



CHILDHOOD CANCER DATA INITIATIVE (CCDI)

Data Access Instructions

7/19/2023

Version	Date	Description
1.0.0	07/19/2023	Initial Version

Introduction and Overview

This document provides information about how to access, query, and process data from the Childhood Cancer Data Initiative (CCDI) stored at the National Cancer Institute's (NCI) Cancer Data Service (CDS). The CCDI studies that submit data to the CDS are registered with the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGaP), which maintains a list of the subject IDs, sample IDs, and consents.

Requesting Access to the Data

The CDS hosts open and controlled access data. Controlled-access data requires authorization through [dbGaP](#). A step-by-step breakdown of the data access request process is located in this guide: [How to Request and Access Data Sets from dbGaP](#). Following authorization, users can analyze CCDI data on the Cancer Genomics Cloud (CGC) through the [Cancer Data Service \(CDS\) Explorer](#). There is also [a tutorial](#) on how to import CDS data.

Contact ncichildhoodcancerdatainitiative@mail.nih.gov with any questions. Seven Bridges also hosts [Office Hours](#) to answer questions about using the CGC site.

Sample Use Case: Accessing CCDI Molecular Characterization Initiative (MCI) Data

One of the CCDI data sets available on the CDS Explorer is from [MCI \(phs002790\)](#). Please see the instructions below on how to access these data. Note that a [list of all CCDI studies](#) released is also available.

1. From the CGC home page at cancer-genomics-cloud.org, click the “CREATE AN ACCOUNT OR LOGIN HERE” link in the center of the page (Figure 1).



Figure 1: CGC home page with “CREATE AN ACCOUNT OR LOGIN HERE” button highlighted in a red box

2. On the login options screen, click on the “Log in with eRA Commons” link (Figure 2).

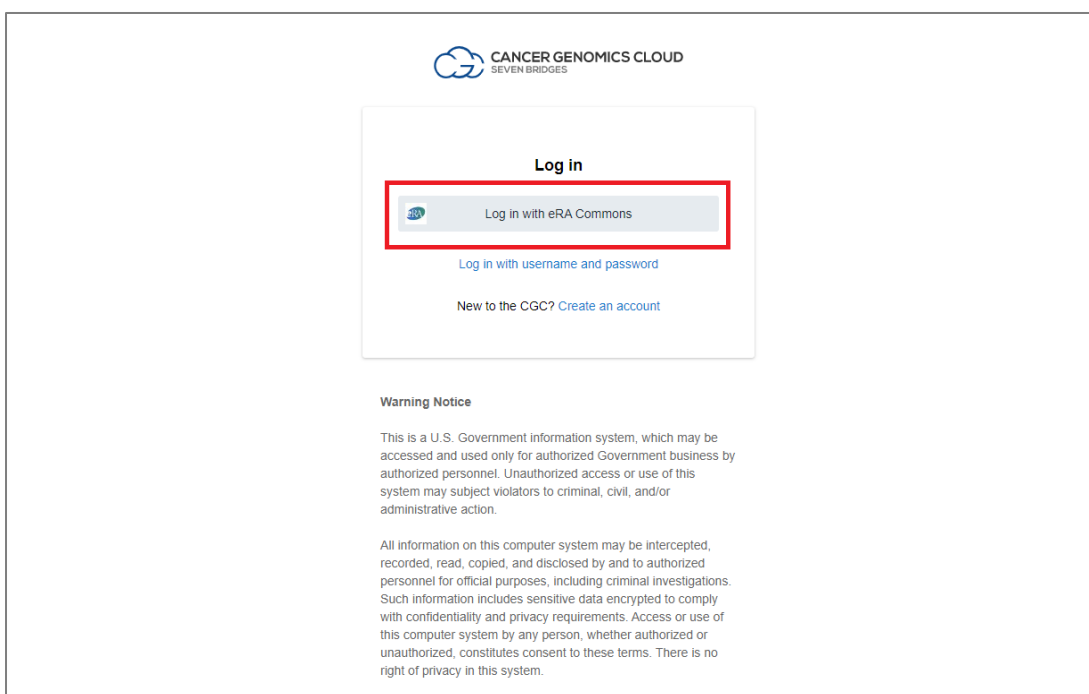


Figure 2: eRA Commons log-in screen for the CGC, with log-in button highlighted in a red box

3. On the login screen, enter the eRA Commons account credentials associated with your approved dbGaP study and click “Sign In” (Figure 3).

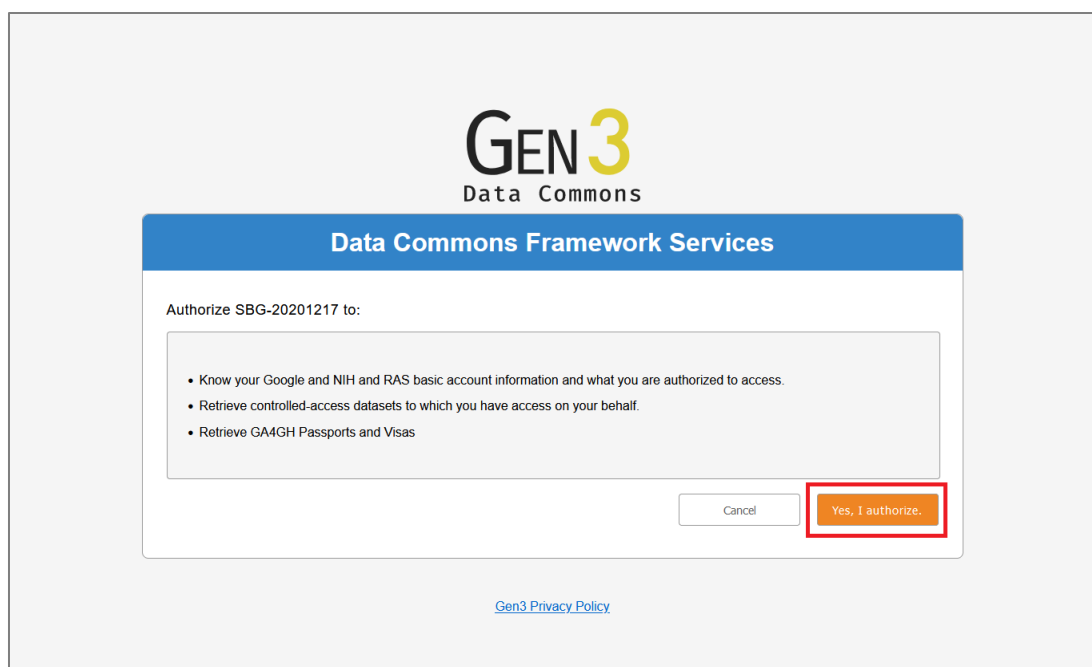
Please note that if you receive an error message when logging in here, you can confirm that your eRA Commons username and password are correct by logging in to the [eRA Commons site](#). (If you’ve previously logged into the eRA Commons site, you may need to clear your web browser’s cache or use the “incognito” mode to ignore cached data and cookies so that you can enter and test your credentials. If you receive an error message on that site as well, you may need to reset your eRA Commons password.

Figure 3: Login page with to access CGC with username and password credentials sections highlighted in a red box.

4. If you agree to the “Consent to Share Information” on the following page, click the “Grant” button to continue (Figure 4).

Figure 4: Consent to share information page with red box highlighting “Grant” button to confirm consent to share information with the CGC

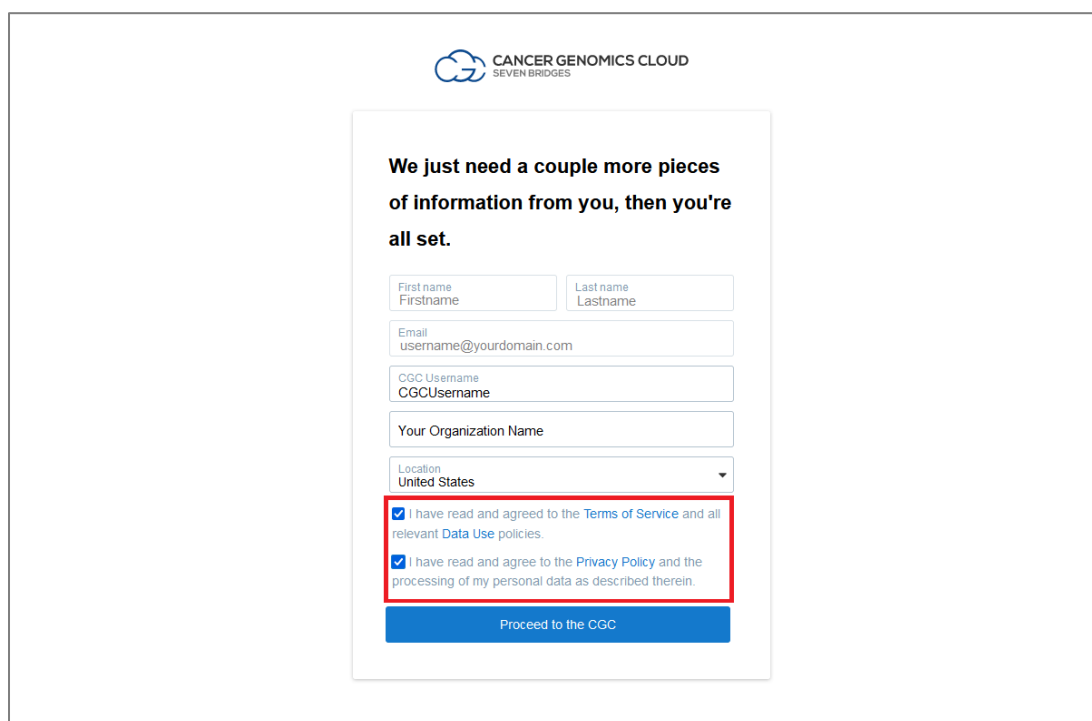
5. If you agree to authorize the Gen3 data Commons Framework Services to share your account and authorization information to access the data sets for which you have been approved, click the “Yes, I authorize” button (Figure 5).



The image shows a web interface for the Gen3 Data Commons. At the top, the 'GEN3 Data Commons' logo is displayed. Below it, a blue header bar reads 'Data Commons Framework Services'. The main content area is titled 'Authorize SBG-20201217 to:' and contains a list of permissions: 'Know your Google and NIH and RAS basic account information and what you are authorized to access.', 'Retrieve controlled-access datasets to which you have access on your behalf.', and 'Retrieve GA4GH Passports and Visas'. At the bottom right of the authorization box, there are two buttons: 'Cancel' and 'Yes, I authorize.'. The 'Yes, I authorize.' button is highlighted with a red rectangular box. Below the authorization box, there is a link for 'Gen3 Privacy Policy'.

Figure 5: Gen3 Data Commons framework services authorization page with “Yes, I authorize” button highlighted in a red box

6. On the next page, confirm that the information listed for you is correct (if this page appears). If you agree to the Terms of Service, Data Use, and Privacy Policies, click the two related checkboxes, and then click on “Proceed to the CGC” (Figure 6).



The image shows a registration confirmation page for the Cancer Genomics Cloud (CGC). The header features the CGC logo and the text 'CANCER GENOMICS CLOUD SEVEN BRIDGES'. The main heading reads: 'We just need a couple more pieces of information from you, then you're all set.' Below this, there are several input fields: 'First name' (with 'Firstname' as placeholder), 'Last name' (with 'Lastname' as placeholder), 'Email' (with 'username@yourdomain.com' as placeholder), 'CGC Username' (with 'CGCUsername' as placeholder), and 'Your Organization Name'. There is also a 'Location' dropdown menu currently set to 'United States'. Below the form fields, there are two checkboxes, both of which are checked. The first checkbox is labeled 'I have read and agreed to the Terms of Service and all relevant Data Use policies.' and the second is labeled 'I have read and agree to the Privacy Policy and the processing of my personal data as described therein.' Both checkboxes and their associated text are highlighted with a red rectangular box. At the bottom of the form, there is a blue button labeled 'Proceed to the CGC'.

Figure 6: Confirmation of terms and policy for CGC registration highlighted in red.

7. If the CGC questionnaire appears, complete it to continue (Figure 7).

Due to technical issues on the external provider side, some users might not be able to access TARGET controlled data. A dedicated team is currently working on fixing the issue. For more information, please contact us at support@sevenbridges.com. Thank you for your patience and understanding.

Projects Data Public Apps Public Projects

cdq-questionnaire-title

The Cancer Genomics Cloud allows you to view and analyze controlled-access data, according to the access granted to you by dbGaP and/or the ICGC DACO. Before you can begin, you will need to answer a few questions about controlled-access data usage. You will only need to complete this questionnaire once. [Learn more](#)

1 On the CGC a controlled project:

☐ A. Serves as a workspace for analyzing controlled-access data

☐ B. Allows all members to access all raw data

☐ C. Allows all members to access all derived data

☐ D. A, B and C

2 As a CGC Certified User I can:

☐ Browse open- and controlled-access cancer genomics data from the Data Browser

☐ Create a controlled project

☐ Add other Certified Users to a controlled project

☐ Access raw files according to my dbGaP and/or ICGC DACO approvals

☐ All of the above

3 I'm studying the influence of lifestyle changes on cancer progression, but some of the variables I'm interested in are not captured. Am I allowed to try to identify participants and ask them follow-up questions?

☐ Yes, absolutely. The participants may decide if they want to ignore this request.

☐ No, absolutely not. All data access agreements specifically prohibit

[Check answers](#)

Figure 7: Screenshot of CGC questionnaire

8. Click on the “Data” dropdown at the top left of the CGC home page and then select “Cancer Data Service Explorer” (Figure 8).

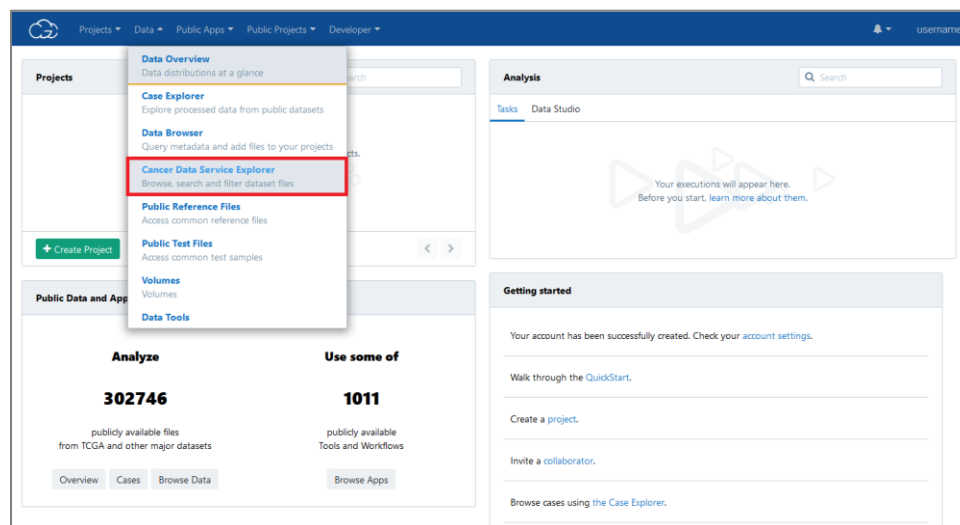


Figure 8: CGC Data dropdown menu with red box around the Cancer Data Services Explorer

9. CCDI studies are marked with “(CCDI)” at the end of the study name. Click on the “PHS002790” link to view basic MCI study information on its dbGaP study page (Figure 9).

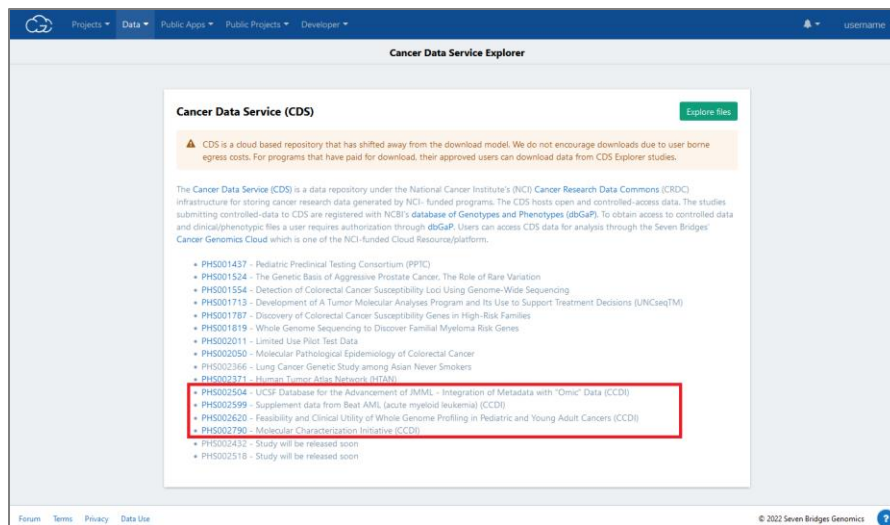


Figure 9: CDS Explorer study list page with red box around CCDI studies

- From the CDS Explorer study list page, click on the “Explore Files” button at the top right of the screen to continue to the CDS Explorer (Figure 10).

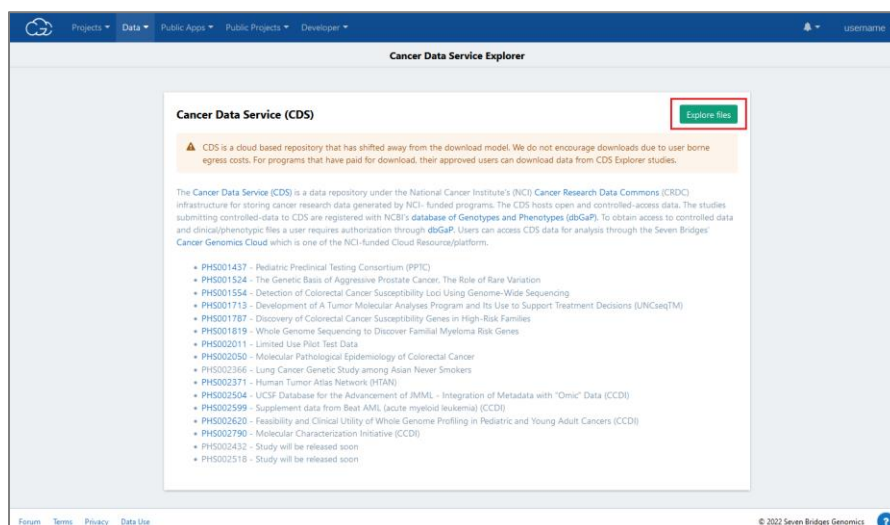


Figure 10: CDS Explorer study list page with red box highlighting "Explore Files" button.

- Use the facets panel on the left side of the screen and click the checkbox next to “PHS002790” under “Access number” to view available data for only that study (Figure 11).

You may use any other facet options to further filter the data set as desired. Any data that you are authorized to access will show a green check mark in the “Authorized” column of the main panel. Data that you are not authorized to access will instead show a red “X” in that column.

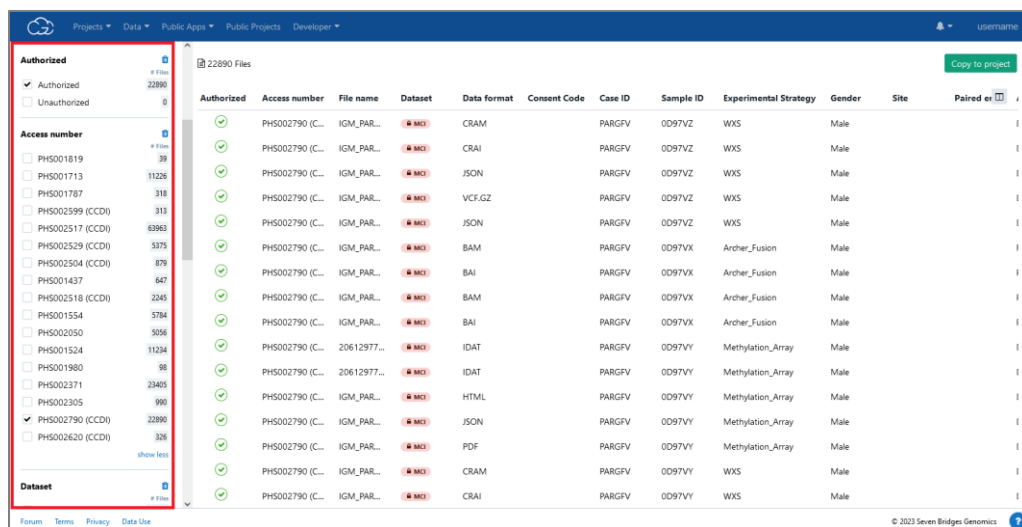


Figure 11: CDS explorer page with red box around left column showing CDS Explorer filters

- Once you've narrowed the data set based on your selections, you can click on the "Copy to project" button at the top right of the page to add your data to a study (Figure 12).

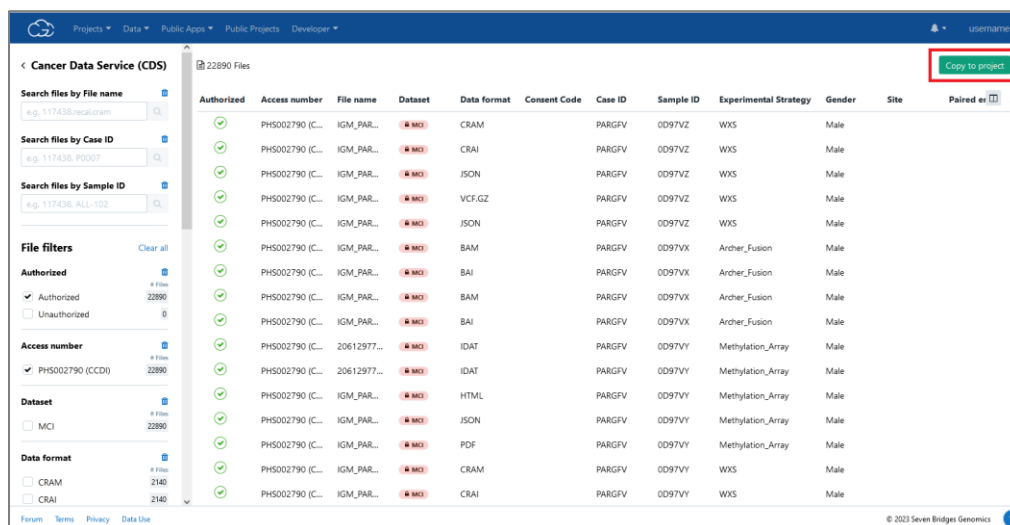


Figure 12: CDS Explorer page with button (upper right) to click to copy selected files to a CGC project indicated with red box

- Create a new project or select an existing one in the pop-up window and then click "Copy" to add the chosen files to that project (Figure 13).

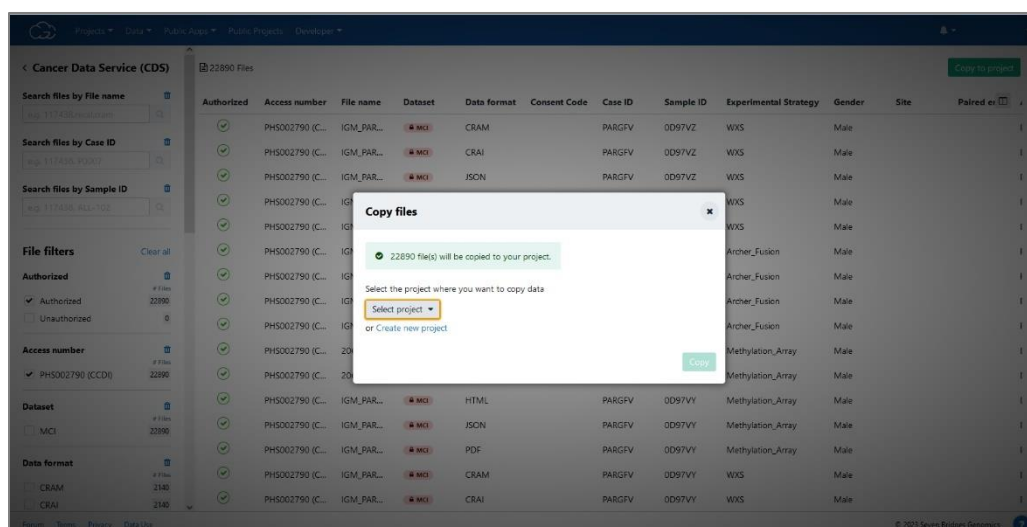


Figure 13: Pop-up showing a dropdown menu with option to copy selected files to a new or existing CGC project

Using the CGC Data Studio

After creating a CGC project, you can then use CGC Data Studio to enter and execute Python, R, or Julia code for conducting additional data analyses on the CGC.

1. From the “Projects” drop down menu, choose the project that contains the data you’d like to analyze (Figure 14).

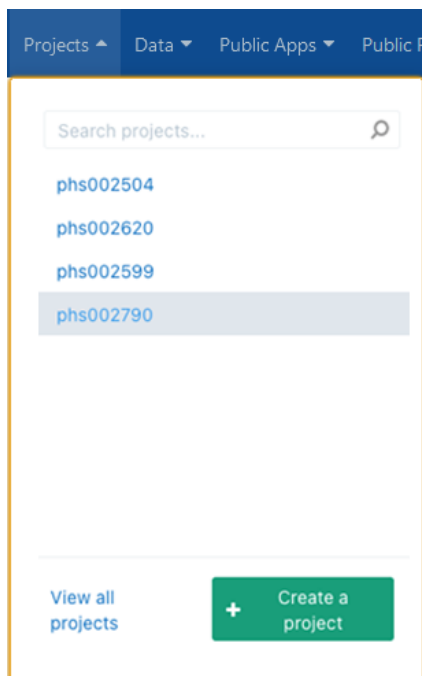


Figure 14: Projects dropdown with List of examples projects with highlighted selection.

2. Once in the project, select the “Data Studio” link from the top of the screen and then click on “Create new analysis” (Figure 15).

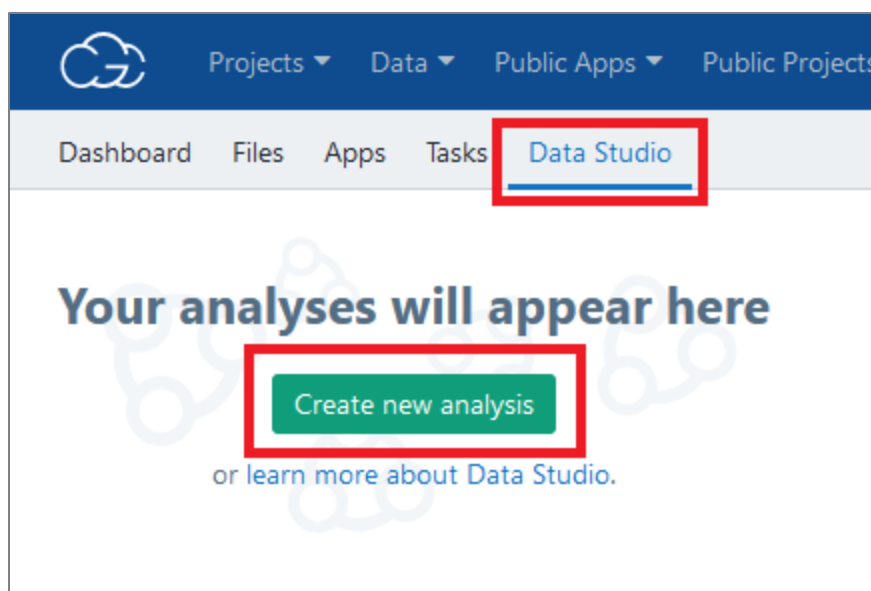


Figure 15: Data Studio menu page with “Creating new analysis” button highlighted in red.

3. Enter a name for the analysis, select “RStudio” or “JupyterLab” under “Environment,” and click the “Start” at the bottom of the dialog box (Figure 16).

Figure 16: Yellow box indicates where to put name of new RStudio analysis

4. A new workspace environment will be created (Figure 17).

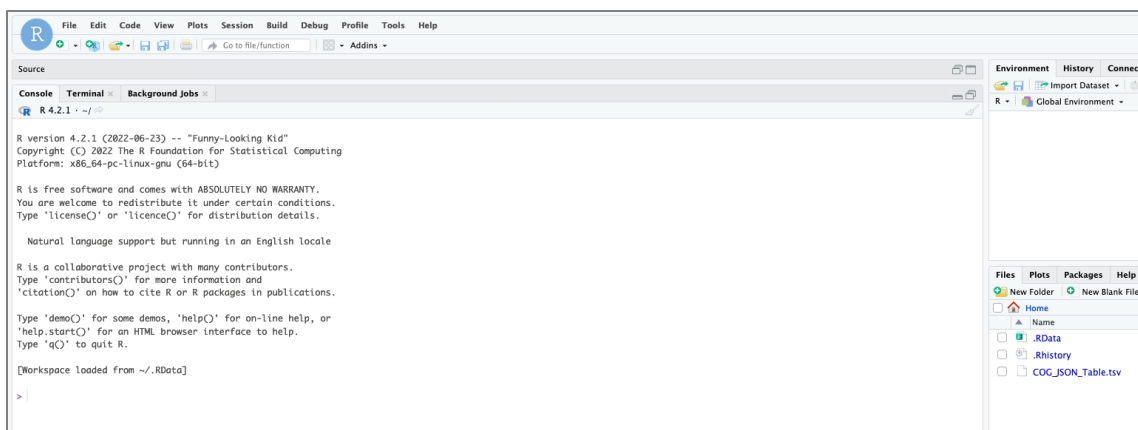


Figure 17: RStudio integrated development environment (IDE)

5. Choose the appropriate instance for analysis. The Instance type list shows the available instances, including their disk size, number of vCPUs, and memory (indicated in brackets). The default instance is c5.2xlarge, which offers 1024 GB of EBS storage, 8 vCPUs, and 16 GB of RAM.
6. Adjust the size of the attached storage. The attached storage consists of disks used by the computation instance for additional storage capacity during task execution. You can choose a size between 2 and 4096 GB. For more information, refer to the [documentation](#).
7. (Optional) Modify the [suspend time settings](#) if desired.
8. Click “Start.” The CGC will initiate the process of acquiring an appropriate instance for your analysis. This may take a few minutes.

Additional Resources About Working with Data at the CGC

- a. CGC Documentation: <https://docs.cancer-genomics-cloud.org/docs>
- b. Importing CDS Data: <https://docs.cancer-genomics-cloud.org/docs/import-cds-data>
- c. Common Workflow Language Workflows and Apps: <https://cgc.sbgenomics.com/public/apps>
- d. Volumes: <https://docs.cancer-genomics-cloud.org/docs/volumes-1>
- e. Tool Editor Tutorial: <https://docs.cancer-genomics-cloud.org/docs/tool-editor-tutorial>
- f. About the Editor: <https://docs.cancer-genomics-cloud.org/docs/about-the-editor>