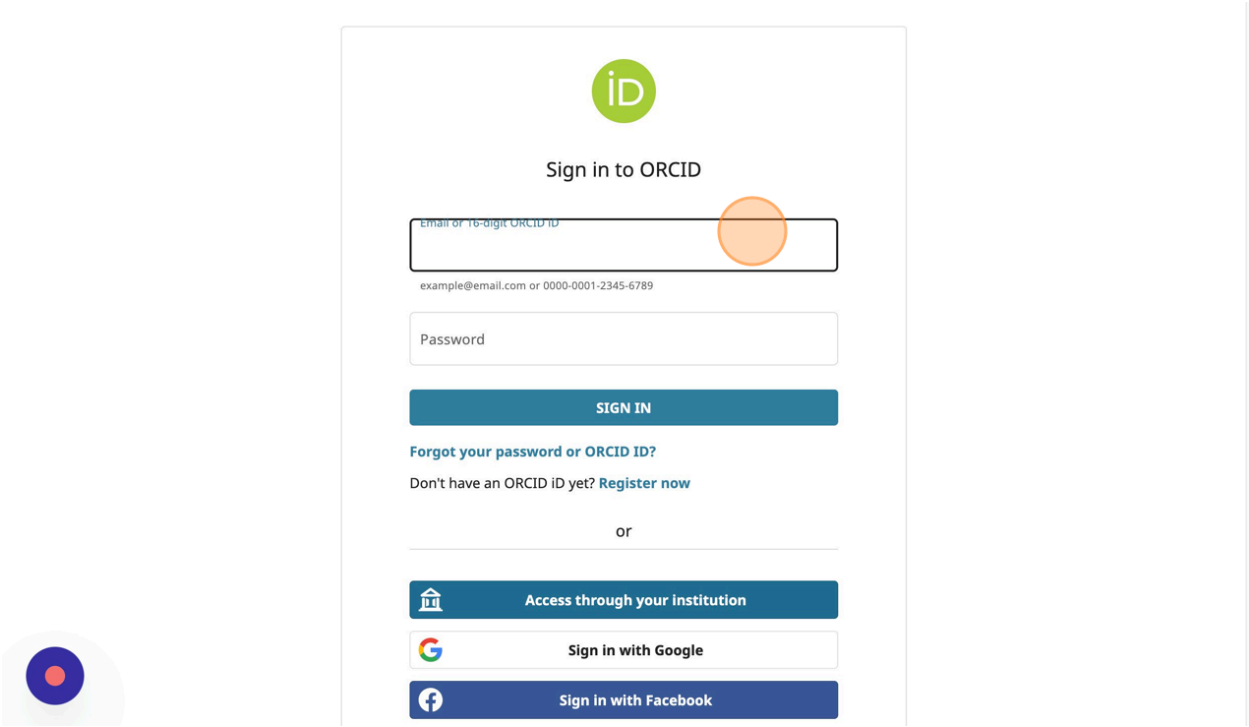


Now, to import data from CFDE portal, navigate to nih-cfde.org and click “Log In” to access the portal.

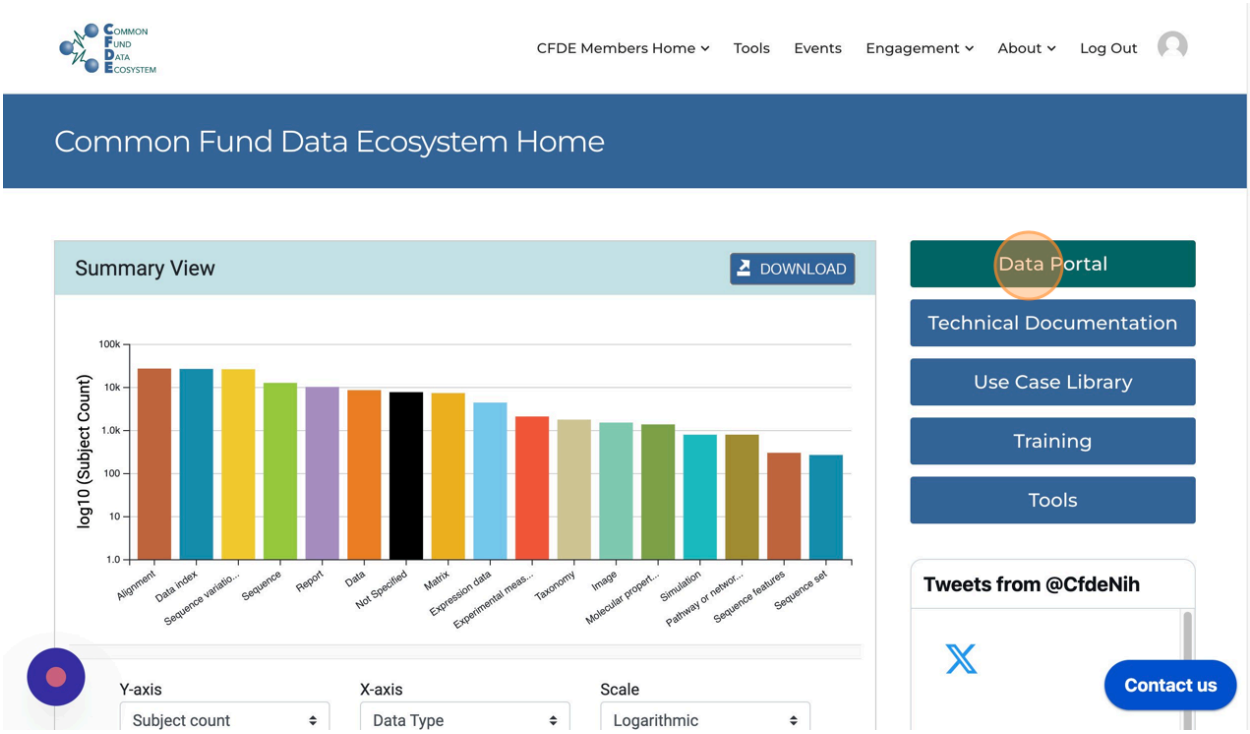


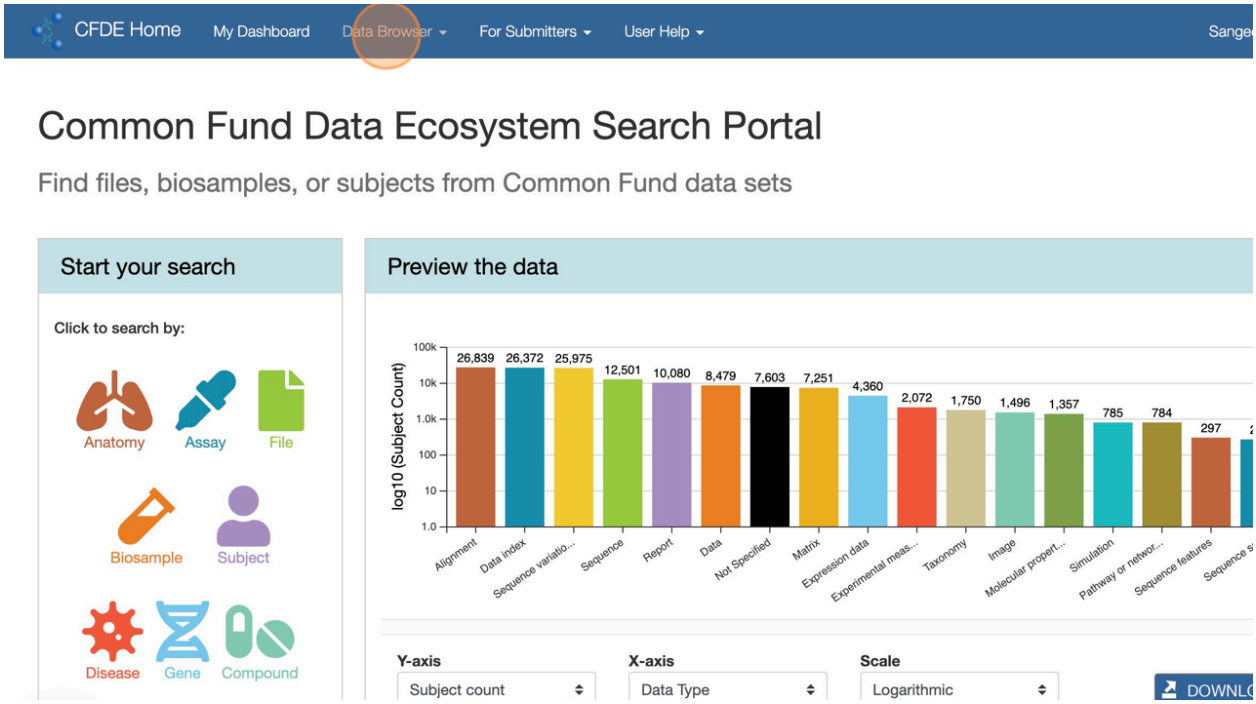
Sign in with ORCID ID



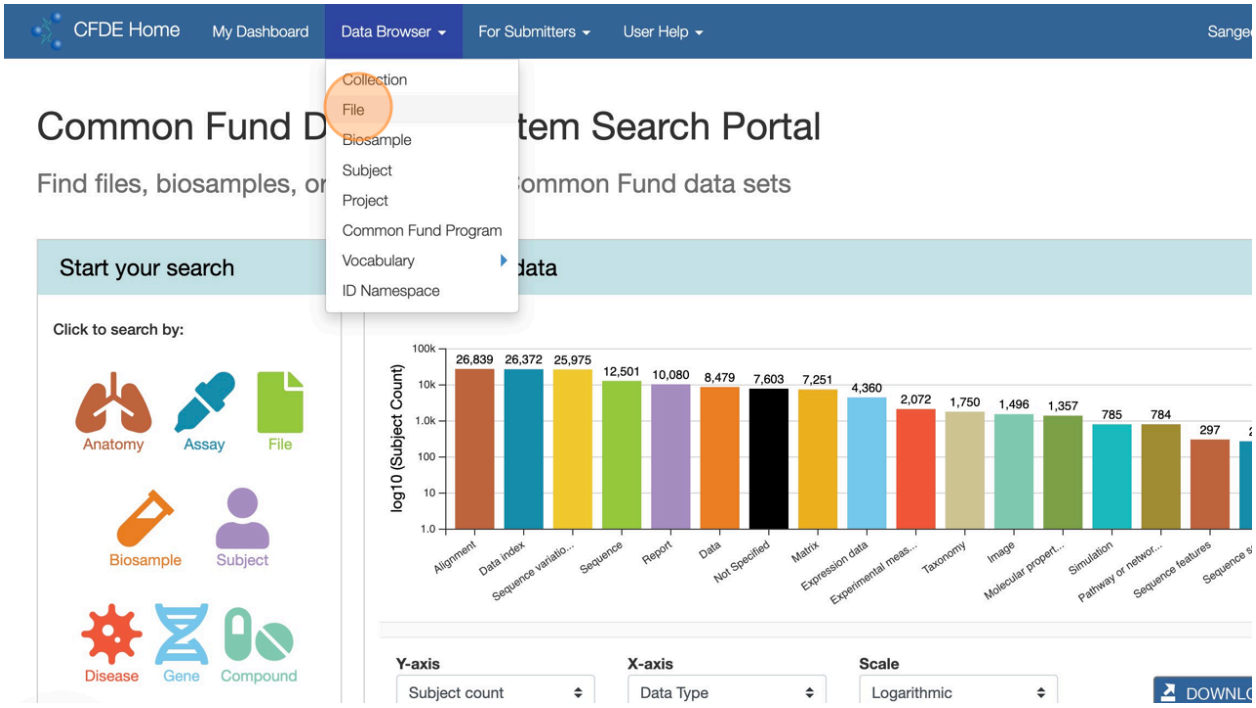


Once signed in, you can browse through available data sets via 'Data Portal'

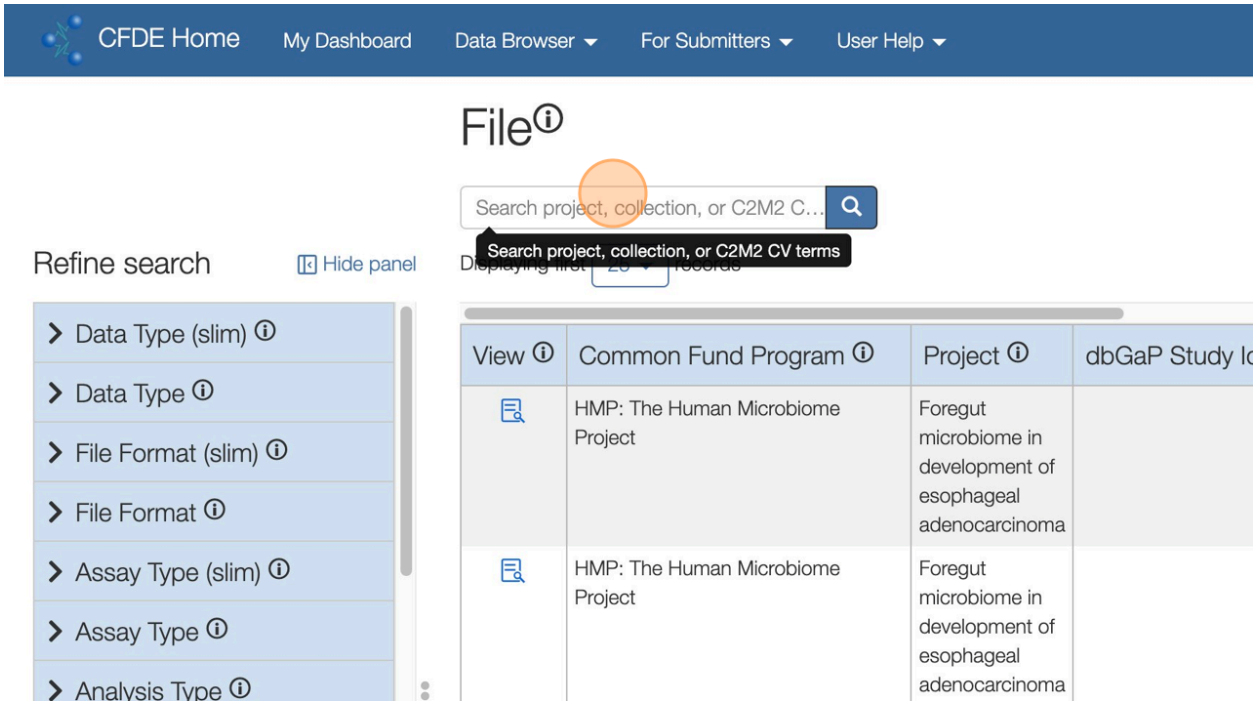


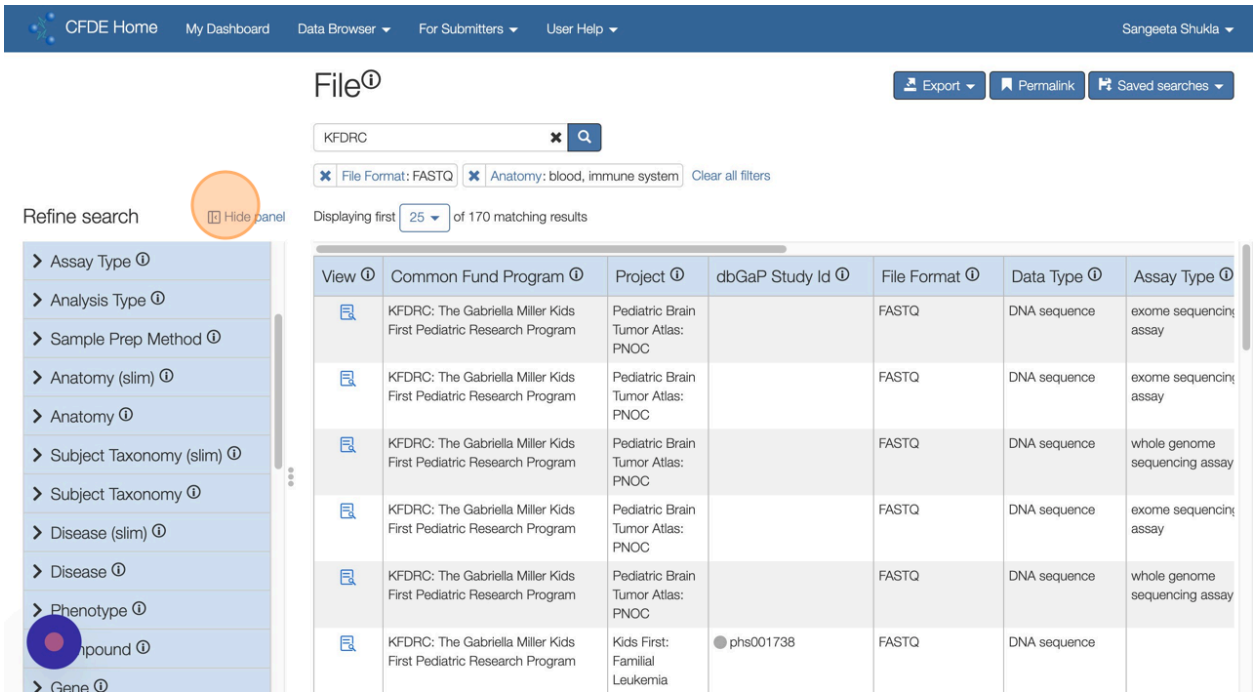


Within the 'Data Browser', you can search for appropriate project or individual file(s) for analysis.

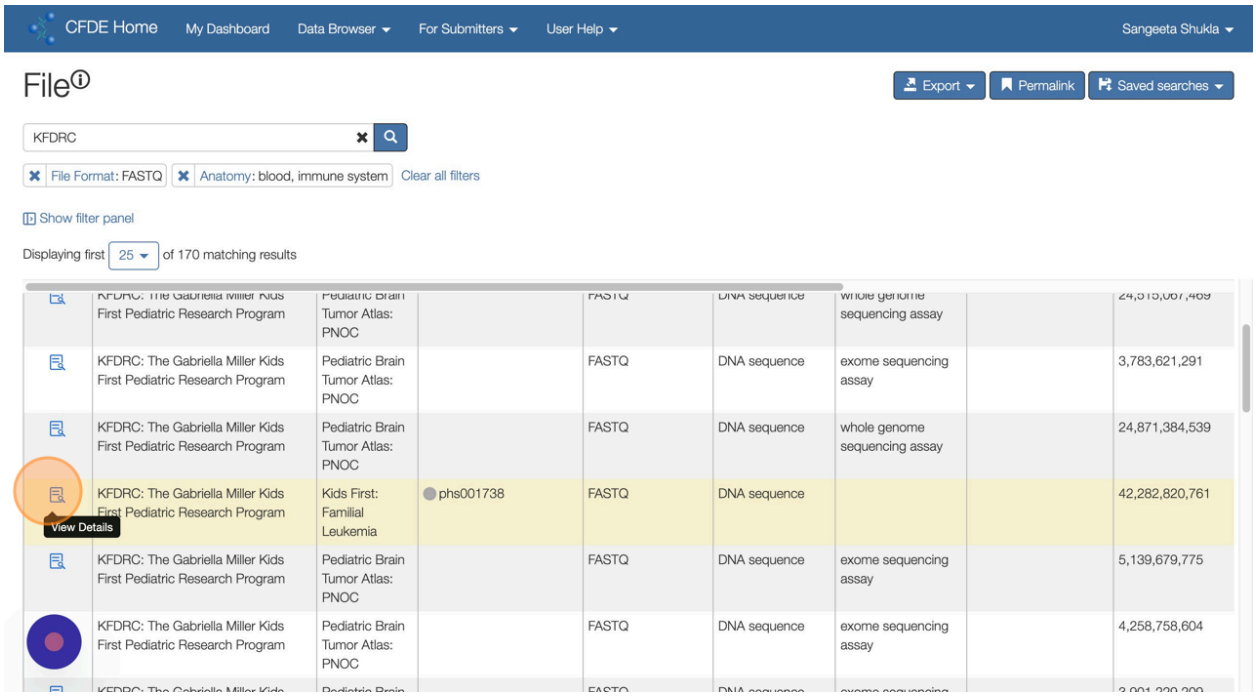


You may also refine your search for the dataset using different parameters.

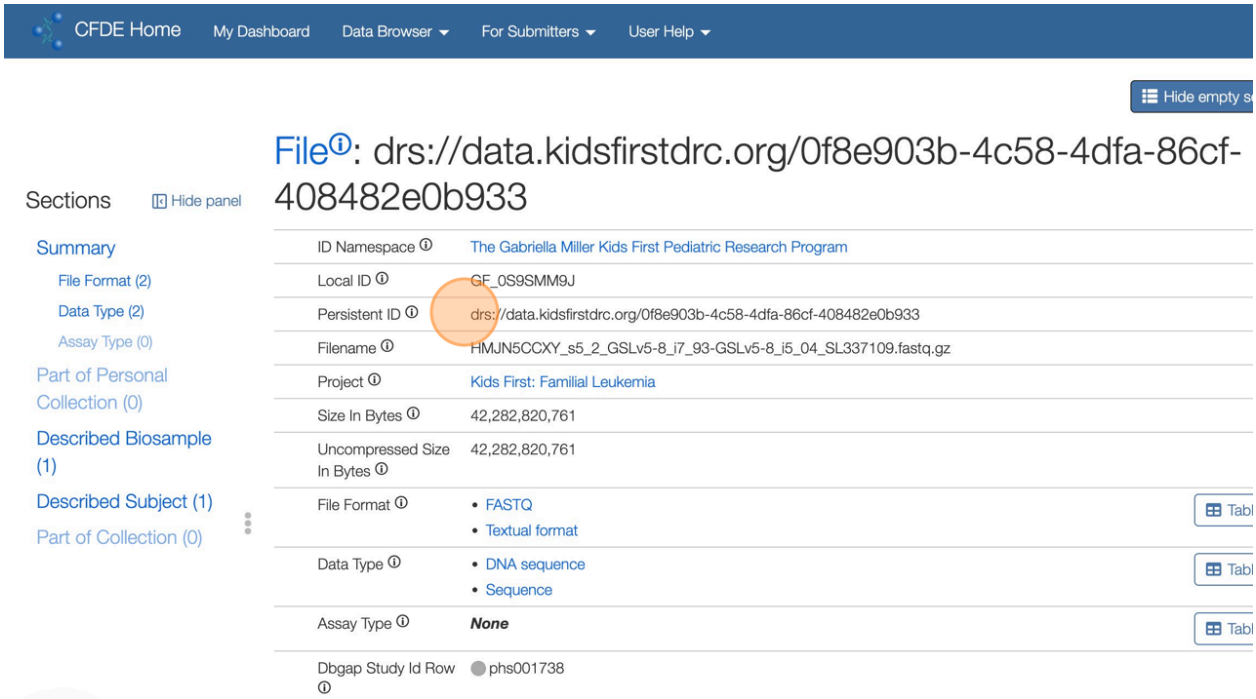




Clicking on 'View Details' for the selected dataset will show a summary, including a DRS link, which can be used on platforms such as CAVATICA, to access the data files, without having to download them or storing them locally.

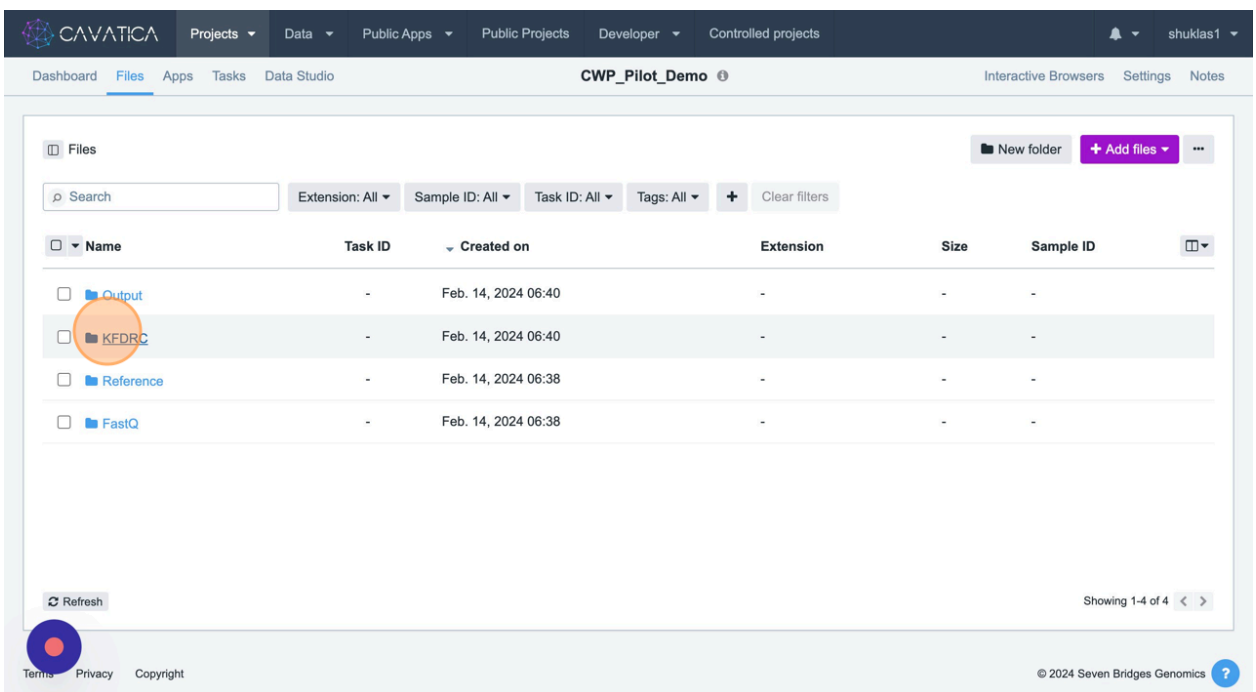


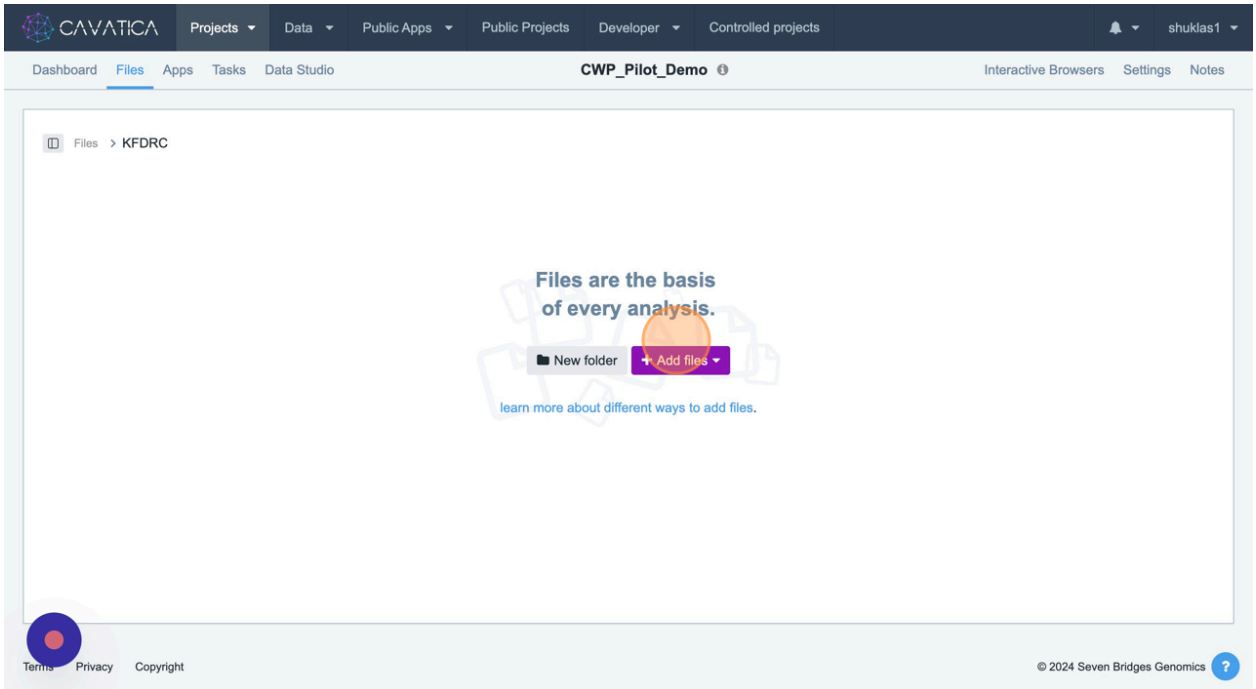
Copy that DRS link and navigate back to the CAVATICA platform, and into the project created earlier.



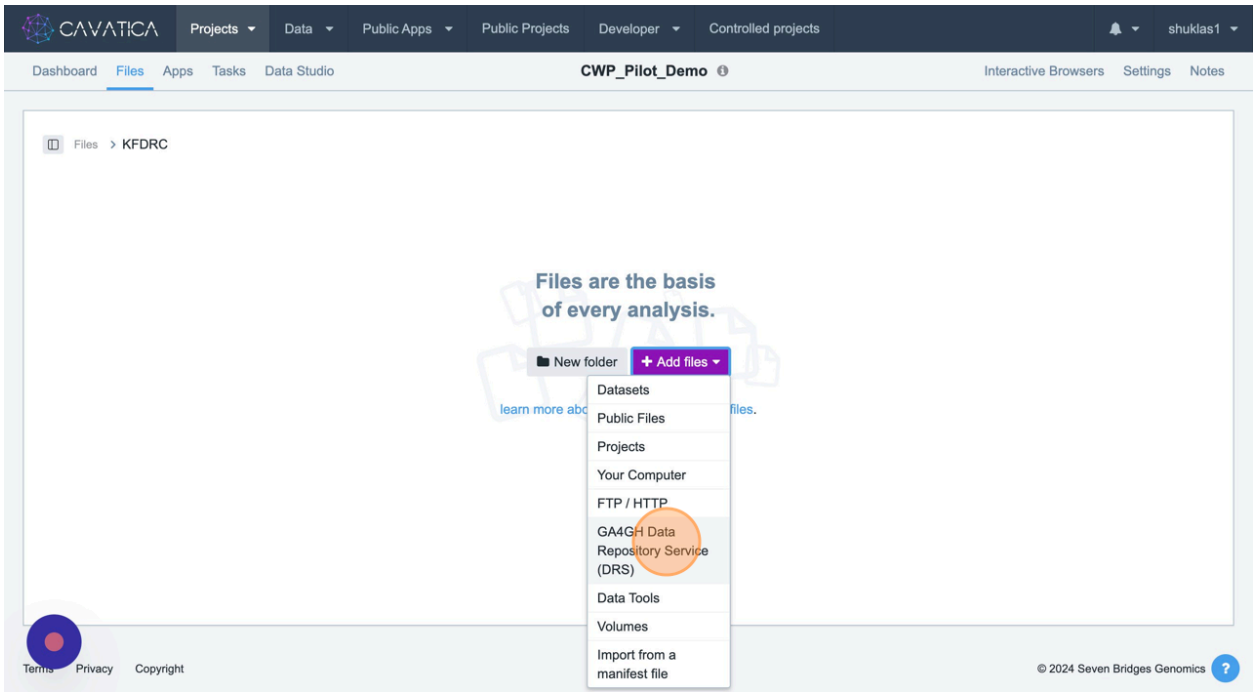
Import the files into the CAVATICA project, using the DRS link by following this guide, outlined below.

Click on “Files” within the project and navigate to the folder to which and click on “Add Files”. If other files exist within the project, you will also see them in the project.

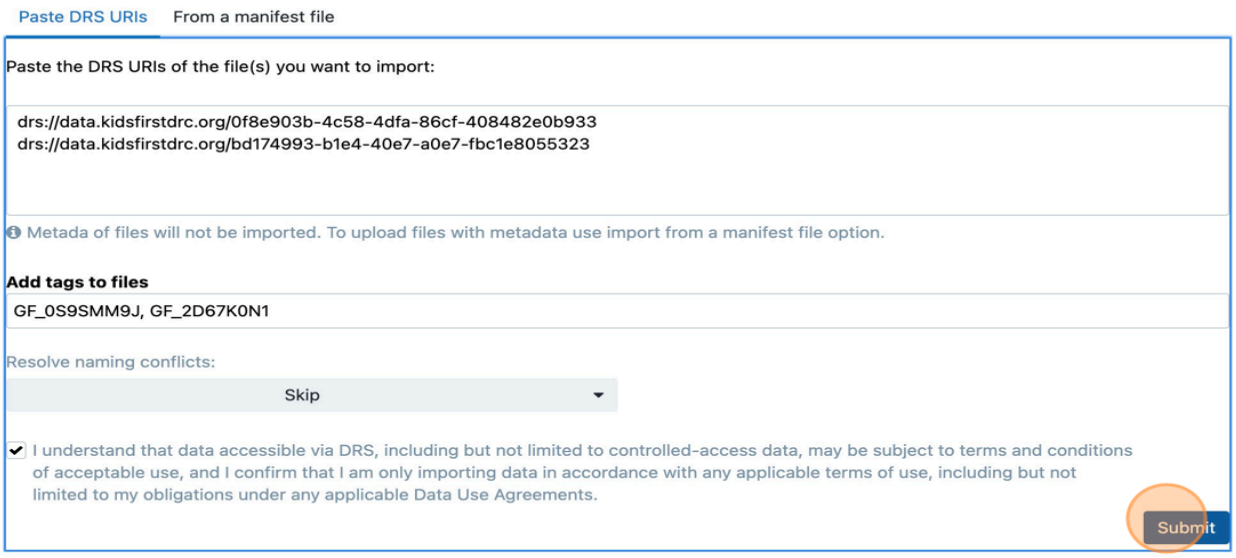




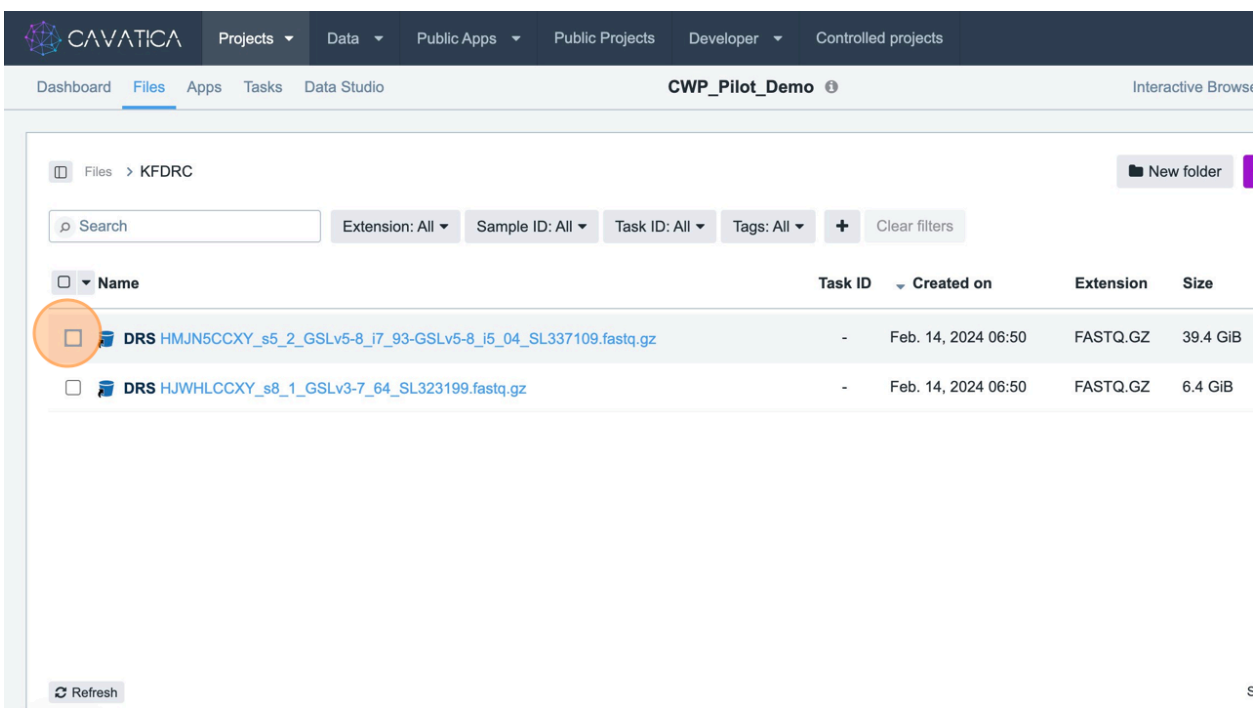
Use the "GA4GH Data Repository Service (DRS)" option, and paste the link(s) copied from the CFDE Data browser.



Add tag(s) for the file(s), and update other options as needed. Finally, click 'Submit'.



You will see the imported file under the 'Files' tab within the appropriate folder.

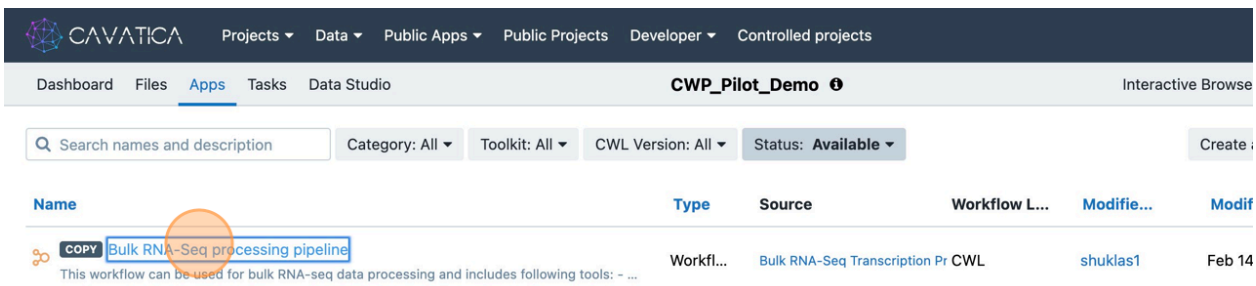


Run a CAVATICA application for Bioinformatics Analysis of data

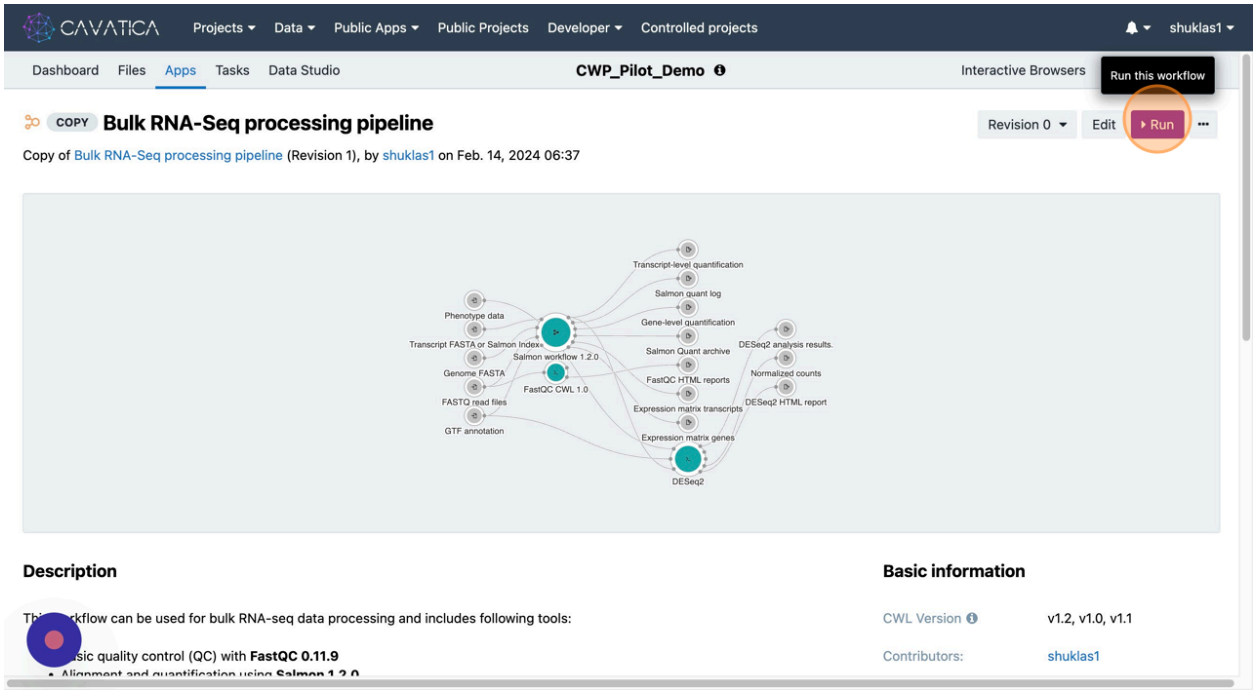
To implement an analysis workflow within the project, user can either import an existing workflow as we did earlier, or create a new one.

Now that we have the necessary data files, let us go ahead and run the analysis though the application.
Before importing and especially before running, take some time to read the documentation for the app, such as types of input and output files, and other parameters for processing.

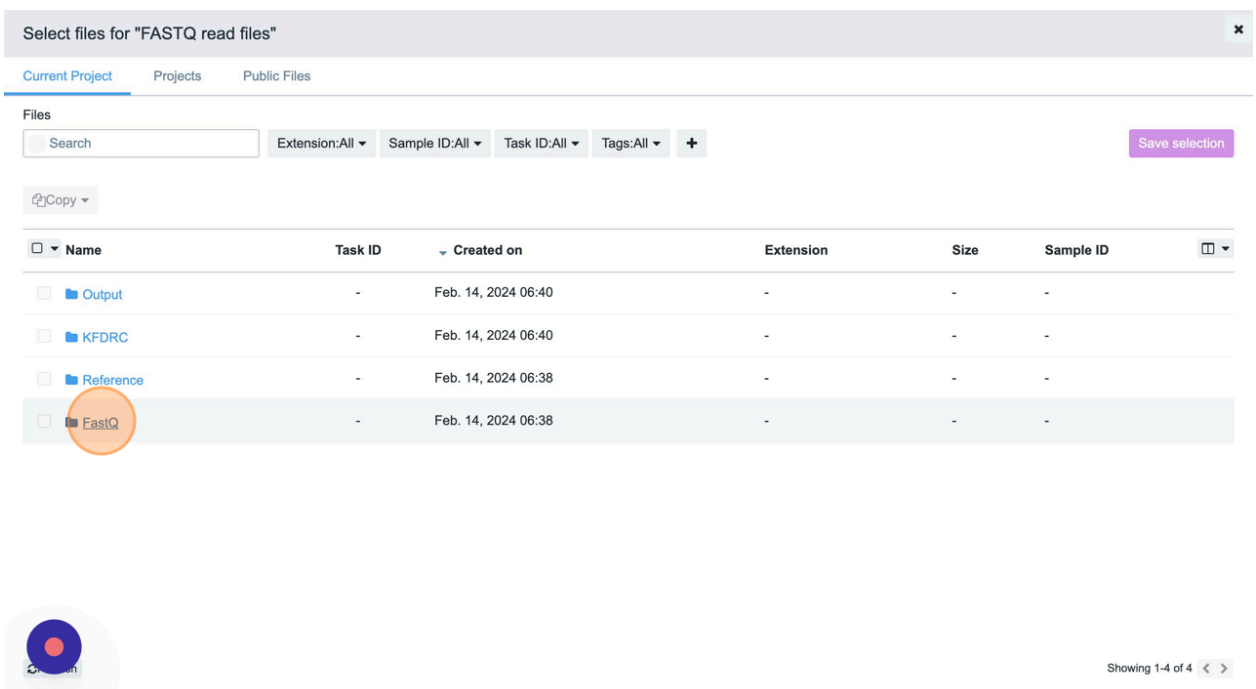
Begin with navigating to the Apps tab on the platform, and click on the app within the project.



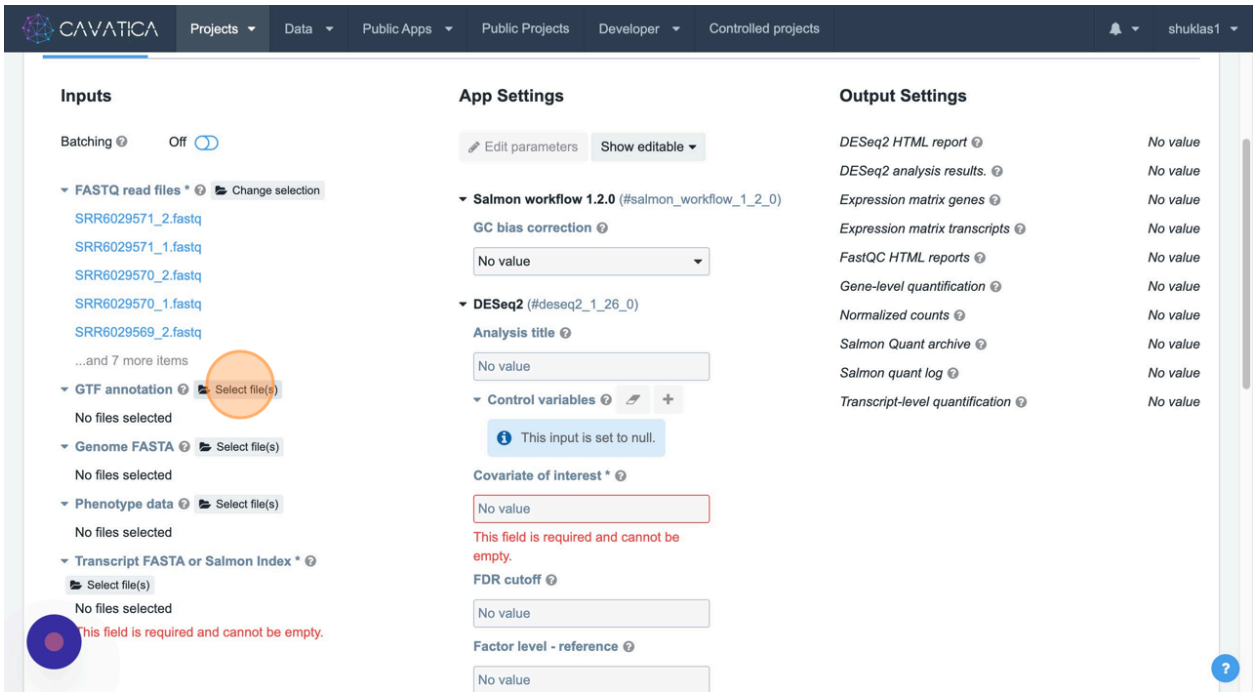
This will open the app workflow, giving the user a general idea of what steps wrapped within the CWL file.
Click on 'Run' and add parameters and link input files from within the project files or other public files within the platform.

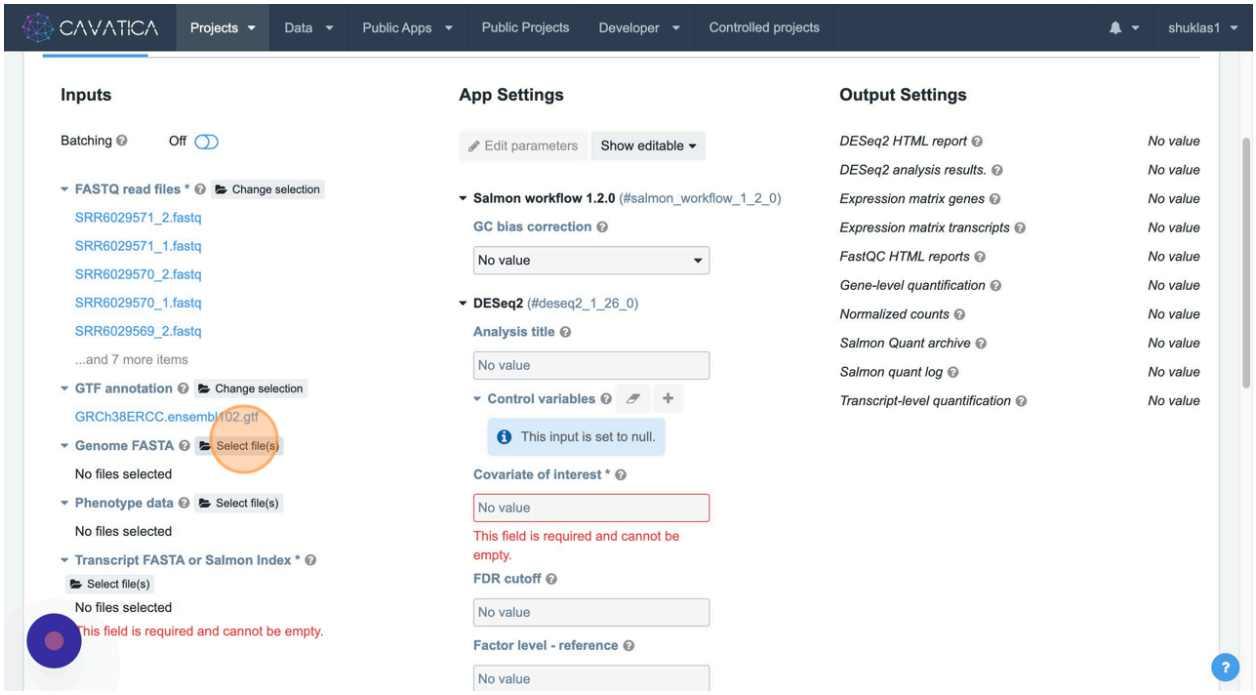


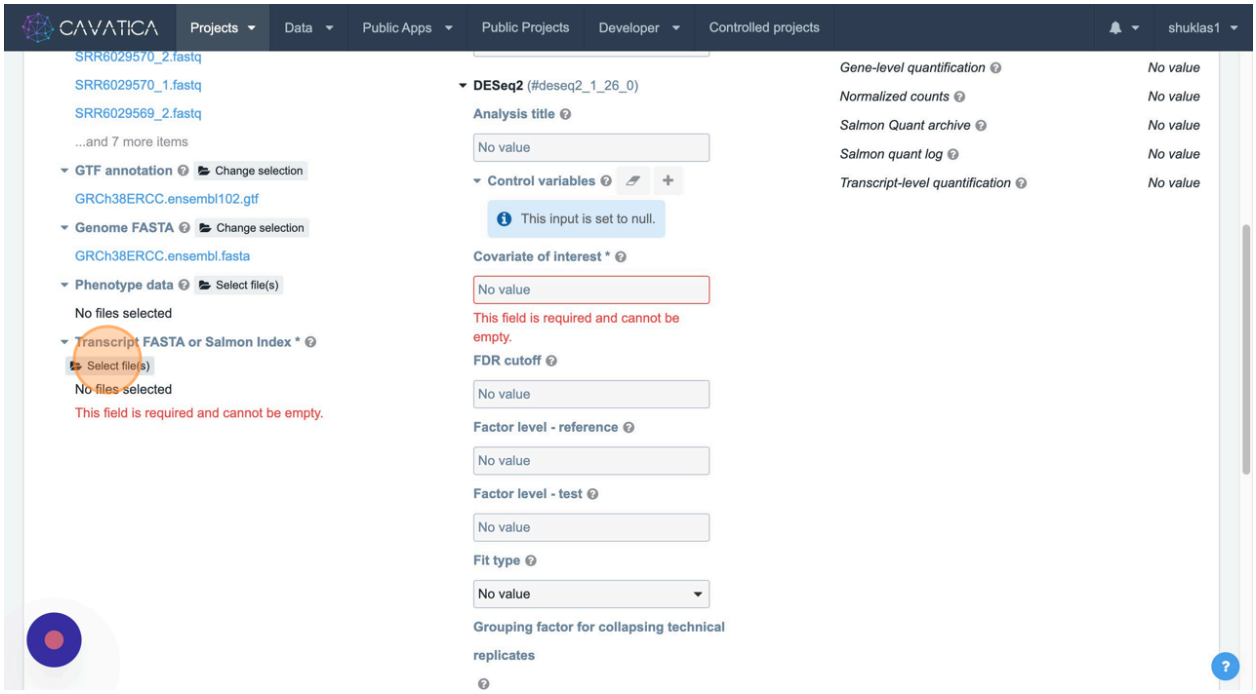
Here, we will select the FASTQ files imported from the original project for a faster implementation.

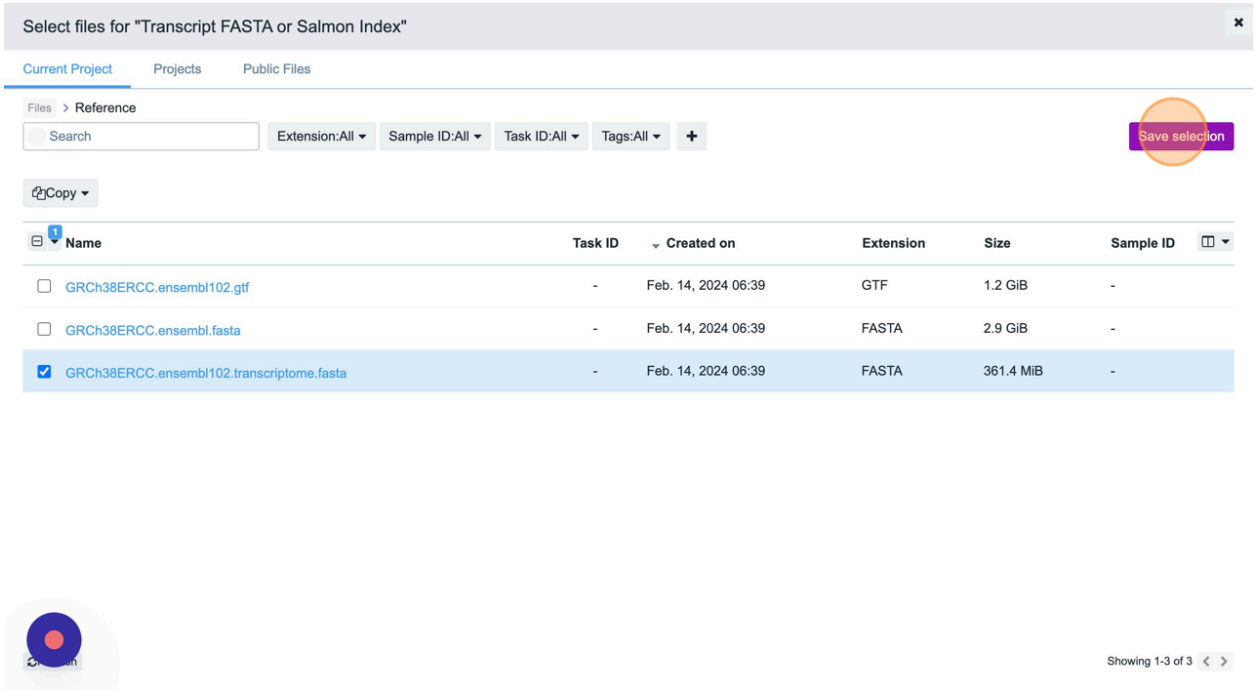


However, you can also use different input files of the same format.

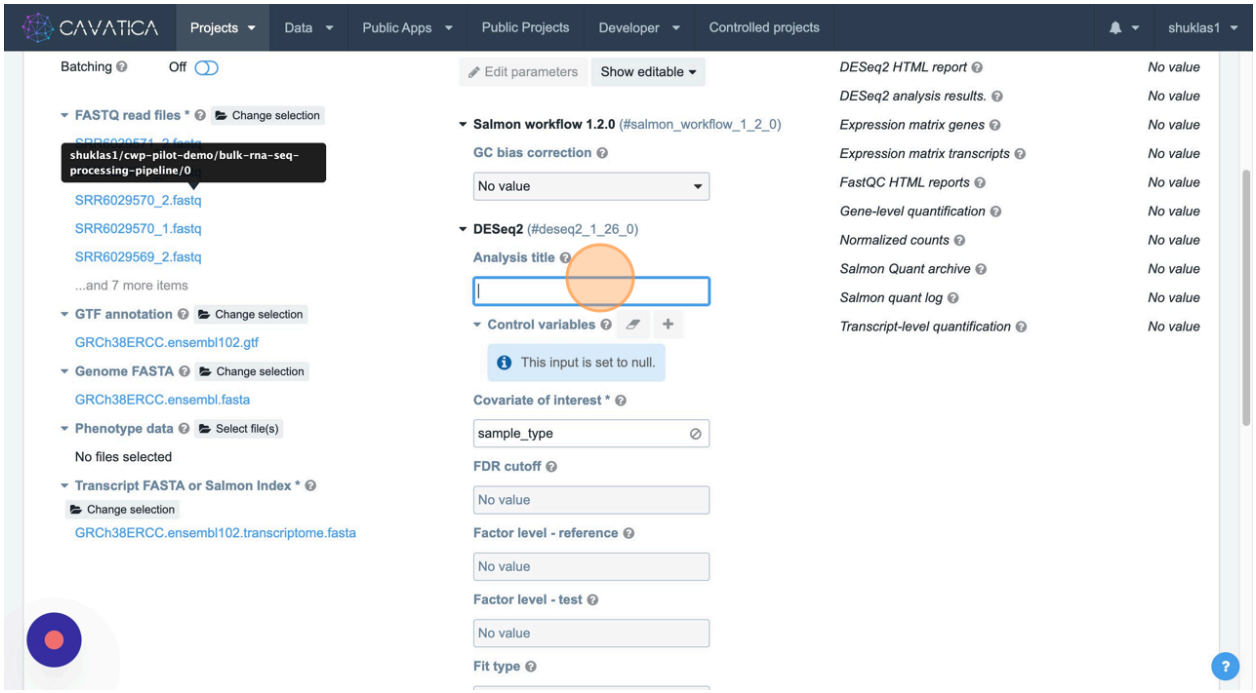

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Inputs
Batching
Off
FASTQ read files *
Change selection
SRR6029571_2.fastq
SRR6029571_1.fastq
SRR6029570_2.fastq
SRR6029570_1.fastq
SRR6029569_2.fastq
...and 7 more items
GTF annotation
Select file(s)
No files selected
Genome FASTA
Select file(s)
No files selected
Phenotype data
Select file(s)
No files selected
Transcript FASTA or Salmon Index *
Select file(s)
No files selected
This field is required and cannot be empty.
App Settings
Edit parameters
Show editable
Salmon workflow 1.2.0 (#salmon_workflow_1_2_0)
GC bias correction
No value
DESeq2 (#deseq2_1_26_0)
Analysis title
No value
Control variables
This input is set to null.
Covariate of interest *
No value
This field is required and cannot be empty.
FDR cutoff
No value
Factor level - reference
No value
Output Settings
DESeq2 HTML report
DESeq2 analysis results.
Expression matrix genes
Expression matrix transcripts
FastQC HTML reports
Gene-level quantification
Normalized counts
Salmon Quant archive
Salmon quant log
Transcript-level quantification
No value


CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Inputs
Batching
Off
FASTQ read files *
Change selection
SRR6029571_2.fastq
SRR6029571_1.fastq
SRR6029570_2.fastq
SRR6029570_1.fastq
SRR6029569_2.fastq
...and 7 more items
GTF annotation
Change selection
GRCh38ERCC.ensembl102.gtf
Genome FASTA
Select file(s)
No files selected
Phenotype data
Select file(s)
No files selected
Transcript FASTA or Salmon Index *
Select file(s)
No files selected
This field is required and cannot be empty.
App Settings
Edit parameters
Show editable
Salmon workflow 1.2.0 (#salmon_workflow_1_2_0)
GC bias correction
No value
DESeq2 (#deseq2_1_26_0)
Analysis title
No value
Control variables
This input is set to null.
Covariate of interest *
No value
This field is required and cannot be empty.
FDR cutoff
No value
Factor level - reference
No value
Output Settings
DESeq2 HTML report
DESeq2 analysis results.
Expression matrix genes
Expression matrix transcripts
FastQC HTML reports
Gene-level quantification
Normalized counts
Salmon Quant archive
Salmon quant log
Transcript-level quantification
No value

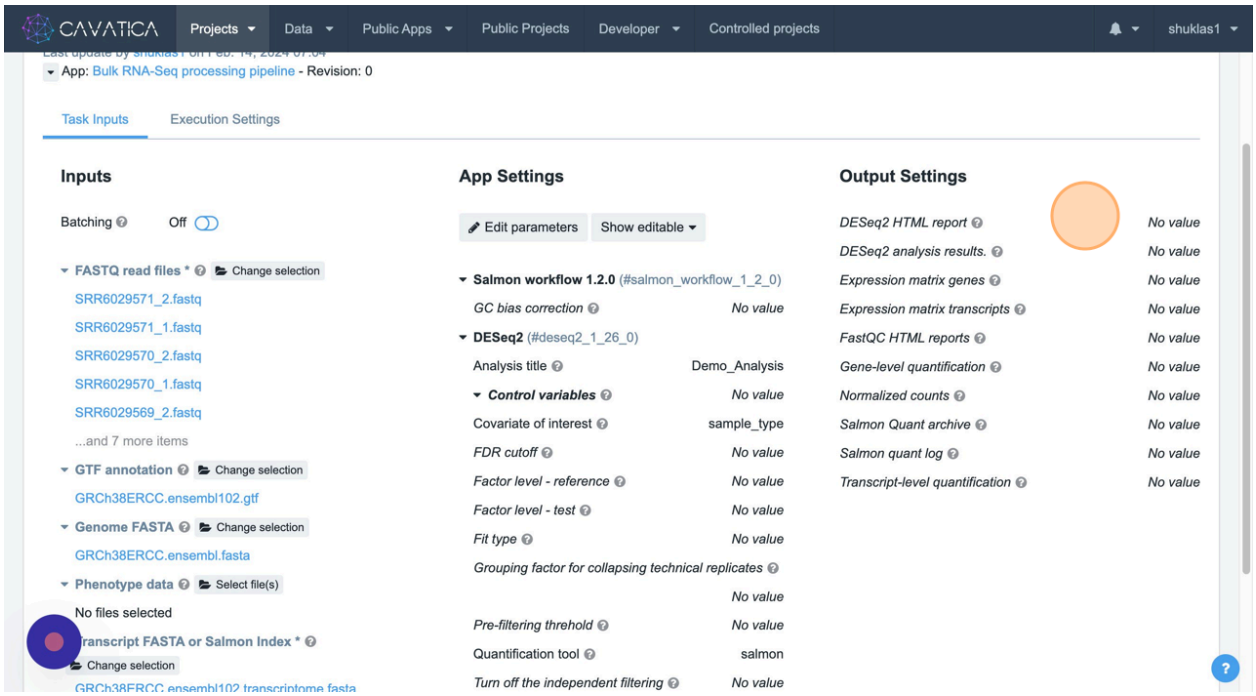




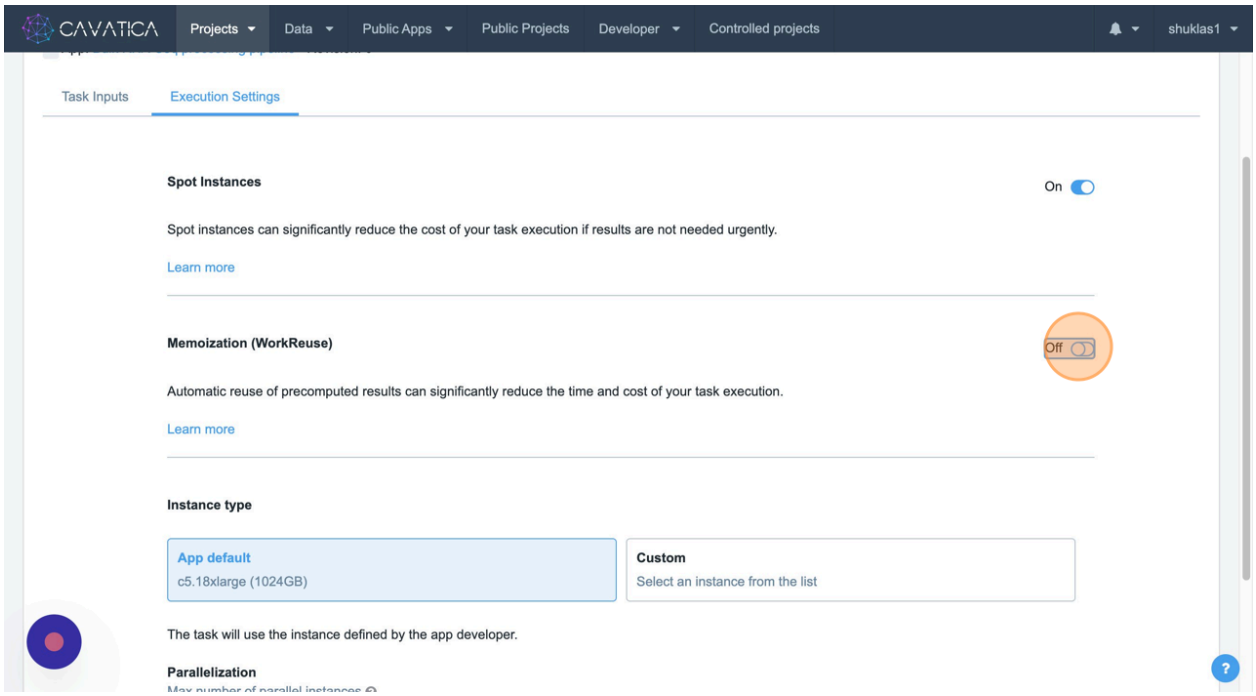
You can also edit the title of the analysis that is specific to the app.

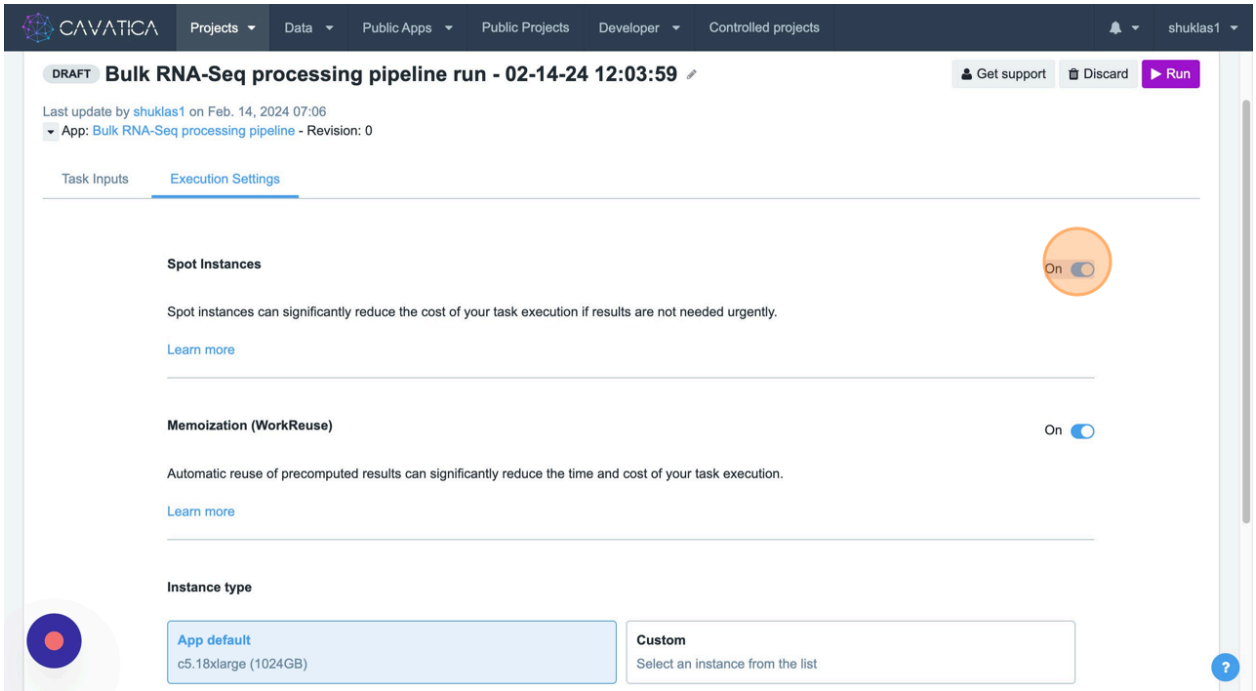


Be sure to include all mandatory parametric values and data files.

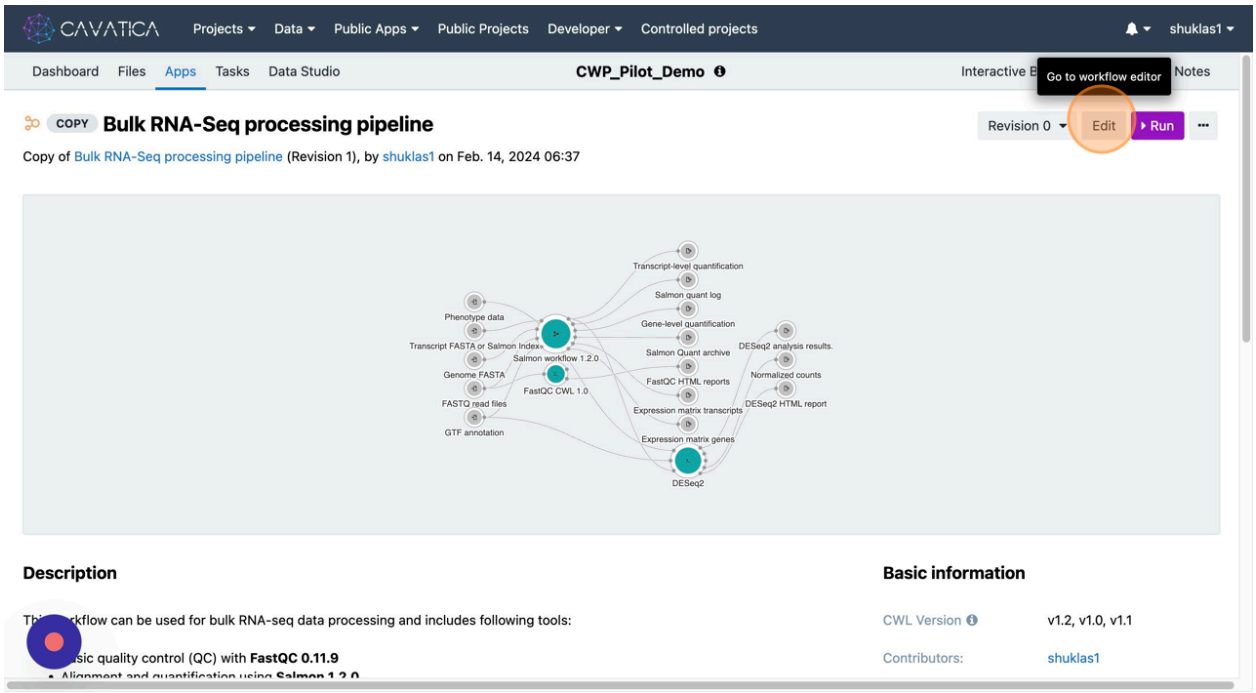


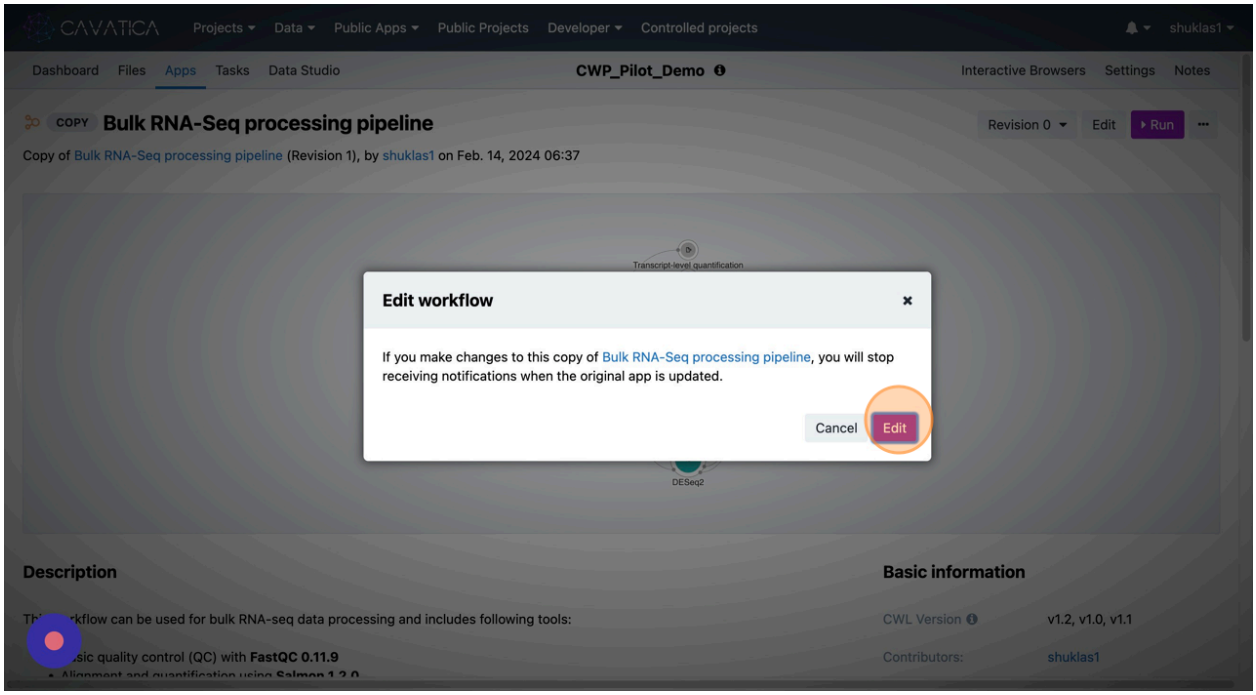
Also review ‘Execution Settings’ for improved efficiency and cost effectiveness.



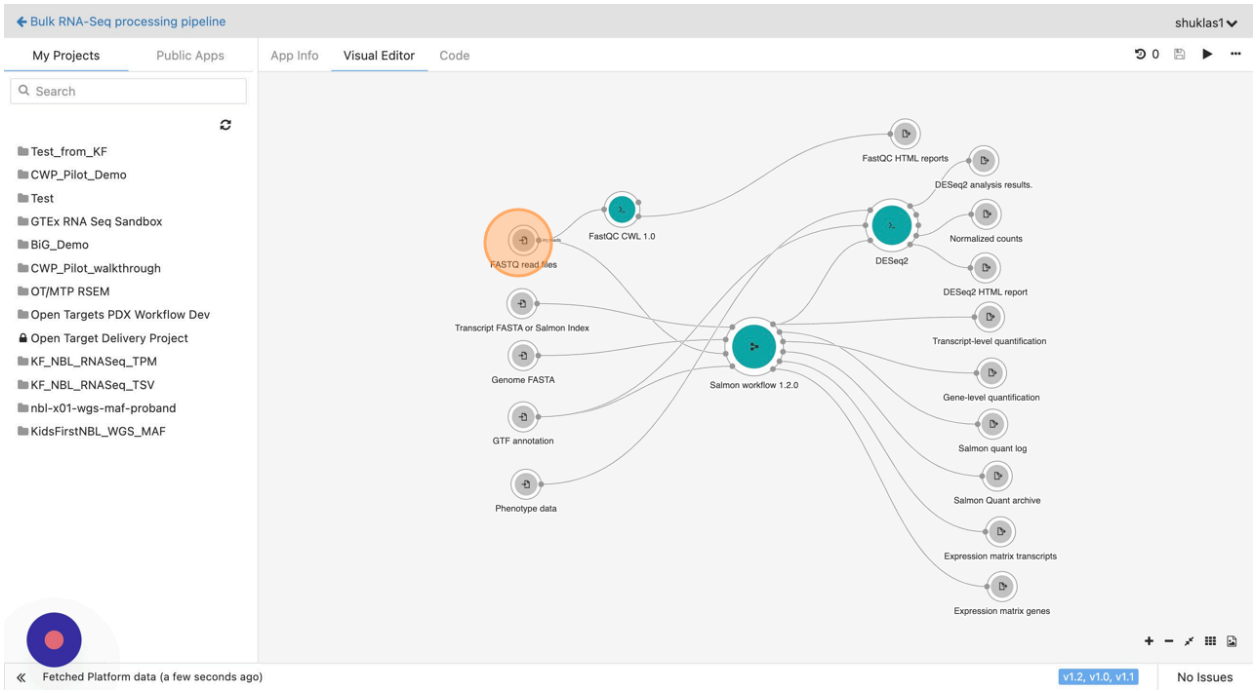


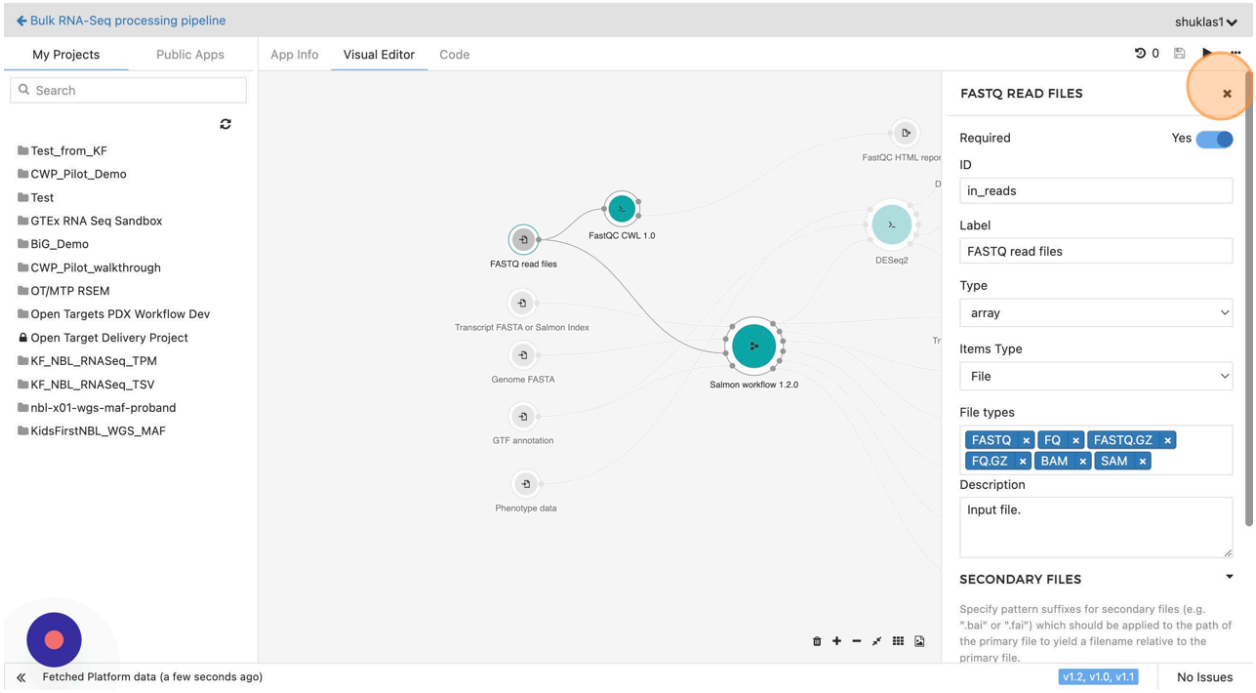
In case you want to edit and customize the application steps, you may do so. Click on the three dots next to the “Run” button, then click on “Edit”.

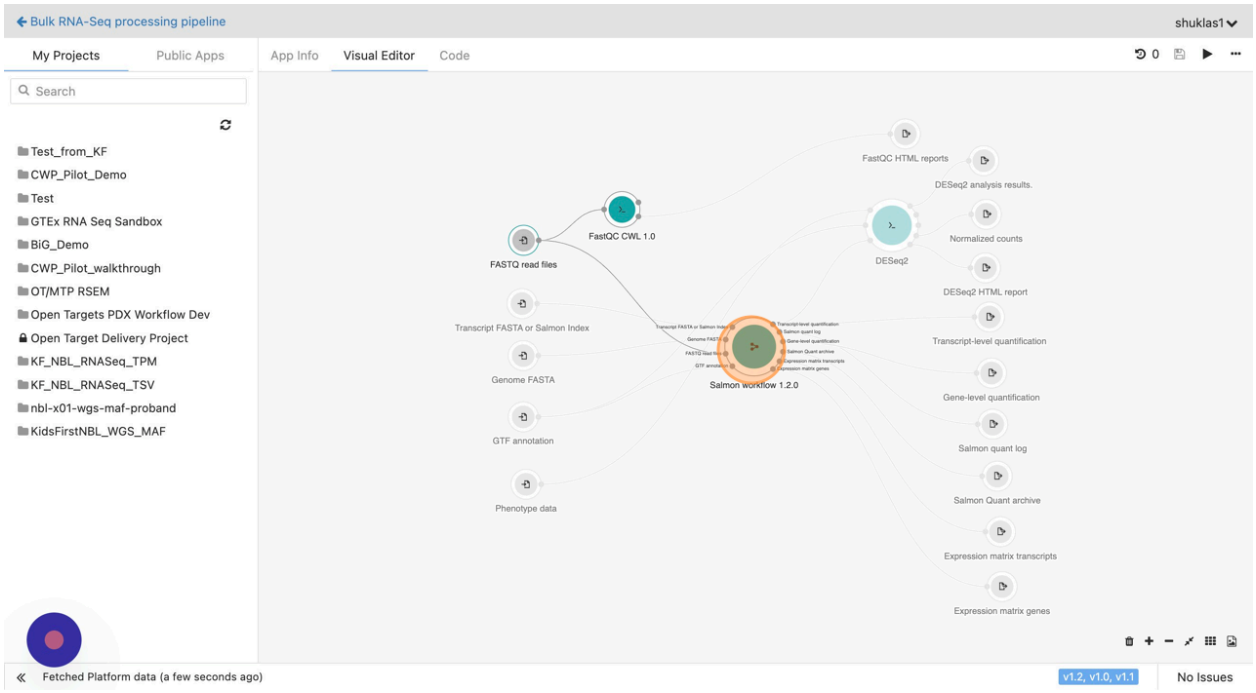




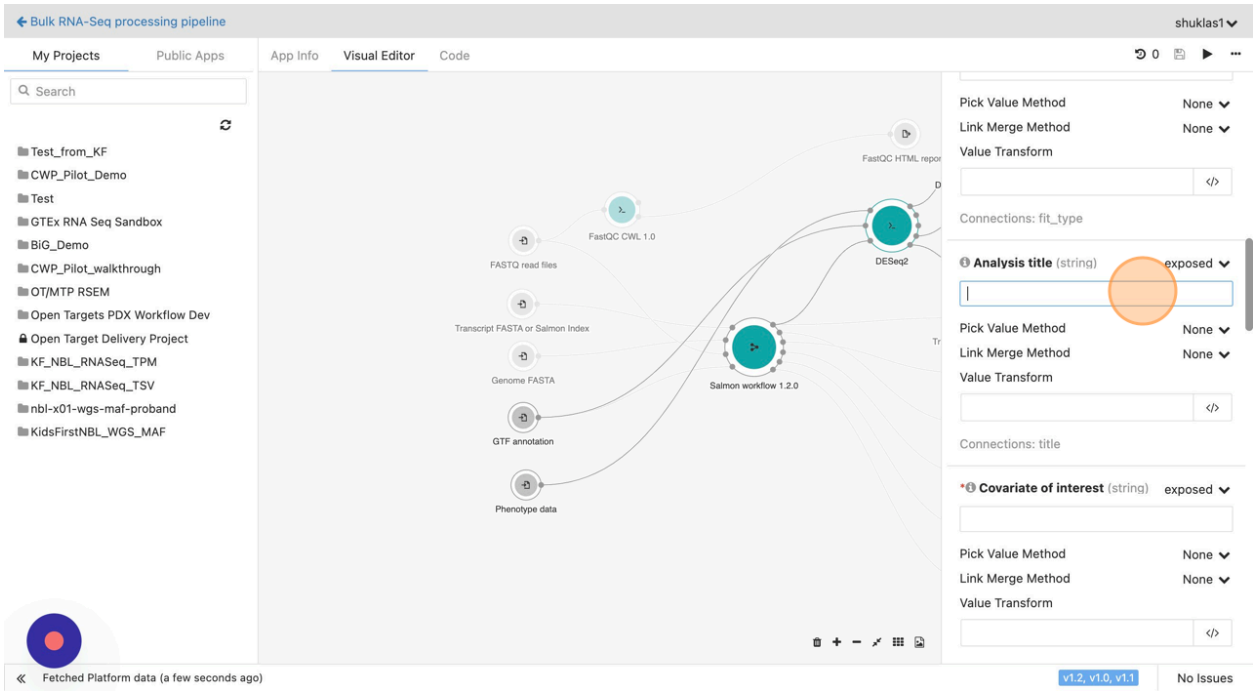
You can double-click on individual steps or blocks of the analytical pipeline to view details.

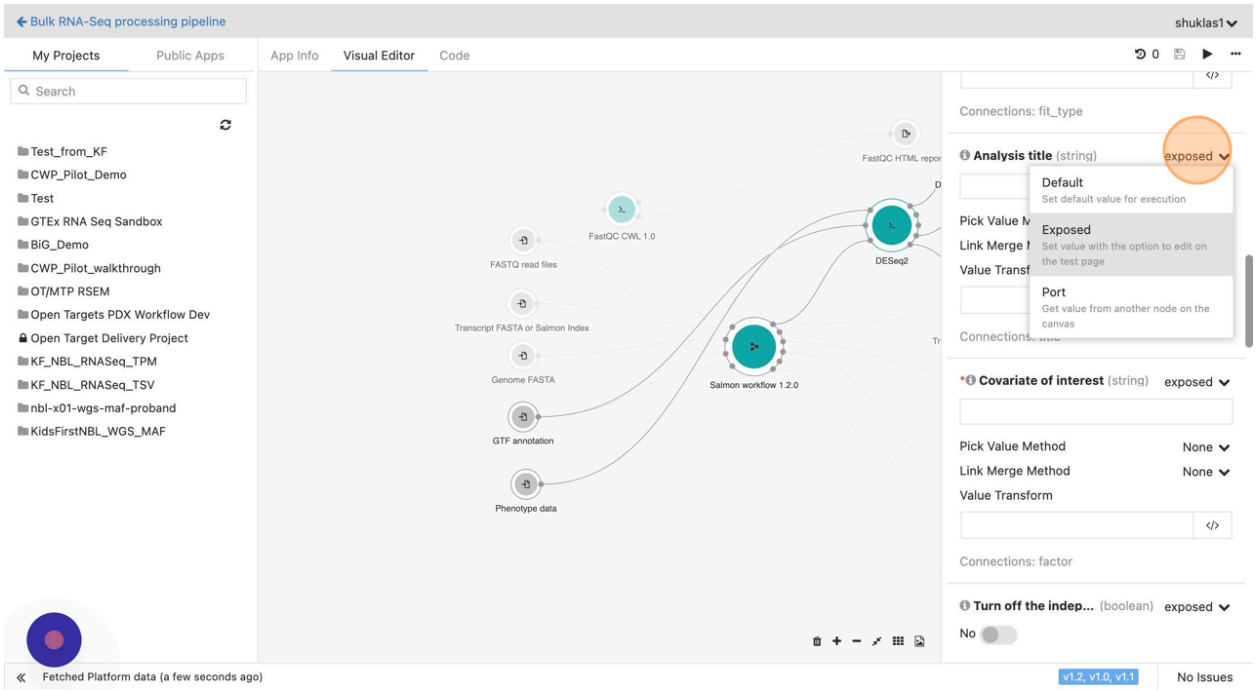

Bulk RNA-Seq processing pipeline
shuklas1
My Projects
Public Apps
App Info
Visual Editor
Code
Search
Test_from_KF
CWP_Pilot_Demo
Test
GTEx RNA Seq Sandbox
BiG_Demo
CWP_Pilot_walkthrough
OT/MTP RSEM
Open Targets PDX Workflow Dev
Open Target Delivery Project
KF_NBL_RNASeq_TPM
KF_NBL_RNASeq_TSV
nbl-x01-wgs-maf-proband
KidsFirstNBL_WGS_MAF
FASTQ read files
FastQC CWL 1.0
Transcript FASTA or Salmon Index
Genome FASTA
GTF annotation
Phenotype data
Salmon workflow 1.2.0
FastQC HTML reports
DESeq2 analysis results.
DESeq2
Normalized counts
DESeq2 HTML report
Transcript-level quantification
Gene-level quantification
Salmon quant log
Salmon Quant archive
Expression matrix transcripts
Expression matrix genes
Fetched Platform data (a few seconds ago)
v1.2, v1.0, v1.1
No Issues


Bulk RNA-Seq processing pipeline
shuklas1
My Projects
Public Apps
App Info
Visual Editor
Code
Search
Test_from_KF
CWP_Pilot_Demo
Test
GTEx RNA Seq Sandbox
BiG_Demo
CWP_Pilot_walkthrough
OT/MTP RSEM
Open Targets PDX Workflow Dev
Open Target Delivery Project
KF_NBL_RNASeq_TPM
KF_NBL_RNASeq_TSV
nbl-x01-wgs-maf-proband
KidsFirstNBL_WGS_MAF
FASTQ read files
FastQC CWL 1.0
Transcript FASTA or Salmon Index
Genome FASTA
GTF annotation
Phenotype data
Salmon workflow 1.2.0
DESeq2
FASTQ READ FILES
Required
Yes
ID
in_reads
Label
FASTQ read files
Type
array
Items Type
File
File types
FASTQ
FQ
FASTQ.GZ
FQ.GZ
BAM
SAM
Description
Input file.
SECONDARY FILES
Specify pattern suffixes for secondary files (e.g. ".bai" or ".fai") which should be applied to the path of the primary file to yield a filename relative to the primary file.
Fetched Platform data (a few seconds ago)
v1.2, v1.0, v1.1
No Issues

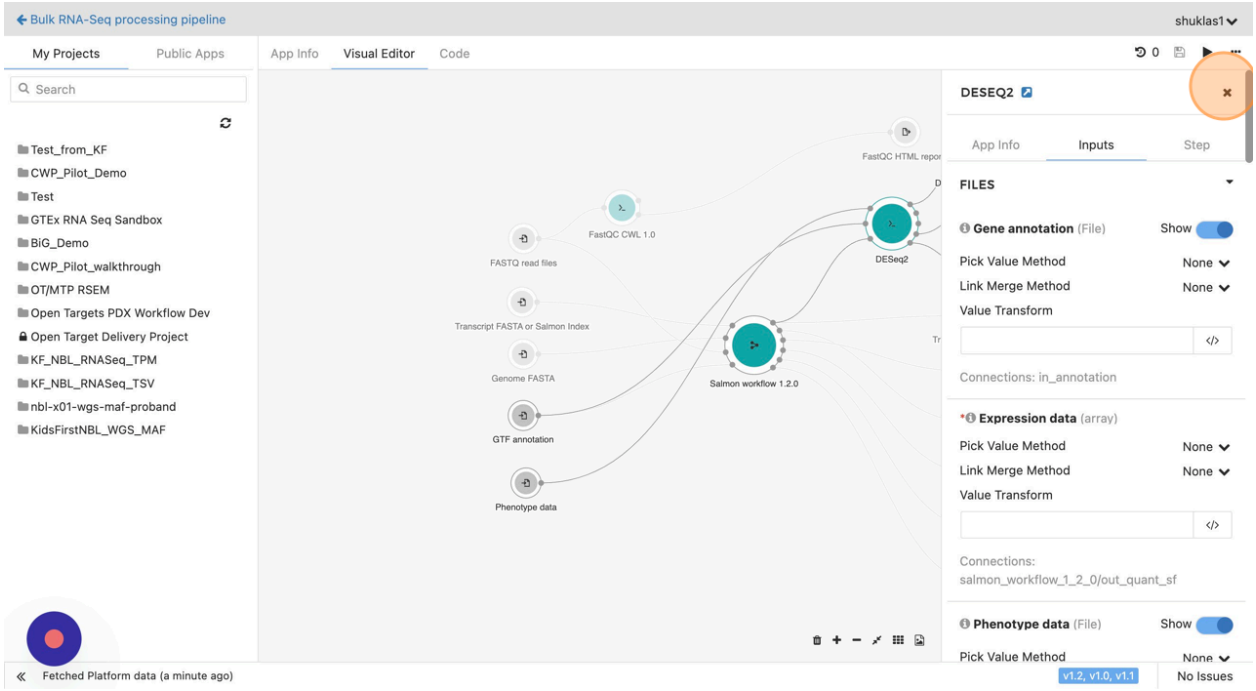


You can also specify parametric values here, and/or change settings for individual steps.

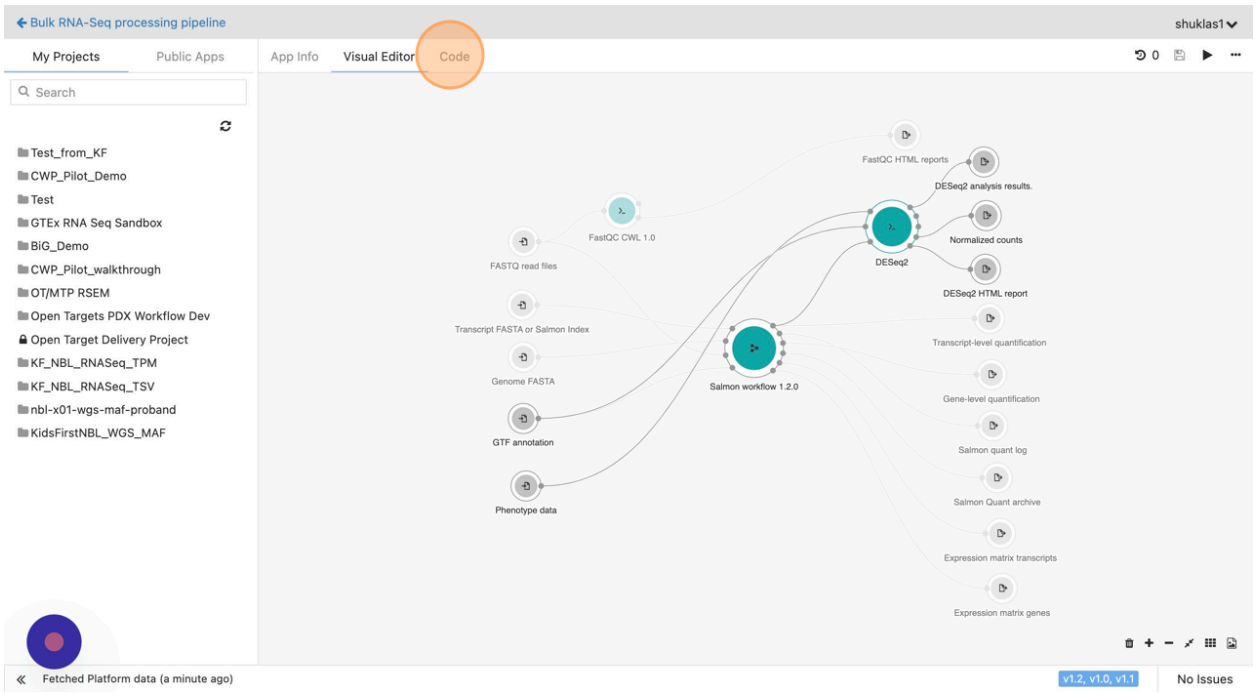


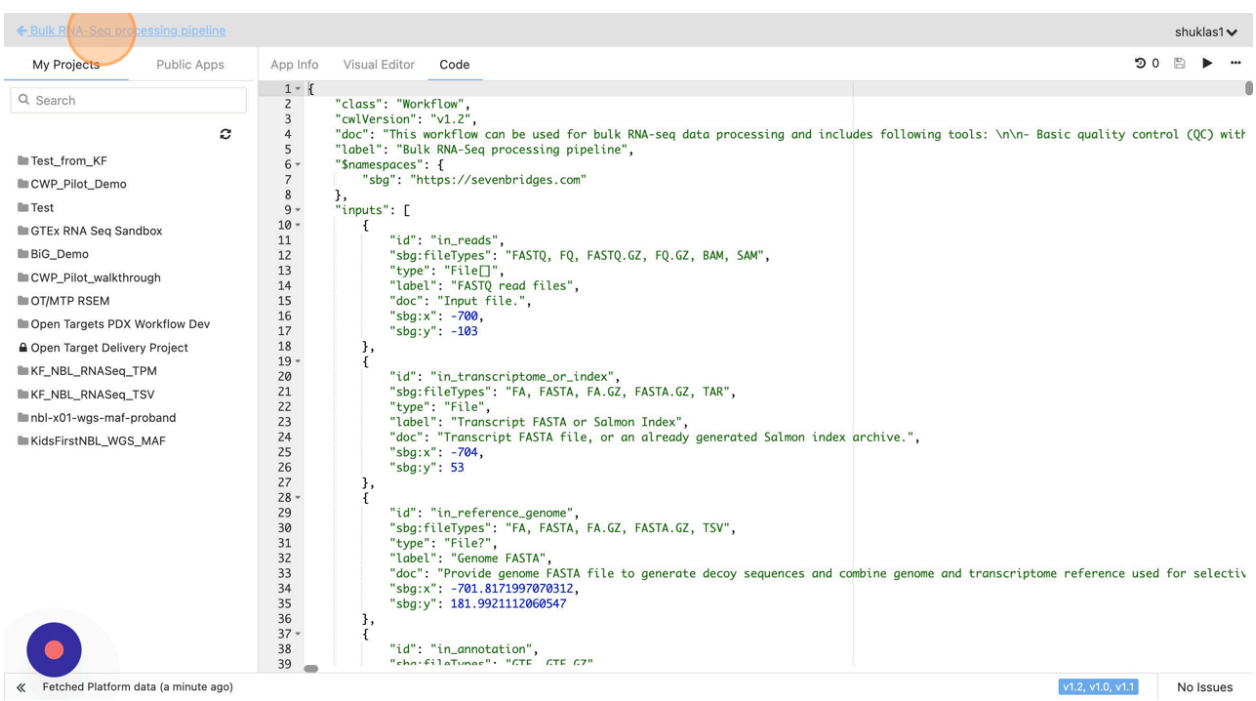


Click to close the editor when you are done reviewing.

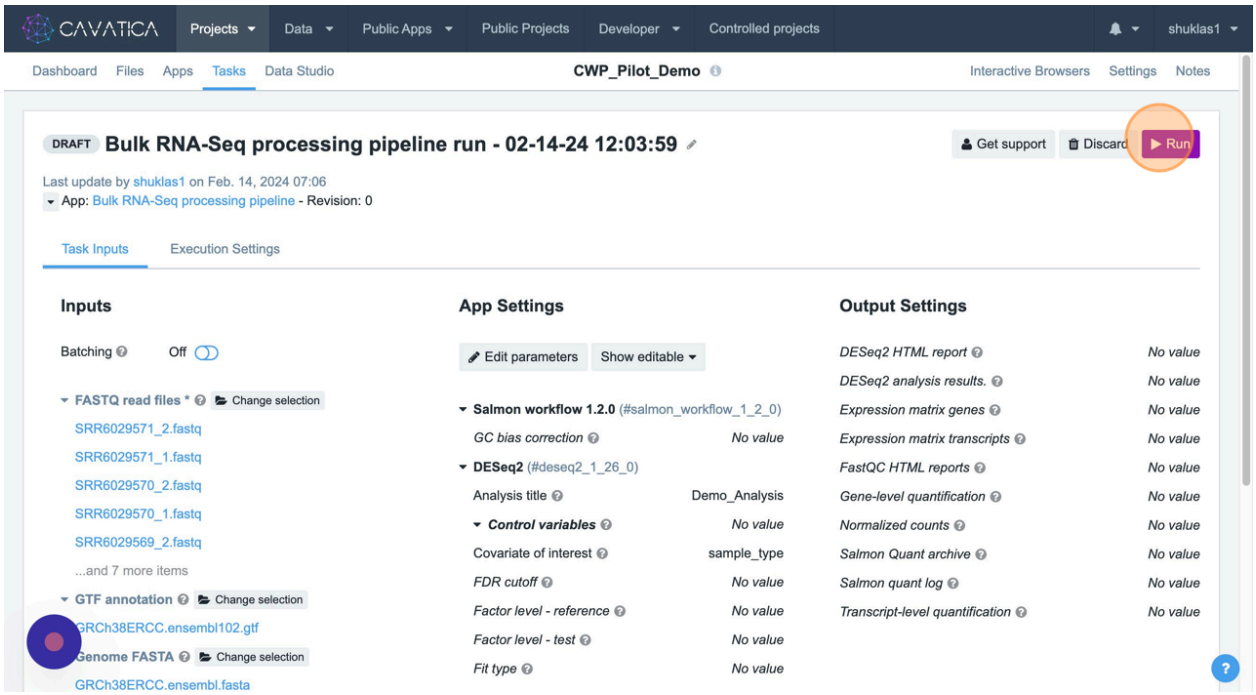


The ‘Visual Editor’ tab shows a summary of the app including information such as reference to the Docker image for the CWL wrapper, arguments expected, et cetera. The platform also allows users to review/edit the app info and CWL script code in the tabs as seen below.





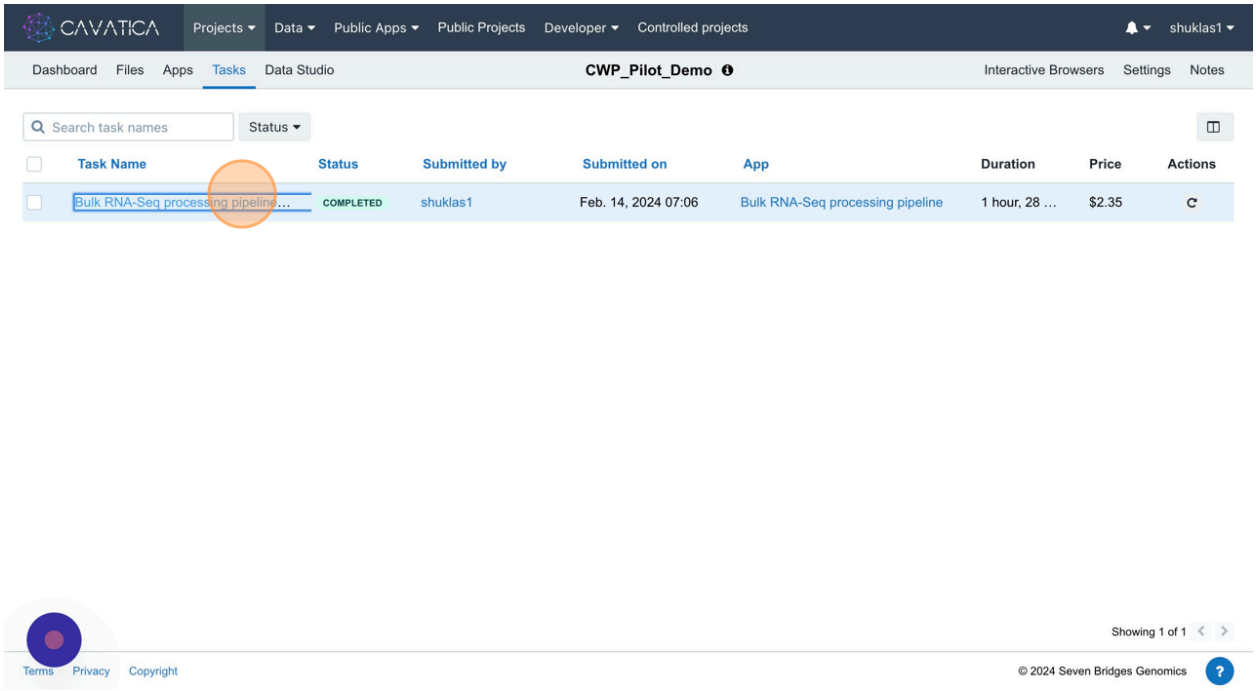
Finally, when everything looks set, click on ‘Run’.



It is also possible to stop a process if necessary. Just click on “Abort”.

While the app is queued and running, you may close the browser window and return later.

If/When the execution completes successfully, you will see the status on the project dashboard as below.

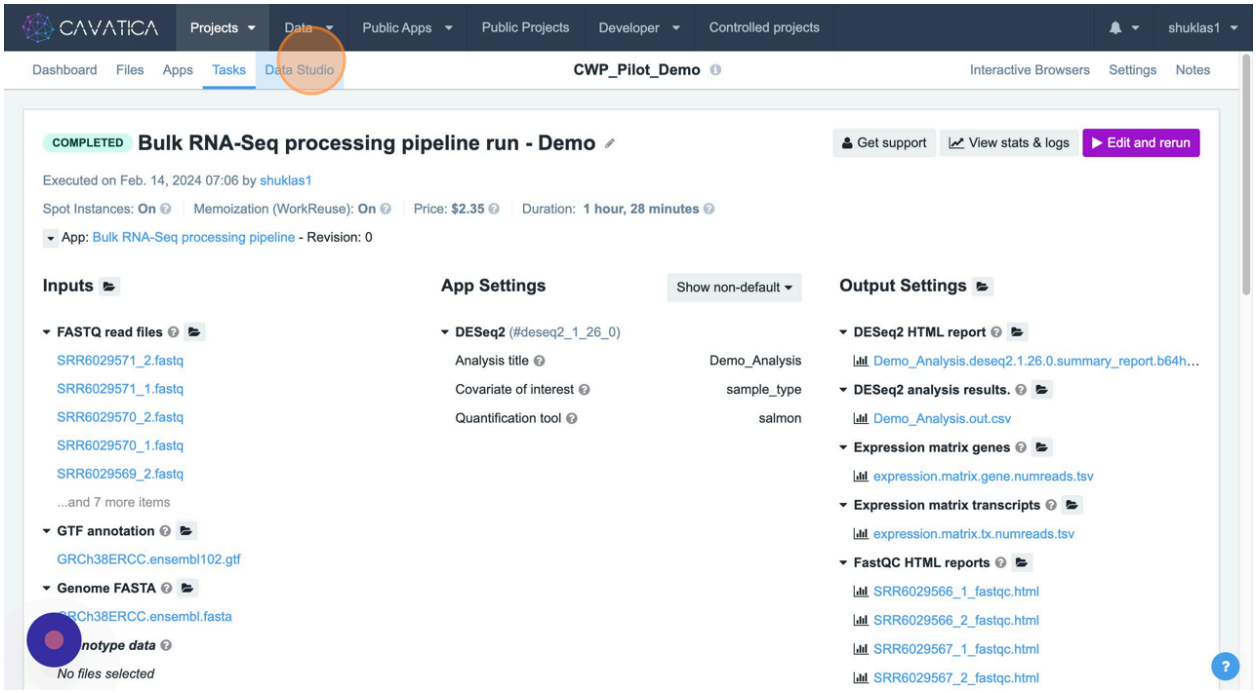


You can see that this analysis ran for about 1.5 hours and cost less than $2.5.

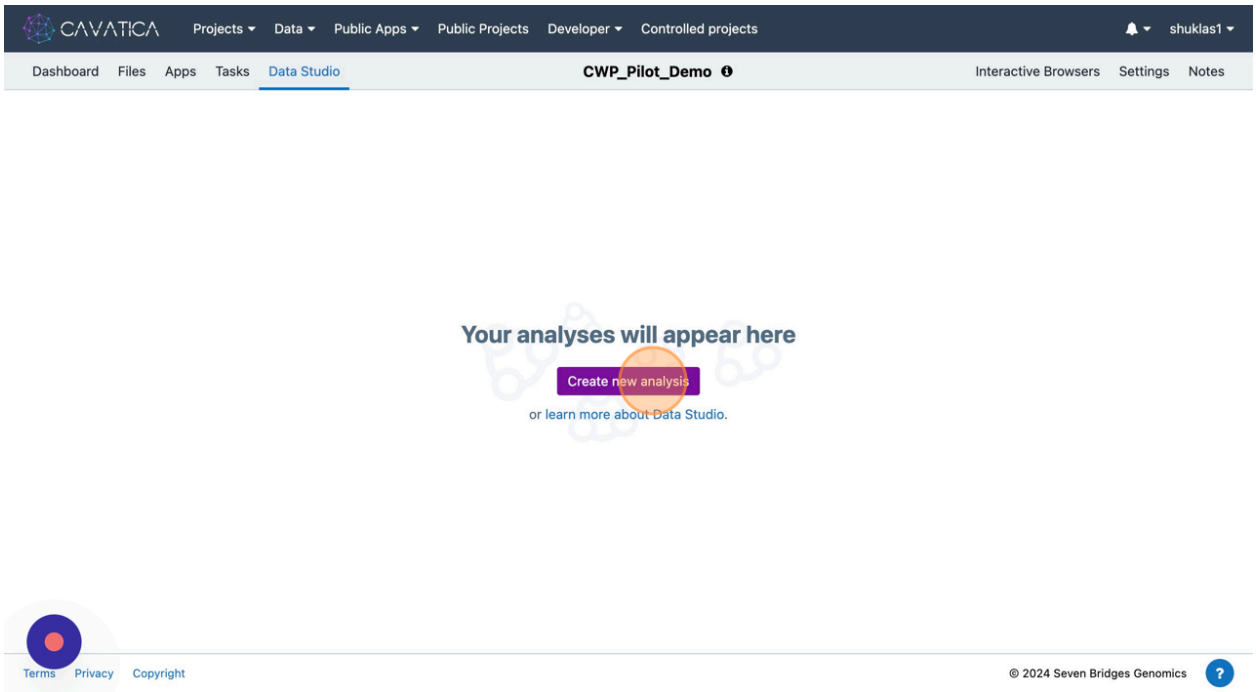
Explore Data using Data Studio

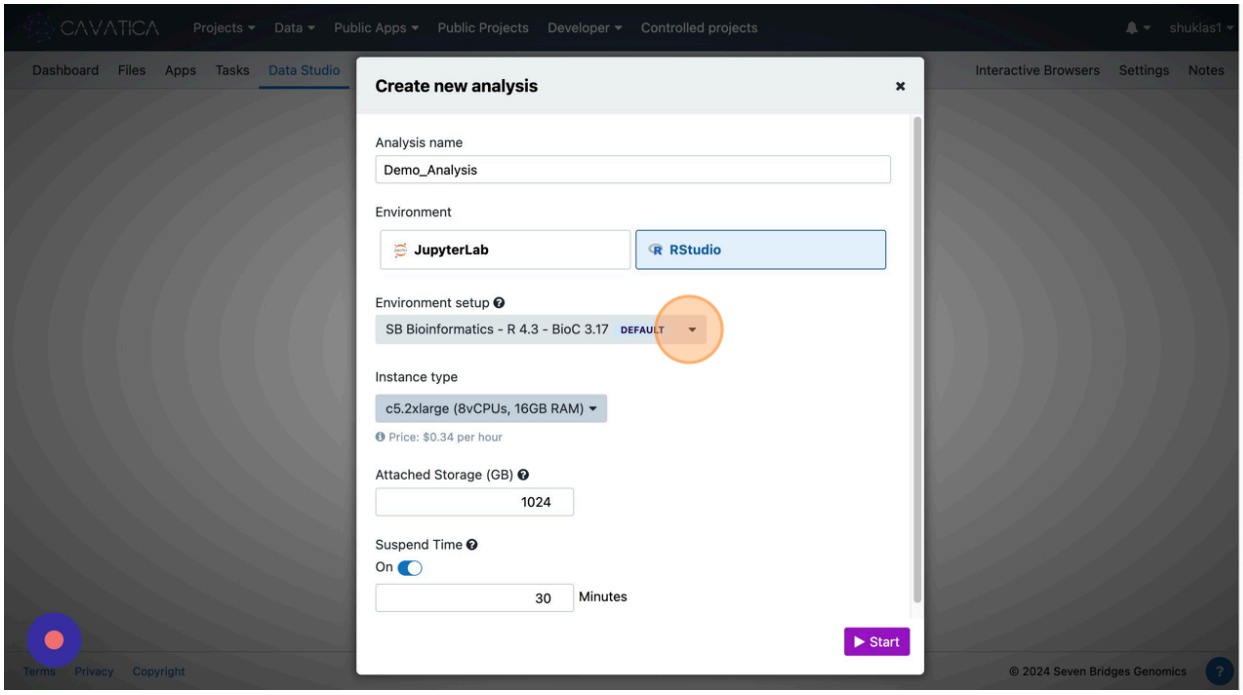
Many users may wish to examine their data using an interactive interface, such as RStudio or Jupyter Lab Notebook. You can run such interactive analyses in Data Studio feature, found in the project toolbar. We will show you how to create a Data Studio environment and analyze some of your data.

When the workflow completes running successfully, and the result files are ready, CAVATICA platform offers visualization and manipulation capabilities to the users via the 'Data Studio' tab. R and Python editors are available within the Data Studio to conduct such downstream analyses.



Note that starting an R or Python instance can take a while.

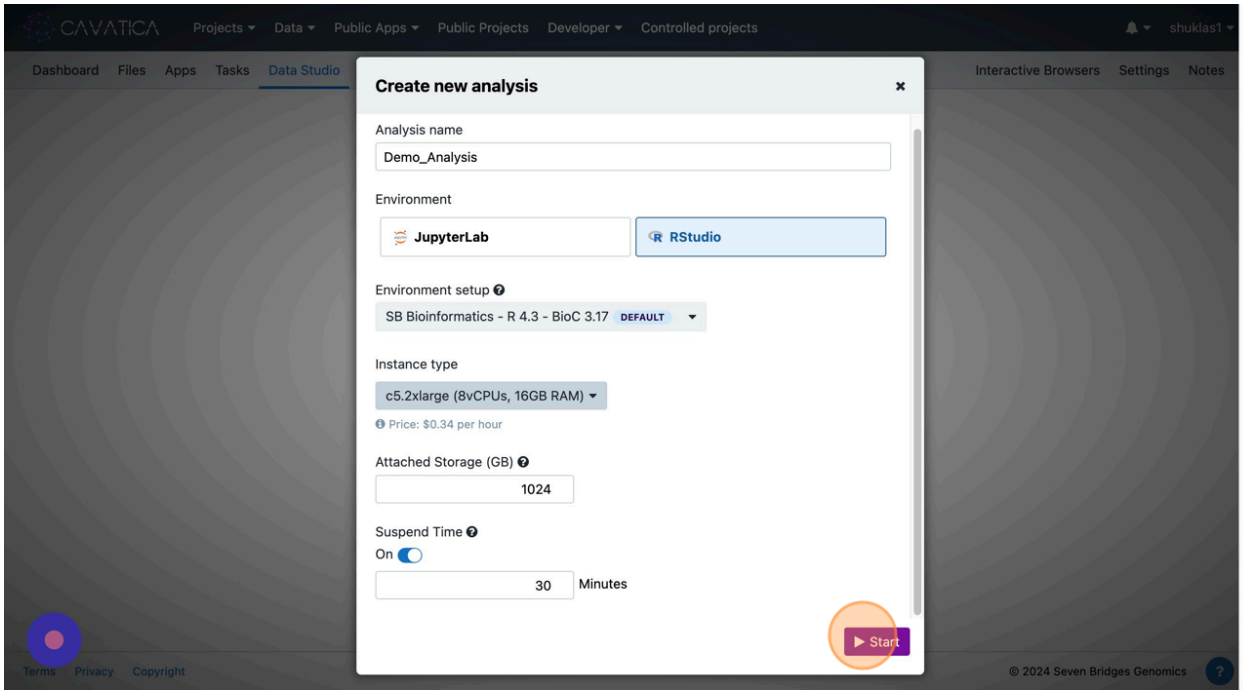




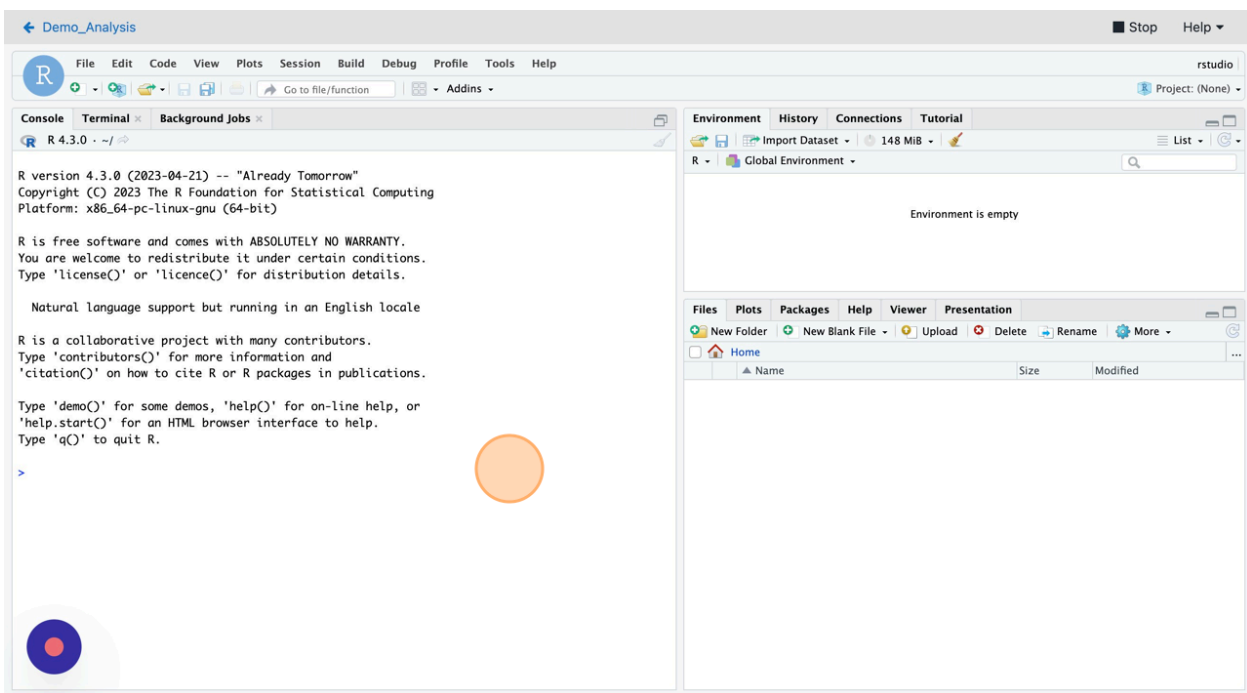
Choose the appropriate Environment Setup you may need for your exploratory data analysis.

Detailed documentation on the different environment specification is available here. Specific R and Python libraries are pre-installed on these environments.

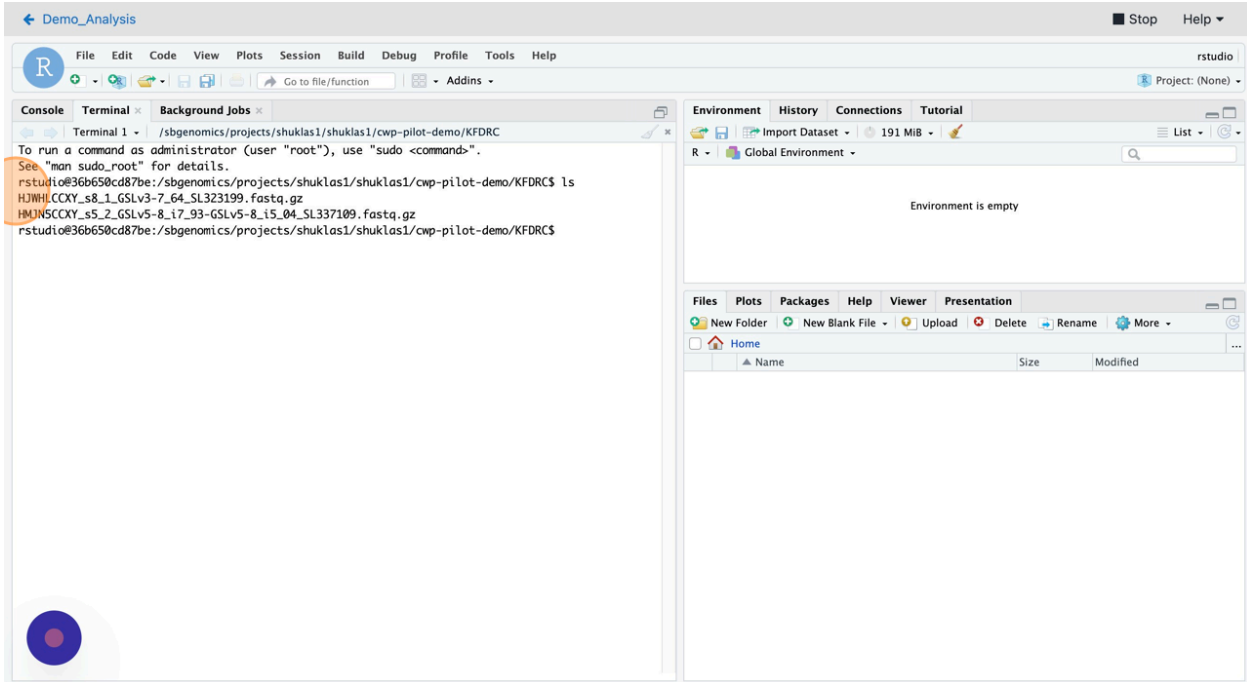
Finally, click on 'Start'.



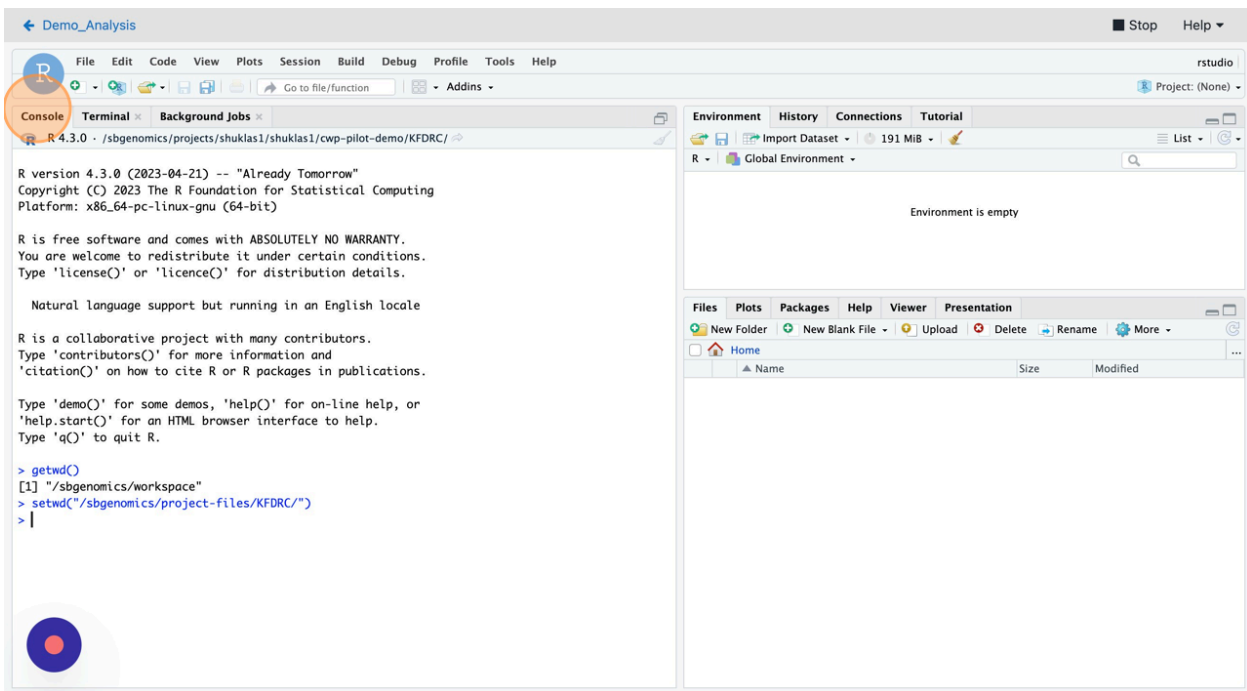
It will instantiate an R environment.

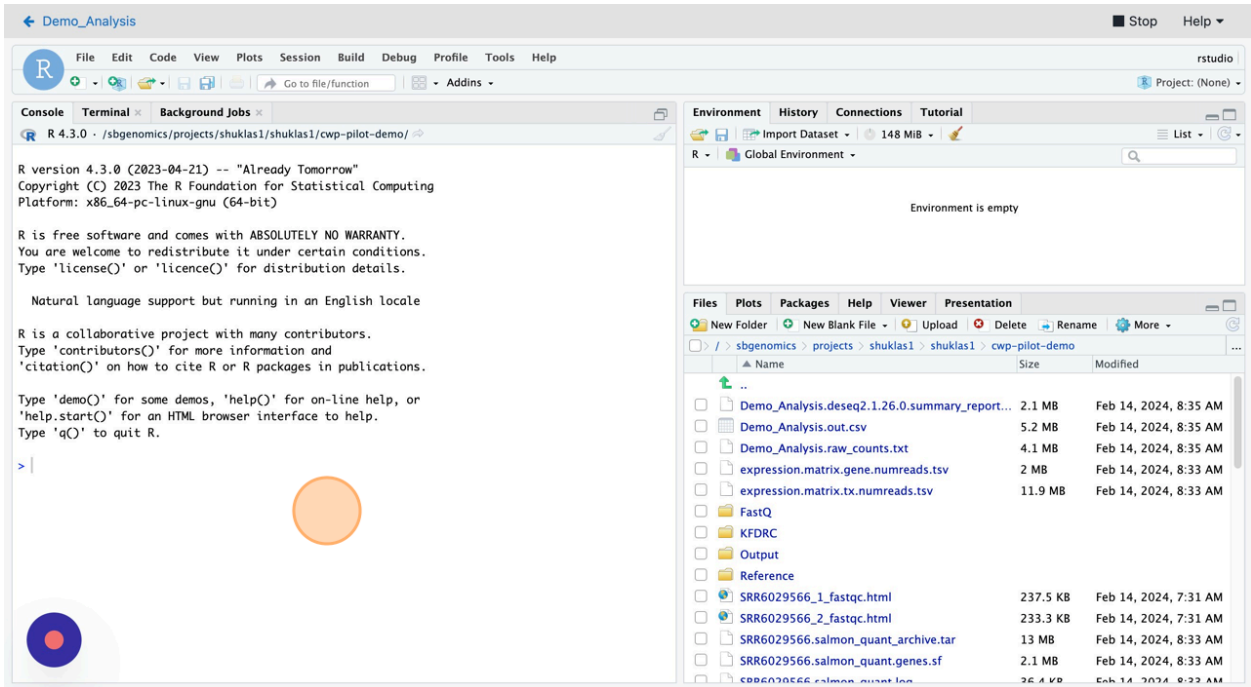


You can investigate the file path and the home directory for this instance using either the R Console, or shell terminal.

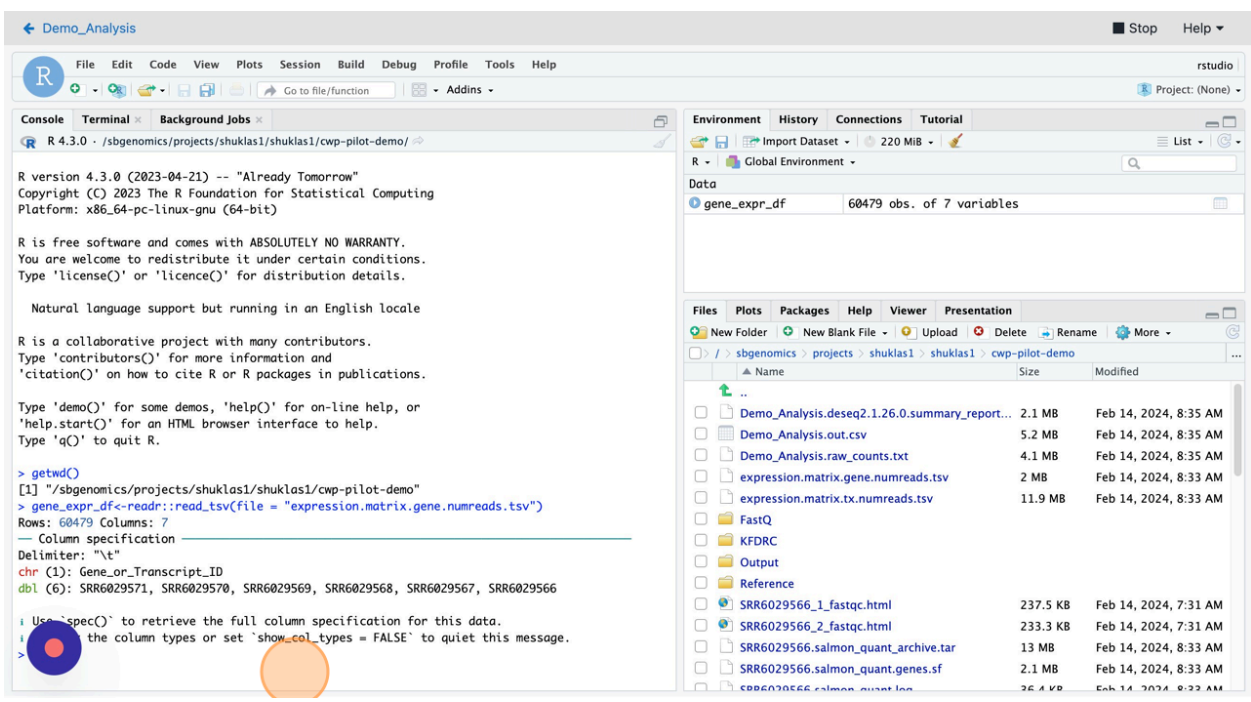


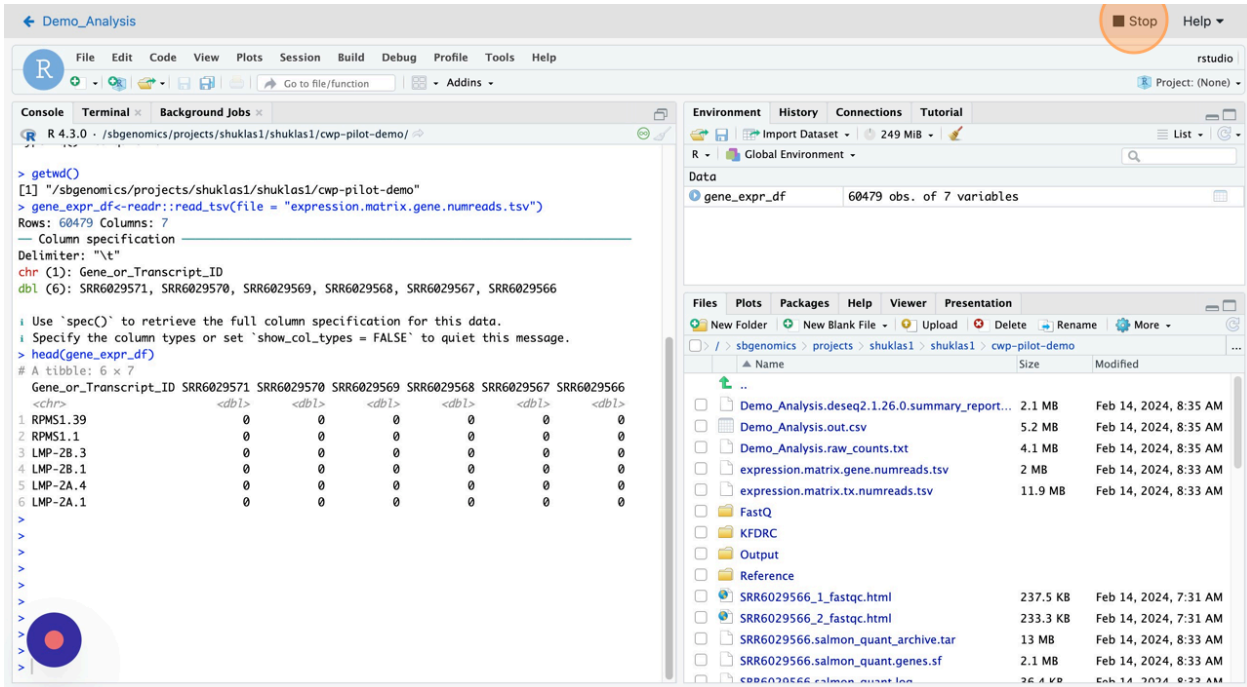
You may also change directories as needed.



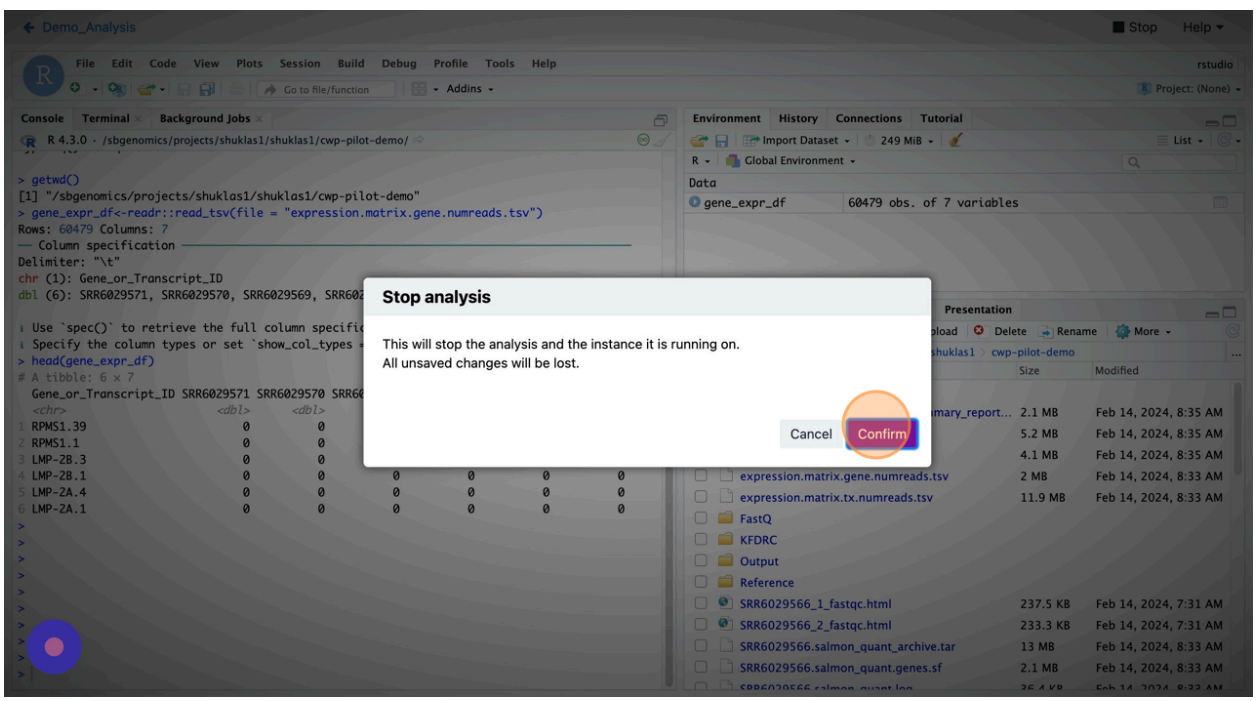


Once you are comfortable navigating to different files and folders, you may conduct exploratory data analysis suitable for your goal.

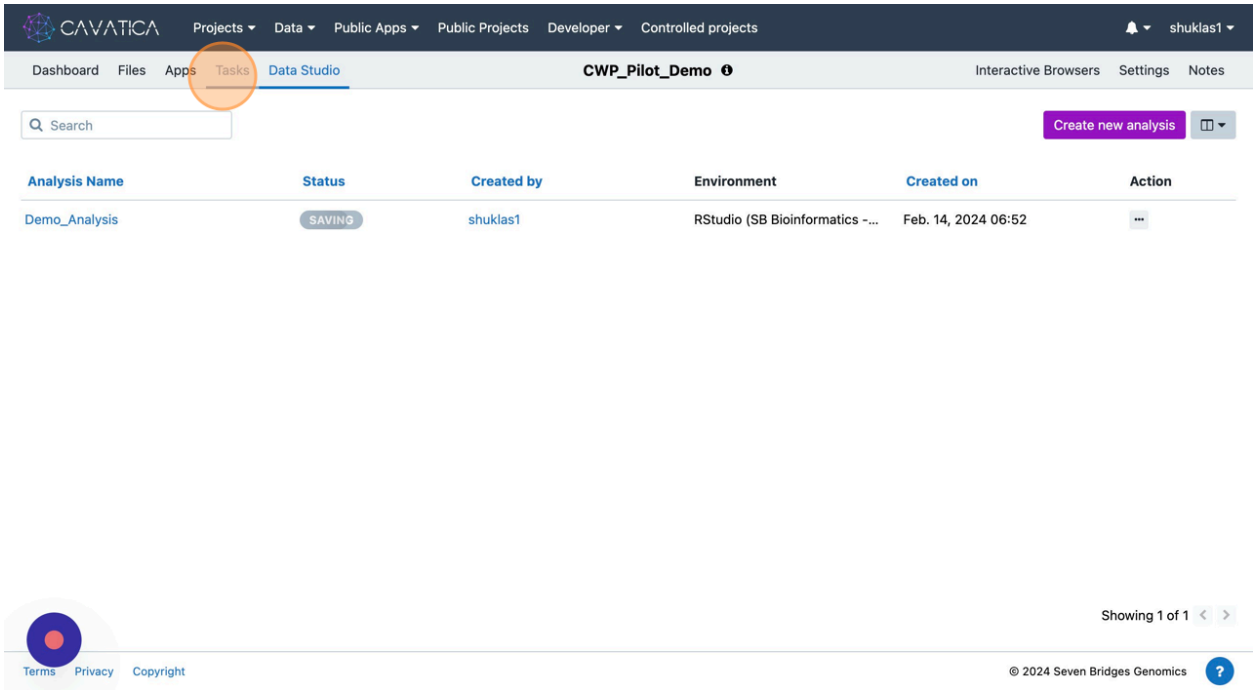




When the work is complete, click on ‘Stop’ to close the Data Studio instance.



You can also reopen the instance later if needed from the Data Studio tab.

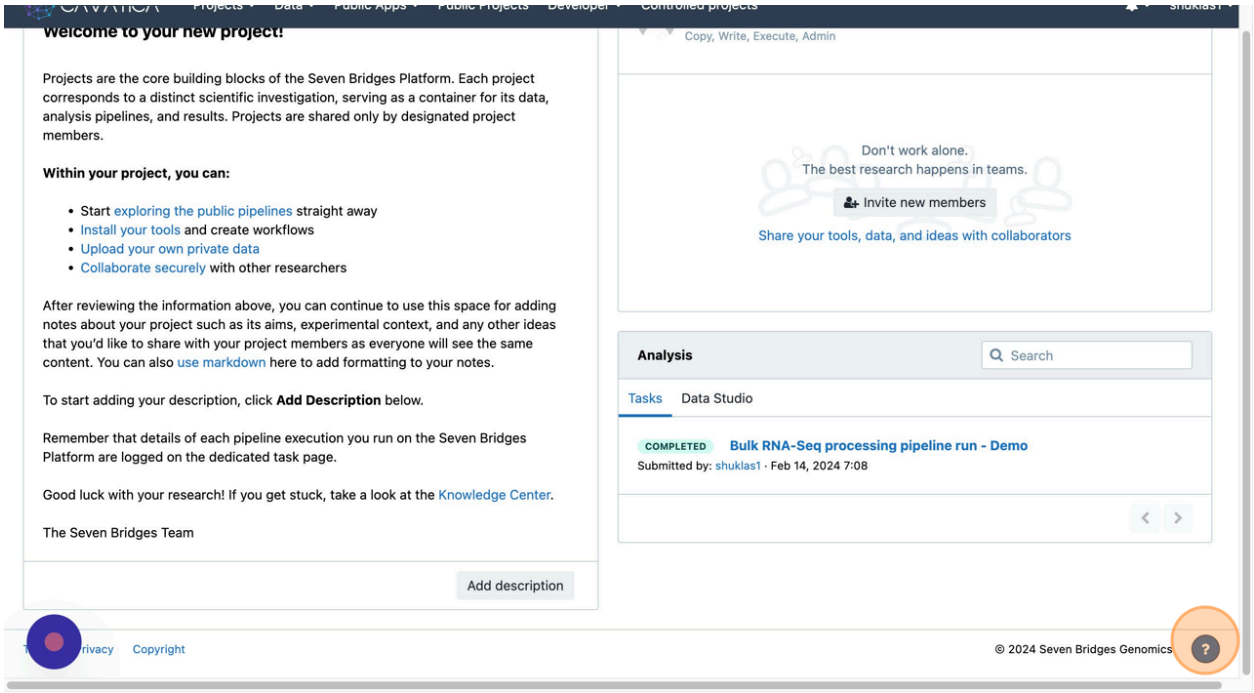

CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
Dashboard
Files
Apps
Tasks
Data Studio
CWP_Pilot_Demo
Interactive Browsers
Settings
Notes
Search
Create new analysis
Analysis Name
Status
Created by
Environment
Created on
Action
Demo_Analysis
SAVING
shuklas1
RStudio (SB Bioinformatics -...
Feb. 14, 2024 06:52
Showing 1 of 1
Terms
Privacy
Copyright
© 2024 Seven Bridges Genomics

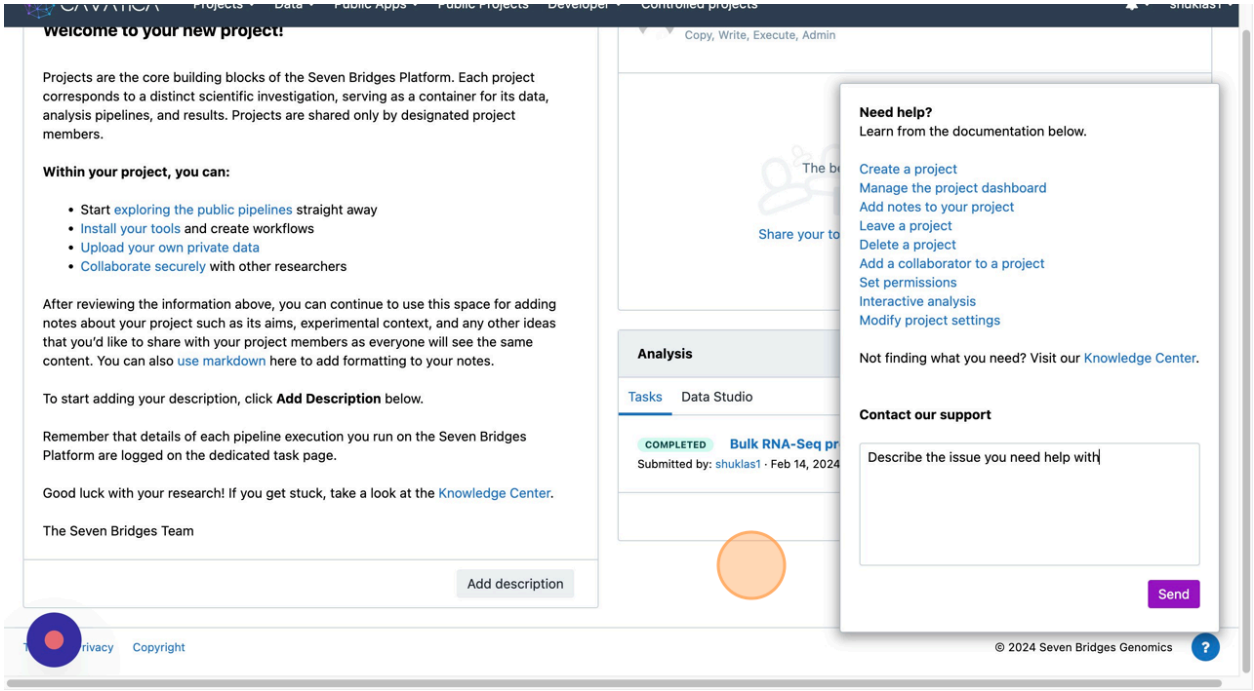
Debugging CAVATICA Application Error

Although the implementation chosen in this documentation ran without a glitch, there is always a possibility that errors in input supplied, or incorrect parameters, or any other cause may lead to the app not running successfully.
In such events, CAVATICA platform offers support and guidance to all users either directly from the platform by opening a support ticket, or through the weekly office hours where users are welcome to join via a Zoom call.

Below screenshots provide information on how to reach out to the support team by opening a support ticket.

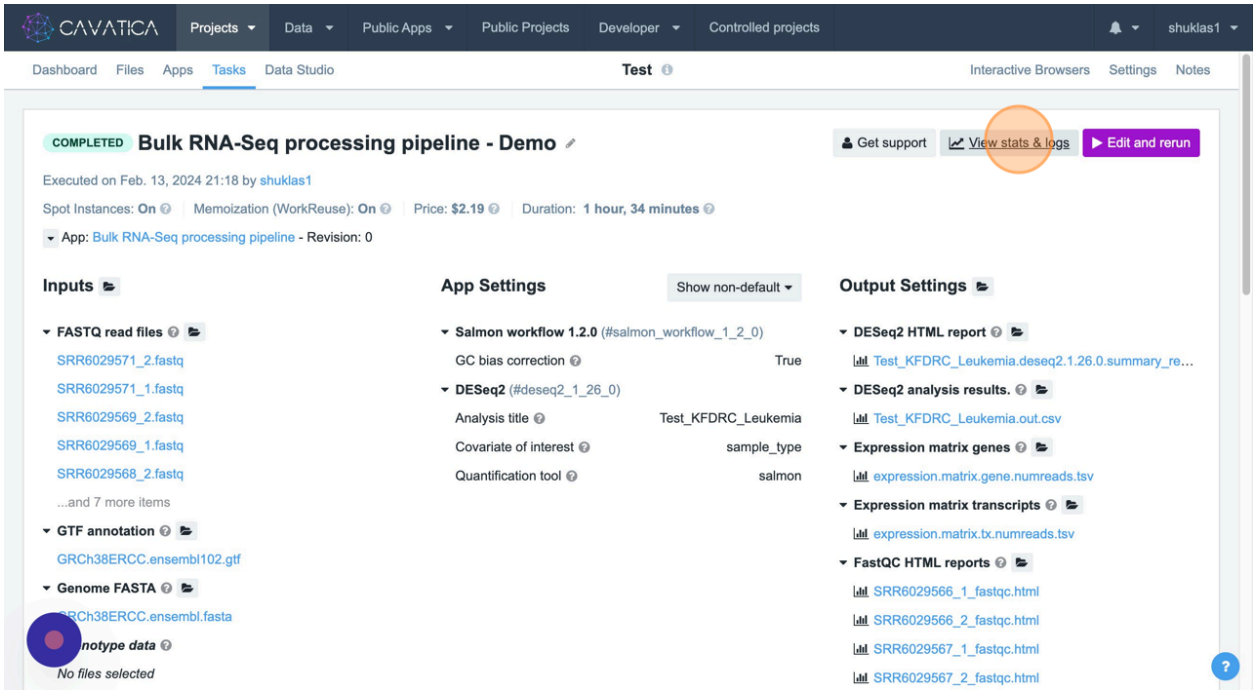
At the bottom of the screen, you will see a question mark, which, on clicking, opens the Help page.



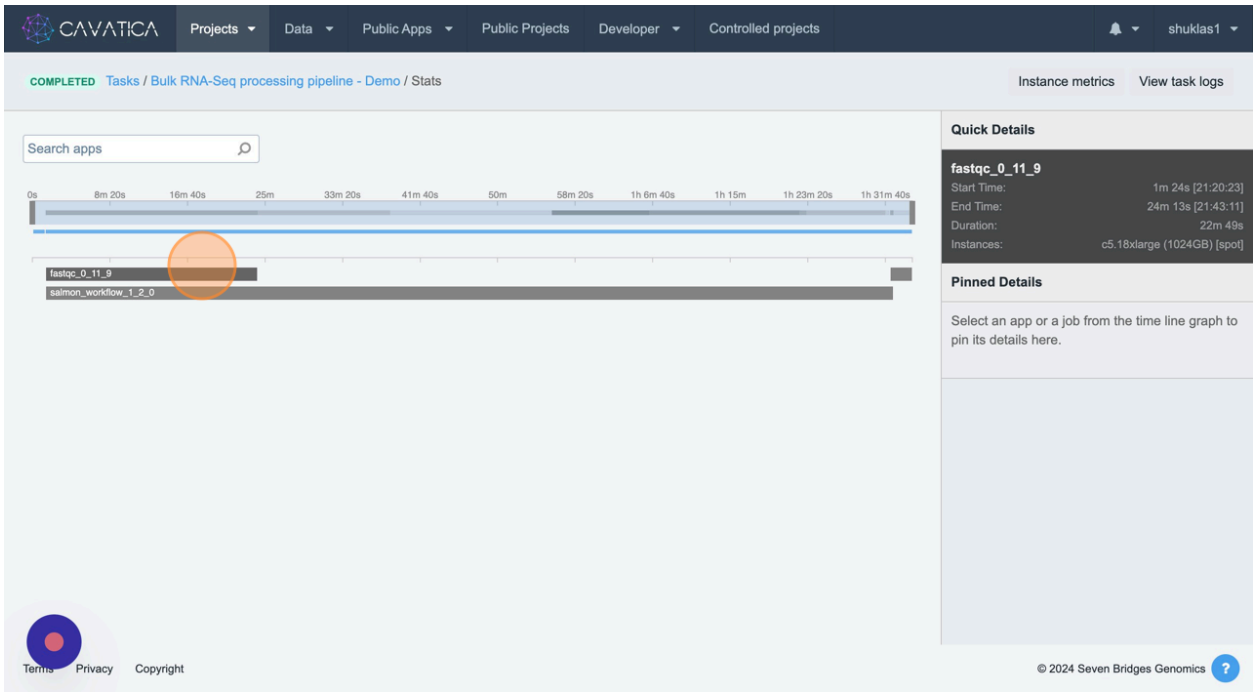


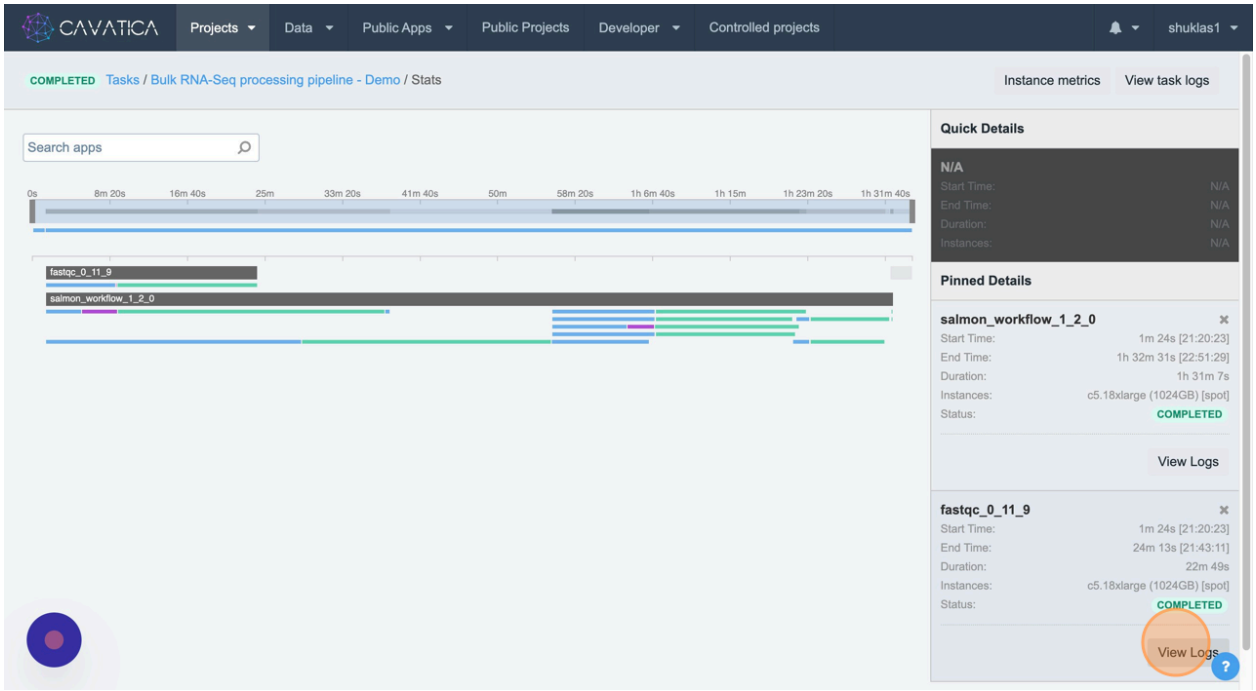
Once the ticket is opened, user receives a confirmation email. The CAVATICA support staff may reach out to the user(s) with specific questions or require additional access permissions. The project owner/admin can edit user access to the project so that the error can be looked into and resolved.

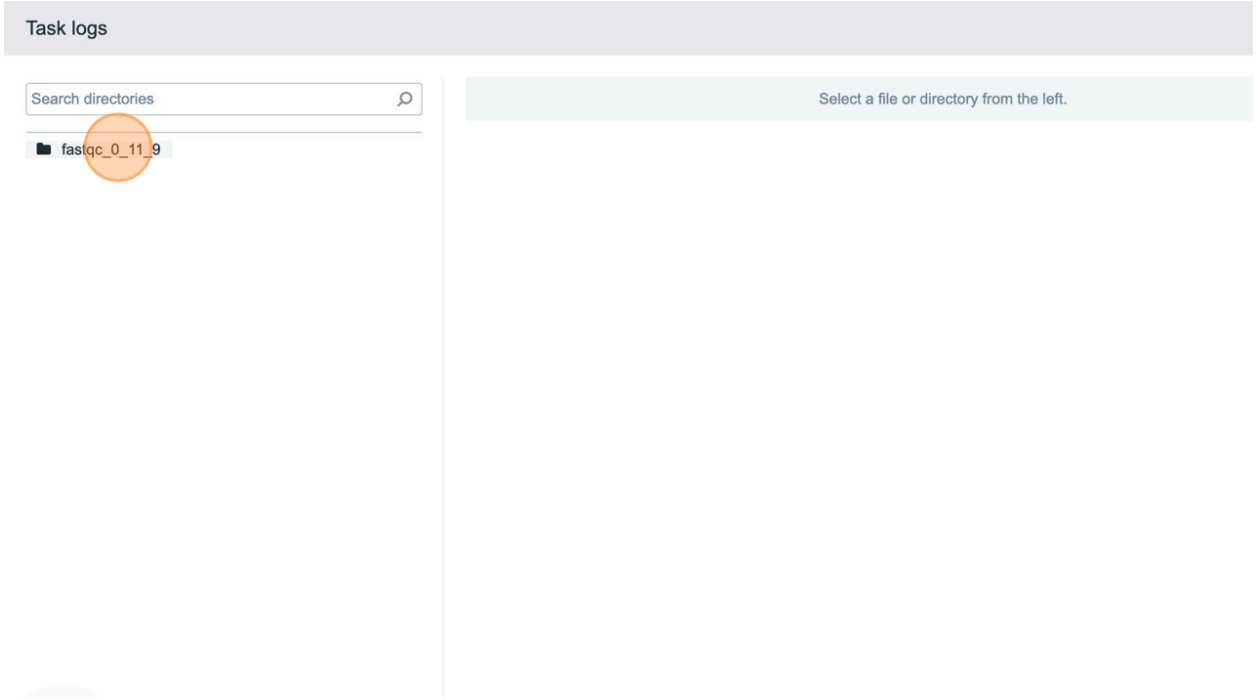
For more technically sound users, they can also look into the Logs and Stats within the project tasks to investigate the cause for failure.

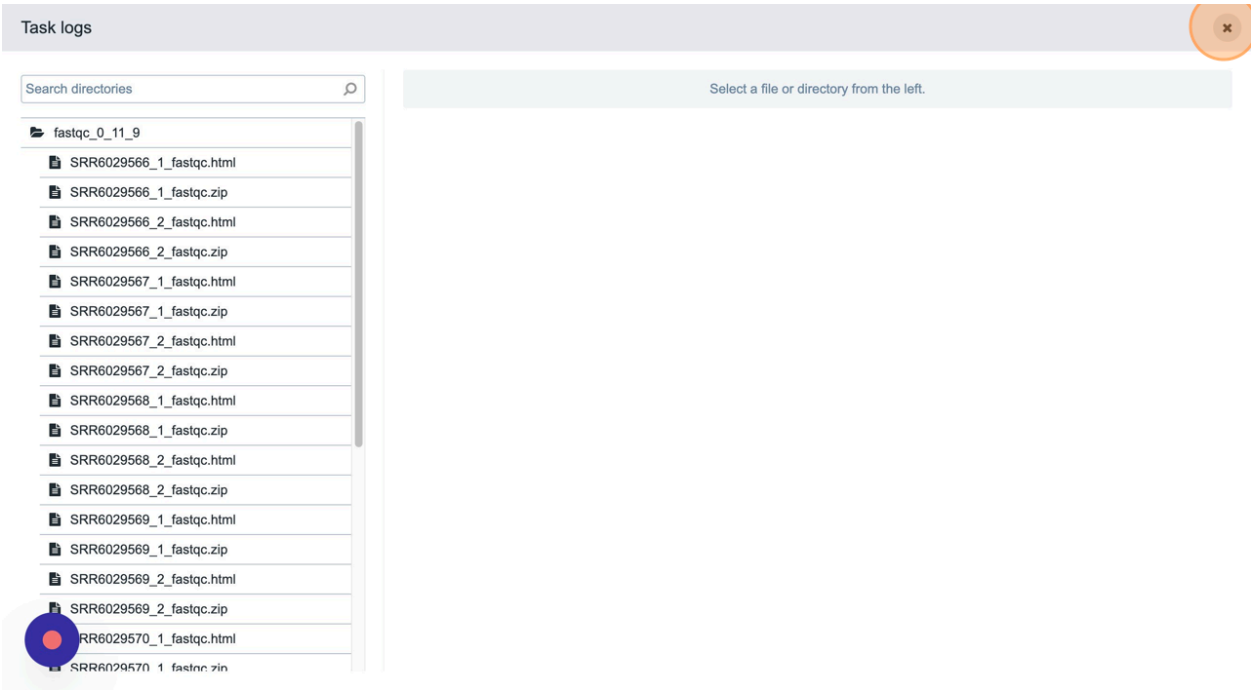


Logs and statistics for individual tasks within the overall workflow can be visualized for detailed investigation.




CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
shuklas1
COMPLETED
Tasks / Bulk RNA-Seq processing pipeline - Demo / Stats
Instance metrics
View task logs
Search apps
0s
8m 20s
16m 40s
25m
33m 20s
41m 40s
50m
58m 20s
1h 6m 40s
1h 15m
1h 23m 20s
1h 31m 40s
fastqc_0_11_9
salmon_workflow_1_2_0
Quick Details
N/A
Start Time: N/A
End Time: N/A
Duration: N/A
Instances: N/A
Pinned Details
salmon_workflow_1_2_0
Start Time: 1m 24s [21:20:23]
End Time: 1h 32m 31s [22:51:29]
Duration: 1h 31m 7s
Instances: c5.18xlarge (1024GB) [spot]
Status: COMPLETED
View Logs
fastqc_0_11_9
Start Time: 1m 24s [21:20:23]
End Time: 24m 13s [21:43:11]
Duration: 22m 49s
Instances: c5.18xlarge (1024GB) [spot]
Status: COMPLETED
View Logs


Task logs
Search directories
fastqc_0_11_9
Select a file or directory from the left.


Task logs
Search directories
Select a file or directory from the left.
fastqc_0_11_9
SRR6029566_1_fastqc.html
SRR6029566_1_fastqc.zip
SRR6029566_2_fastqc.html
SRR6029566_2_fastqc.zip
SRR6029567_1_fastqc.html
SRR6029567_1_fastqc.zip
SRR6029567_2_fastqc.html
SRR6029567_2_fastqc.zip
SRR6029568_1_fastqc.html
SRR6029568_1_fastqc.zip
SRR6029568_2_fastqc.html
SRR6029568_2_fastqc.zip
SRR6029569_1_fastqc.html
SRR6029569_1_fastqc.zip
SRR6029569_2_fastqc.html
SRR6029569_2_fastqc.zip
RR6029570_1_fastqc.html

CAVATICA Documentation and Resources

The intention and goal of this document is to equip CFDE users to get started with expanding the scope of their research using CFDE data on the CAVATICA platform while minimizing redundancy, improved efficiency with shared infrastructure, all while offering fast and cost-effective processing capabilities.

However, for more details on topics covered in the document and for additional resources, users are encouraged to refer to CAVATICA Docs.

The range of topics covered include how to get started, tutorials, access, projects, apps, files, metadata, archiving, secure collaboration, options to bring in custom tools, tool editor and wrapping tips, among others.

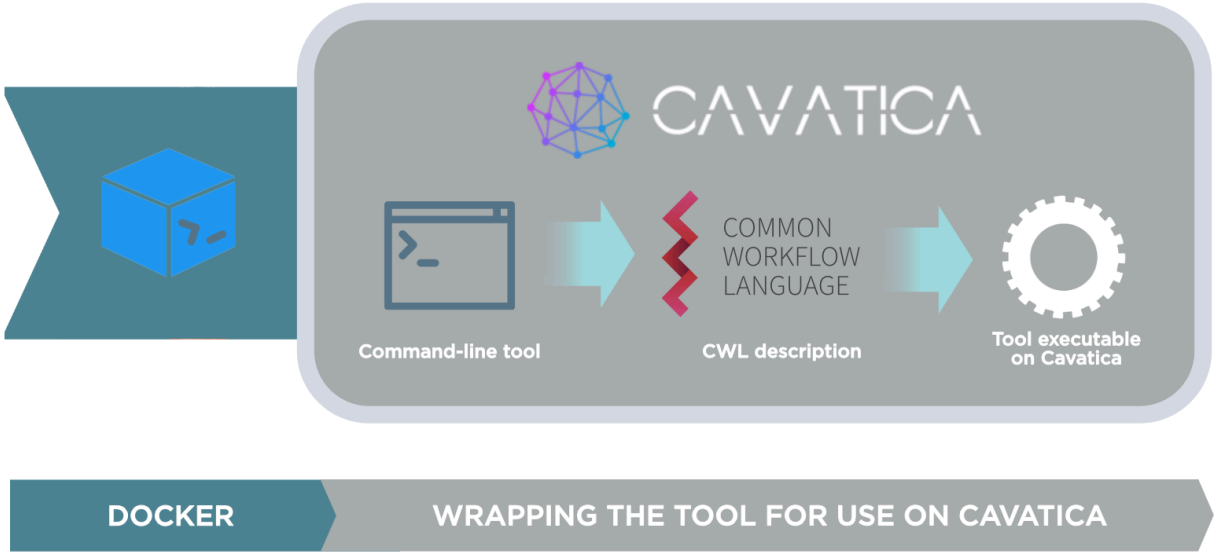
Scope and Potential for CFDE Users

With the wide scope of utility options that CAVATICA platform offers, users are welcome to add their research tools to the platform portfolio.
There is no limitation to the programming language for such custom tools. Any scripts can be wrapped in Common Workflow Language(CWL), and CAVATICA offers flexibility to include options for default or exposed parameters, input and output files.

Create CWL for custom scripts

CAVATICA allows you to bring your own tools and execute them on the Platform. This is done through our Software Development Kit (SDK) and the process consists of the following steps:

1. Create a Docker image containing the tool and its dependencies. Push the image to the CAVATICA Image Registry.
2. Use the tool editor on CAVATICA to create a description of the tool's functionalities. The description is automatically transcribed into the Common Workflow Language (CWL). This process is also known as *wrapping*.



This means that there is no need to reconfigure your existing command line tools to meet any proprietary format. Additionally, the tools remain runnable across a diverse range of infrastructures should you want to use them on different platforms.

To get your first hands-on experience with CWL, please read the Common Workflow Language User Guide which will take you from writing your first simple tool using CWL, to creating a workflow that contains several different interconnected steps. By reading this guide, you should be able to understand how each of the CWL tasks is isolated and that there is an explicit definition of its inputs and outputs. It is the explicitness and isolation that allow tools and workflows described with CWL to be **flexible**, **portable** across different CWL implementations and CWL-compliant execution engines and **scalable** from simple local execution to large-scale complex execution environments.

CAVATICA App and Dockerfile

Docker is an application that allows tools and their dependencies to be packaged into discrete runtime environments. These environments, **containers**, are instantiated from **images** and are stored inside an **image registry**.

For an overview of Docker, please see the Docker website. Learn more about Docker images, containers and image registries below.

Docker images uploaded to the Cavatica Image Registry are further organized into repositories. Once the images are uploaded to the Cavatica Image Registry, you can run these tools on Cavatica. Workflows will execute the tools in series inside their Docker containers.

You can also execute tools on Cavatica that are contained in images stored in Docker Hub – the Docker Image Registry. However, storing your images in the Cavatica Image Registry rather than in Docker Hub will speed up processing time on Cavatica, since the tools will be executed closer to the data they are processing.

Create Public Projects and Apps

Users can publish their tools and workflows to the platform's Public Apps Gallery instantly by publishing the project containing it. Anyone with access to the URL can then view and copy the contents. Any changes made to the Public app are also reflected to the Public Apps gallery immediately. Users may Contact us at support@sevenbridges.com to publish your project.

Published apps are tagged to indicate you as the publisher and appear in the Public Apps gallery. However, project files in the published project do not appear in the Public Reference Files repository, and your project is not listed as a public project on CAVATICA.

It is however important to note that although the project or the app may be public, and that the data files within the project may be visible to users who access the project, users must also have authorized access to those data files. To understand this more clearly, user must have authorized access to the KidsFirst portal and specific study that is the source of the data file. User must also have a valid eRA Commons account, and an ORCID id to access the CAVATICA platform, CFDE portal, Kids First portal and any other compatible data platforms and studies such as dbGAP, TCGA, et cetera.