CHILDHOOD CANCER DATA INITIATIVE (CCDI)

Data Access Instructions

2/02/2024

Content

Background

The guiding principle of the National Institutes of Health (NIH) Data Sharing Policy is to make data available in a timely manner to the largest possible number of investigators. For human data, data are made available under terms and conditions consistent with the informed consent provided by individual participants, and the confidentiality of the data and the privacy of participants are protected.

For the Childhood Cancer Data Initiative (CCDI), some resources contain open-access data, while others contain registered-access or controlled-access data sets.

Open Access: For public access; requires no special credentials.

Examples: Childhood Cancer Data Catalog, Molecular Targets Platform

Registered Access: For anyone registered with the repository; usage may be monitored.

Example: National Childhood Cancer Registry

Controlled Access: For credentialed users who have applied to access the data.

Example: CCDI genomic data stored at the Cancer Data Service

Reach out to the CCDI mailbox with any questions.

Introduction and Overview

This document provides information about how to find, request, access, download, and analyze controlled-access data from CCDI.

CCDI studies are summarized and indexed in the CCDI Hub Explore Dashboard, where row-level metadata for CCDI participants, samples, or files can be exported. The shopping cart feature on the dashboard allows users to select and manage files of interest and download a comma-separated values (CSV) file manifest. This manifest file can be uploaded to the Cancer Genomics Cloud (CGC) for downstream data analysis or used locally. CGC is a flexible cloud platform, enables the analysis, storage, and computation of large cancer datasets.

The data is hosted in NCI's Cancer Research Data Commons (CRDC), a cloud-based infrastructure where the data are hosted. For controlled-access studies, CRDC collaborates with Data Commons Framework Services (DCFS) to provide authentication and authorization services for the database of Genotypes and Phenotypes (dbGaP). To gain access to controlled data, researchers must first have an NIH eRA Commons account for authentication, after which they will need to obtain authorization (via an active DCFS login account) to access the data in the NIH dbGaP.

Below is a guide to help you understand how these platforms are used to connect different components of a CCDI study.

Platform	Data Types
database of Genotypes and Phenotypes (dbGaP)	CCDI study information, list of CCDI study subject IDs, sample IDs, and consents to register the controlled-access studies in dbGaP system.
Data Commons Framework Services (DCFS)	Globally Unique Identifiers (GUIDs) for digital objects and authentication and authorization services.
CCDI Hub Explore Dashboard	Basic deidentified information on participant, samples, files, etc. to build cohorts.

Cancer Genomics Cloud (CGC)	Tools, computing resources, etc. to analyze the data.

database of Genotypes and Phenotypes (dbGaP)

CCDI studies with controlled-access data are registered with the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGaP), which maintains a list of the studies' subject IDs, sample IDs, and consents.

Eligible investigators interested in obtaining a controlled data set should watch the instructional video on applying for controlled access data and consult the Tips on Preparing a Successful Data Access Request. A step-by-step breakdown of the data access request process is located in this guide.

NCI Data Commons Framework Services (DCFS)

Data Commons Framework Services (DCFS), powered by Gen3, facilitates data authorization in a secure and scalable manner. DCFS's Indexd service provides permanent digital IDs for data objects. These IDs can be used to retrieve the data or query the metadata associated with the object.

CCDI Hub Explore Dashboard

The CCDI Hub Explore Dashboard is a tool that allows for the exploration of individual-level participant, sample, and file information for CCDI-managed data sets. The Explore Dashboard enables researchers to find CCDI data within a single study or across multiple studies and create synthetic cohorts based on filtered metrics (i.e., demographics, diagnosis, samples, etc.). Users can review the open-access information and determine which data sets are applicable to their research questions. To access the controlled data, users must request them at the controlled-access login page.

Finding Participants, Samples, and Files

The CCDI Hub Explore Dashboard provides row-level metadata for CCDI study participants and their data objects for review with a filtered search, select visualizations, and an exportable table of results. Here's how to find and filter information on the Explore Dashboard:

1. From the CCDI Hub, navigate to the Explore Dashboard (Figure 1).

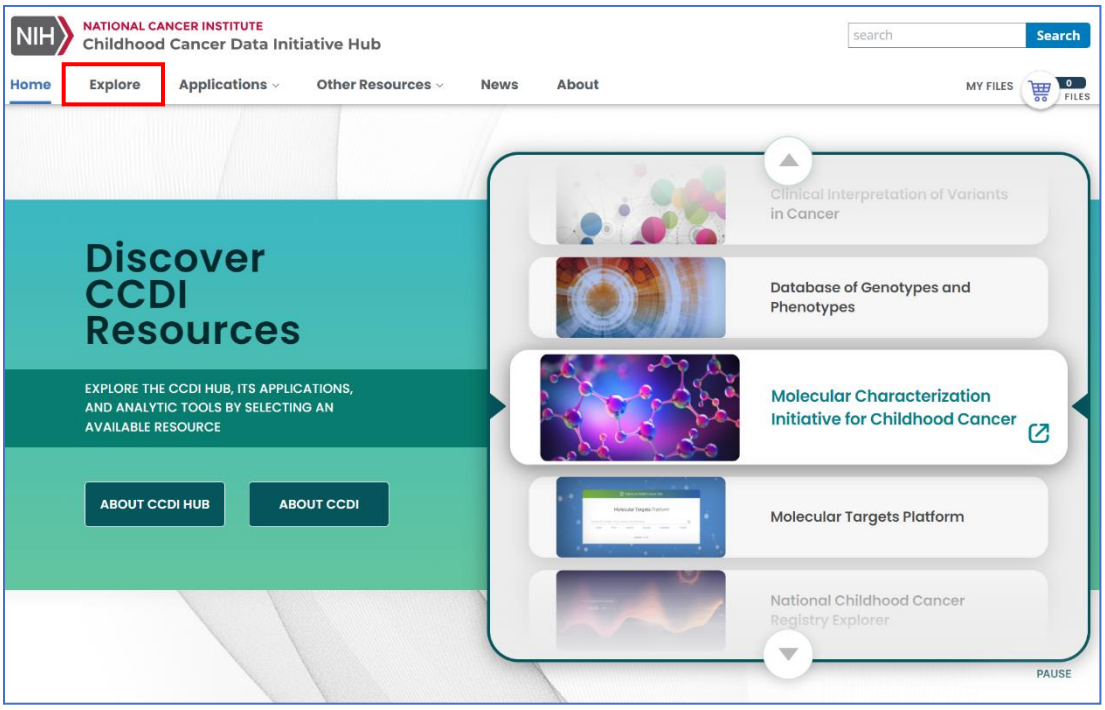

Figure 1: CCDI homepage with red box highlighting the Explore Dashboard menu bar link

2. On the Explore Dashboard, you can filter row-level data and view them as visualizations (Figure 2). The Explore Dashboard is participant-centric, meaning that filtering criteria and results return deidentified information about a participant in addition to information about the participant's collected samples or created file.

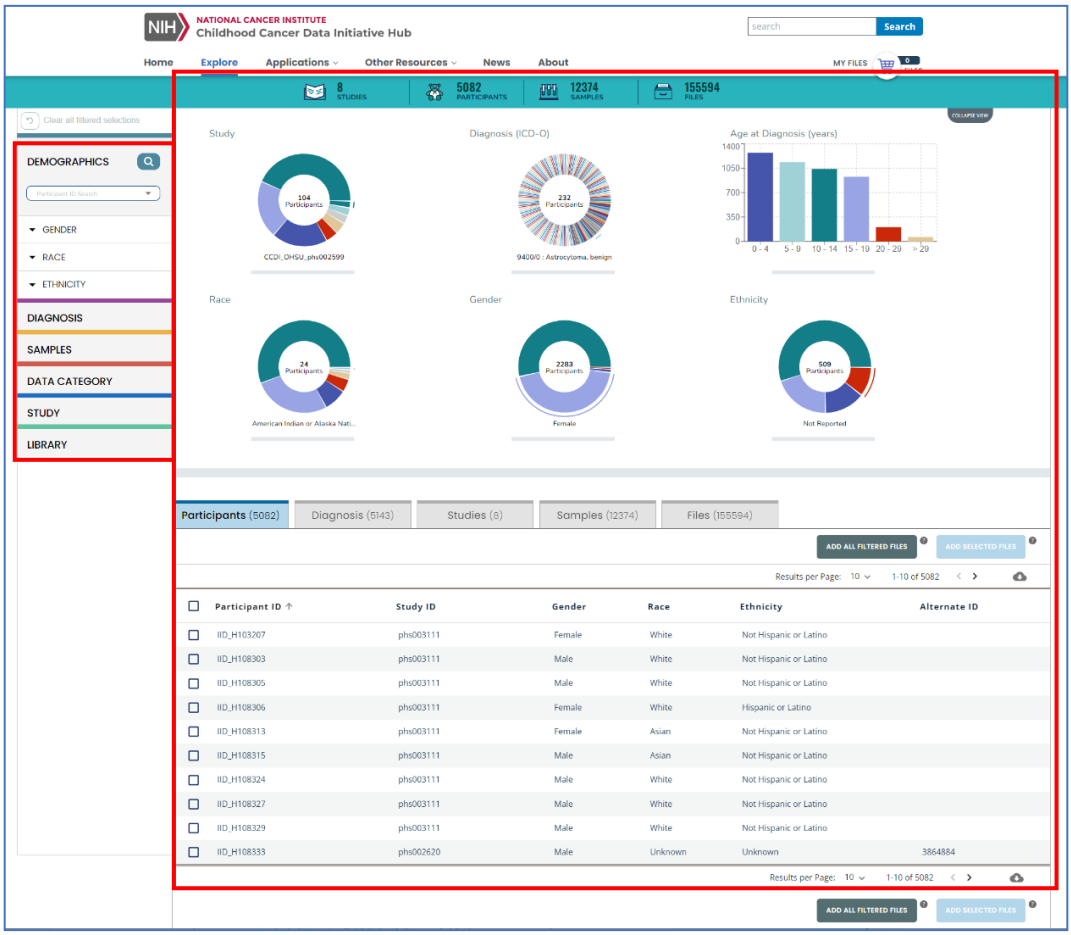

Figure 2: Explore Dashboard page with red boxes highlighting the search filters and results

3. You can apply multiple filtering criteria at the same time in a search (Figure 3).

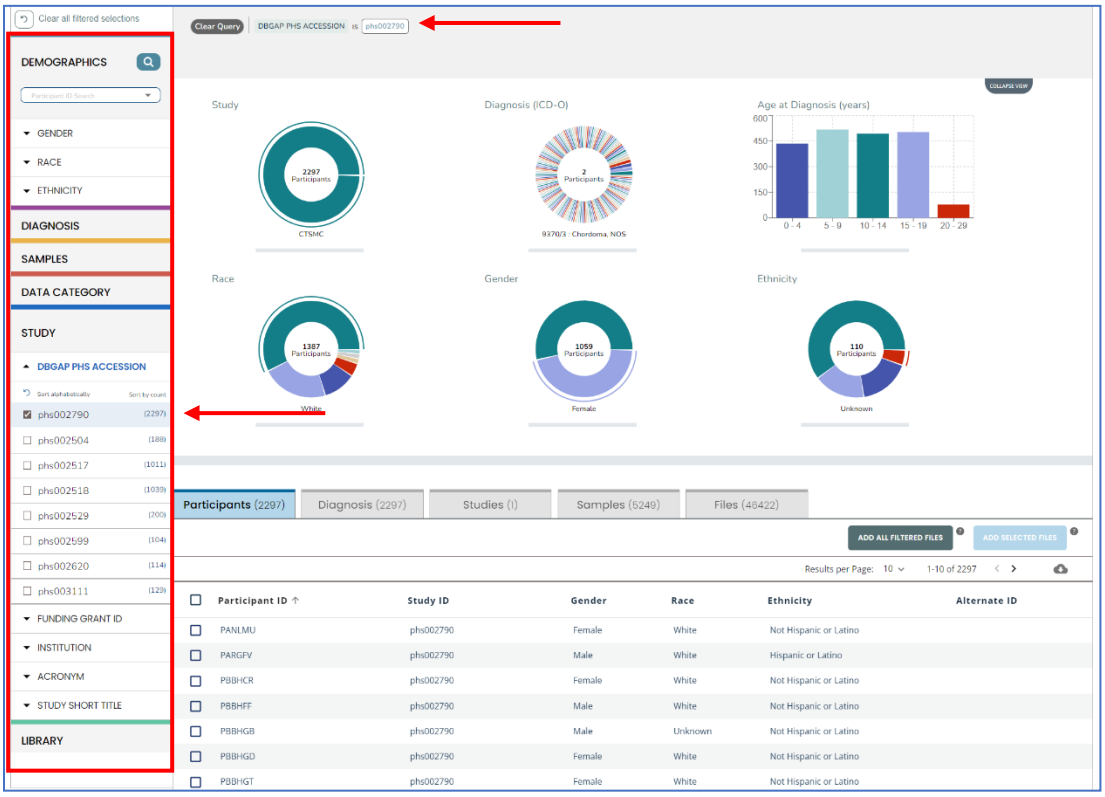

Figure 3: A filtered search for study phs002790

4. Filtering your search will update the Explore Dashboard's visualizations and the results tables (Figure 4). Each results table will be updated with information on the participants, samples, or files that meet the filtered criteria. Information on each table is described below:

    a. **Studies**: Studies that are a part of the Explore Dashboard. Participants, samples, and files all belong to a CCDI study.
    b. **Participants and Diagnosis**: Characteristics of a participant in the Explore Dashboard. Participants belong to a study, and they may have one or more samples or files associated with them.
    c. **Samples**: Samples available from participants within the Explore Dashboard. Samples belong to a participant and can be associated with one or more files.
    d. **Files**: Files available from studies, participants, and samples within the Explore Dashboard. Files may belong to a study and may be associated with one or more participants or samples.

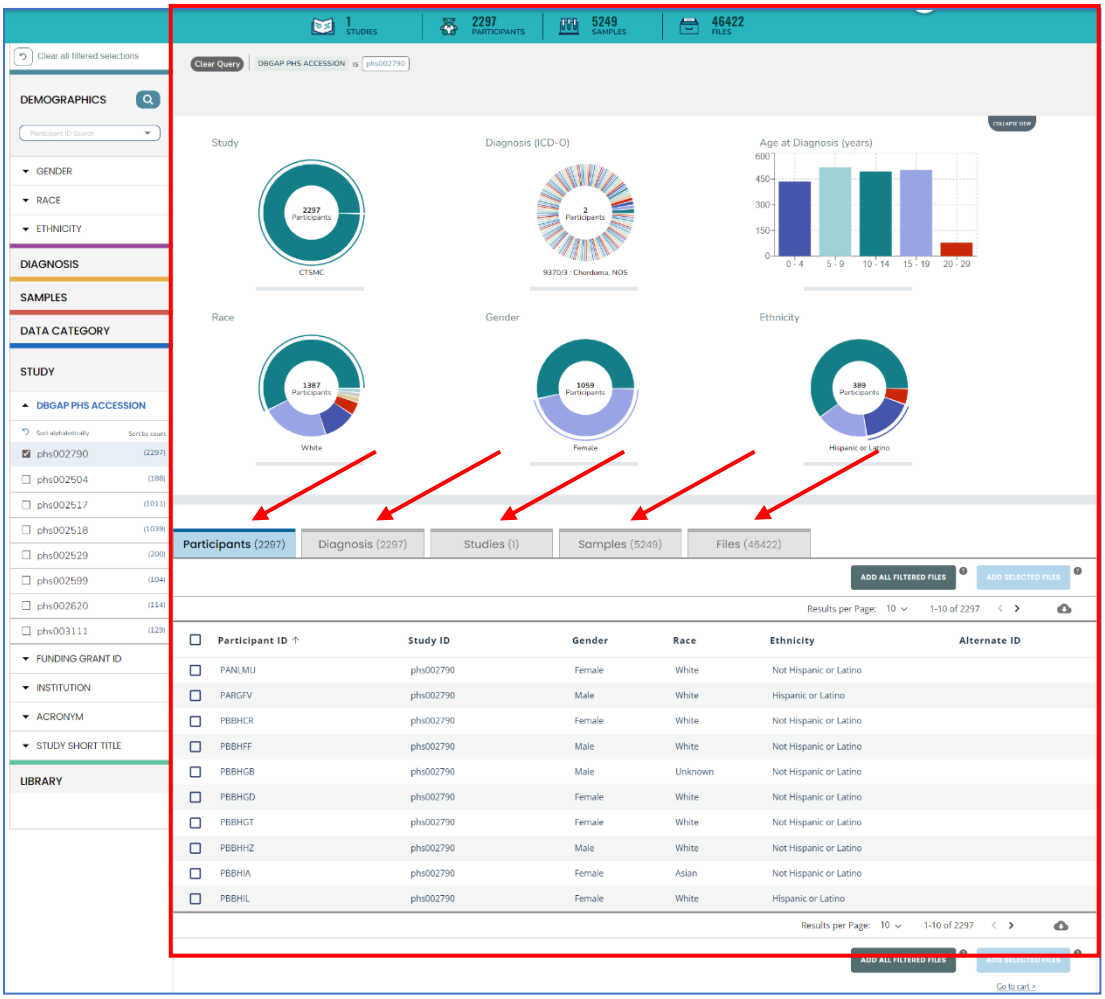

Figure 4: Explore Dashboard visualizations and results tables with arrows pointing to the available informational table

Creating an Exportable File Manifest

From the CCDI Hub Explore Dashboard, you can export each row-level metadata element for CCDI participants, samples, or files. Here's how to create a manifest file of filtered information on the Explore Dashboard:

1. On the results tables of the Explore Dashboard, you can select a row of metadata using the checkbox at the start of the row. Multiple rows can be selected within a table, even across pages of the table. Use the checkbox at the top of the checkbox column to select or deselect all rows. After selecting desired rows, add files for that element to the My Files shopping cart (Figure 5) by clicking the "Add Selected Files" or "Add All Filtered Files" button. Rows of each table add different files to the cart:

    a. **Participants and Diagnosis**: Selecting an item means every file associated with a participant will be added to the My Files shopping cart.
    b. **Studies**: Selecting an item means every file in a study will be added to the My Files shopping cart. For every participant in a study, every associated file and any clinical files associated with the study will also be added.
    c. **Samples**: Selecting an item means every file associated with a collected sample will be added to the My Files shopping cart.

d. **Files**: Adds a selected file.

You can also add all the files for a specific cancer, across all pages of files available, to the My Files shopping cart by clicking the “Add All Filtered Files” button.

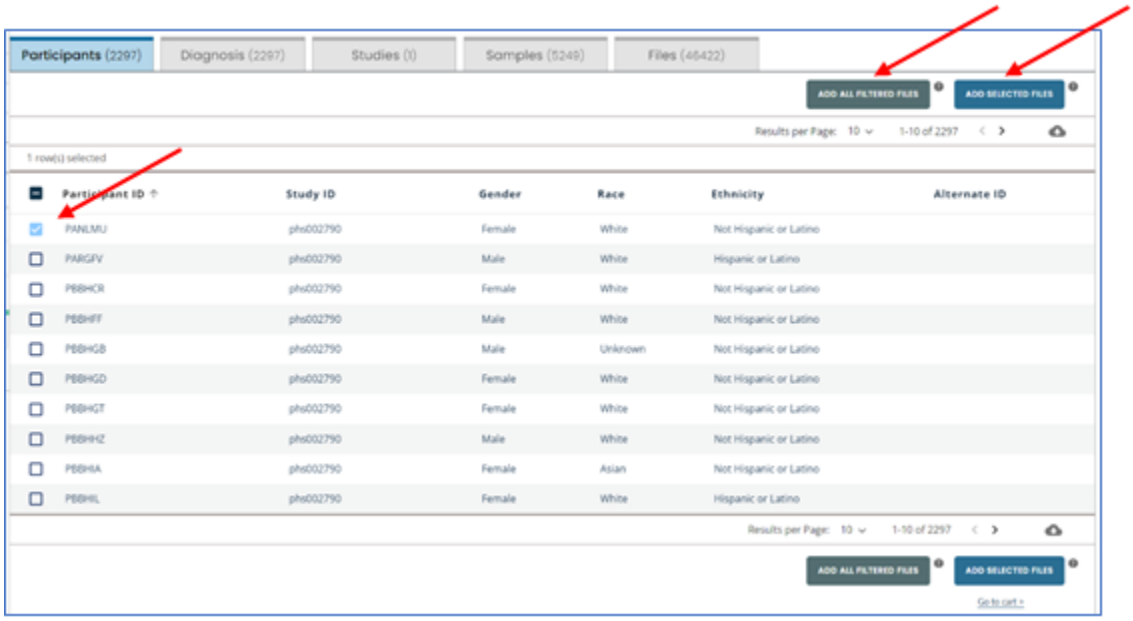

Figure 5: Selection checkboxes and buttons to add files to the cart for the Participants table

2. To navigate to the shopping cart, select “My Files” or the shopping cart icon on the menu bar (Figure 6).

Figure 6: CCDI Hub menu bar with red box highlighting the My Files shopping cart

3. The shopping cart feature enables you to select and manage files. It’s a simple way to keep track of data and files during your session. Selecting the “Download Manifest” button (Figure 7) will produce a comma-separated values (CSV) file manifest of the items within the cart. You can then upload this manifest file in the Cancer Genomics Cloud (see [Creating a Project in CGC & Importing Explore Dashboard Manifest Files]) or use it locally.

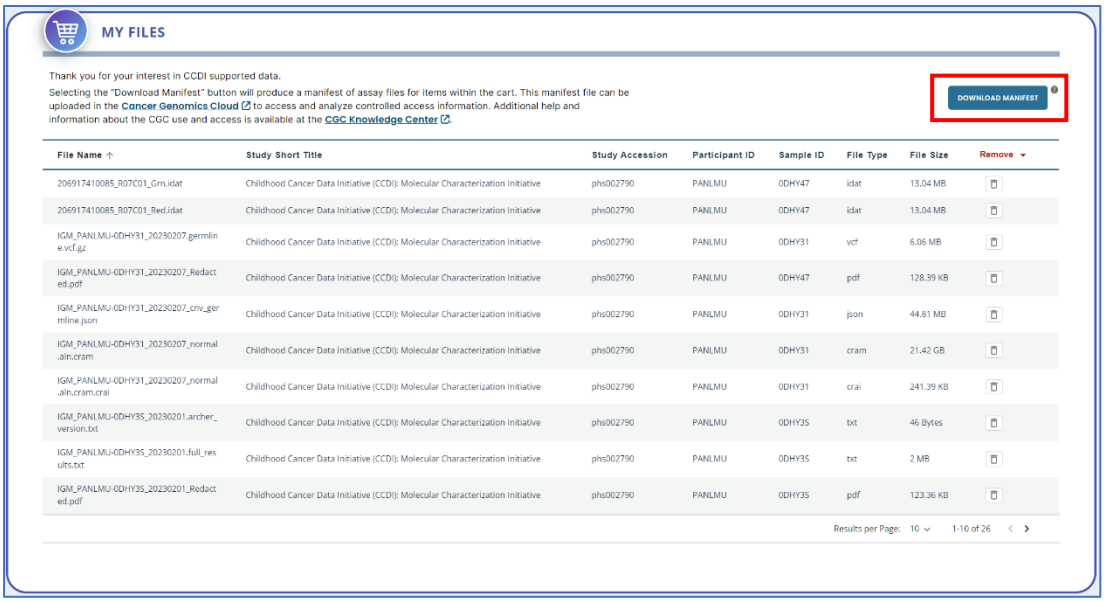

Figure 7: The Explore Dashboard shopping cart page with red box highlighting the "Download Manifest" button

Cancer Genomics Cloud (CGC)

The Seven Bridges Cancer Genomics Cloud (CGC), powered by Velsera and funded by NCI, is a flexible cloud platform that enables analysis, storage, and computation of large cancer data sets. The CGC provides a user-friendly portal to access and analyze cancer data without having to learn how to program.

There are two ways to sign up for the CGC. Please follow the procedure to "Register via an external account" to access dbGaP-regulated controlled data. Velsera also hosts office hours to answer questions about using the CGC site.

Creating a Project in CGC & Importing Explore Dashboard Manifest Files

You can access CCDI data by creating a Project and importing the Explore Dashboard manifest created on the CCDI Hub to the CGC. The following instructions describe the process to create a new project and import the manifest file using the CGC Data Repository Service (DRS) import tool.

1. Under the Projects section, select the “Create a project” button to start a new project. A pop-up will appear with a search bar and short list of studies (Figure 8).

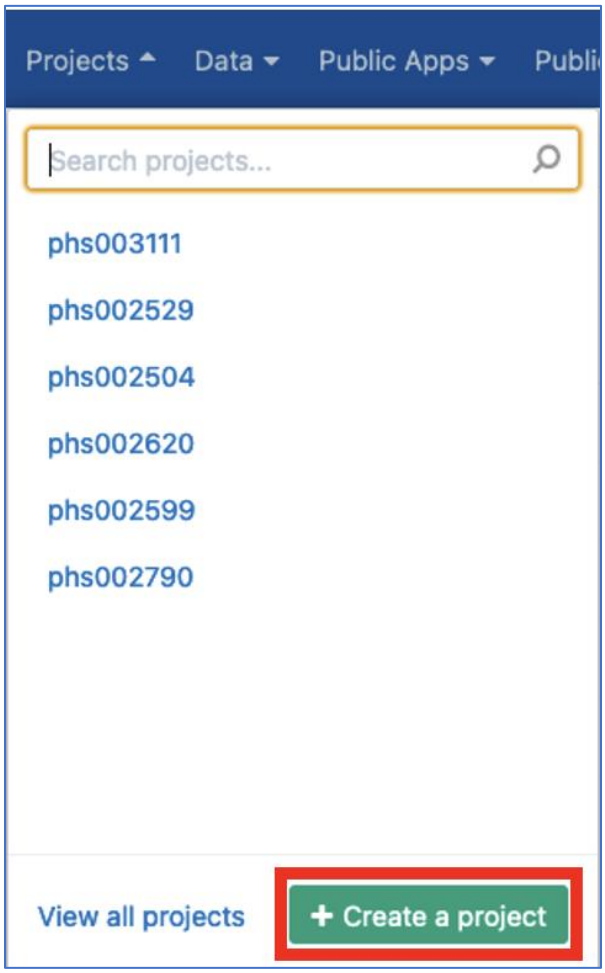

Figure 8: CGC project menu with red box highlighting the “Create a project” button

2. A new project creation prompt will appear (Figure 9). After naming the new project, further options can be selected or left at default values. Once the project configuration is set up, clicking “Create” will take you to the new project dashboard.

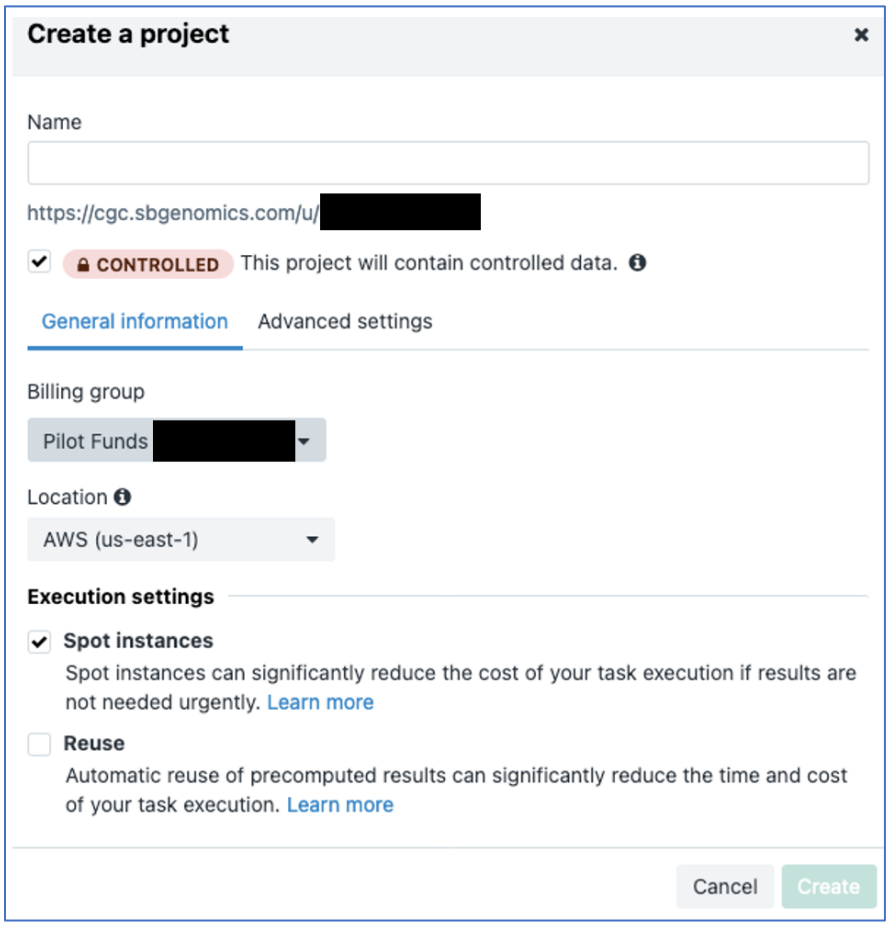

Figure 9: A new project creation prompt

3. On the project's dashboard page, select the "Files" option in the toolbar (Figure 10). You will now see the current files available in the project. For newly created projects, there will only be the options to create a "New folder" or "Add files" (Figure 11).

Figure 10: Project menu bar with "Dashboard" selected

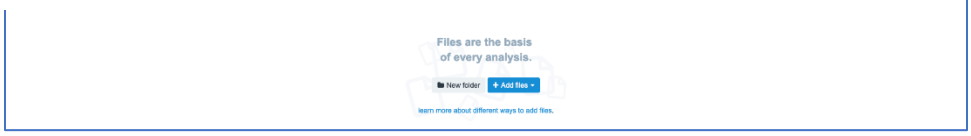

Figure 11: Empty files page of a new project

4. Select the “Add files” button, then “GA4GH Data Repository Service (DRS)” (Figure 12).

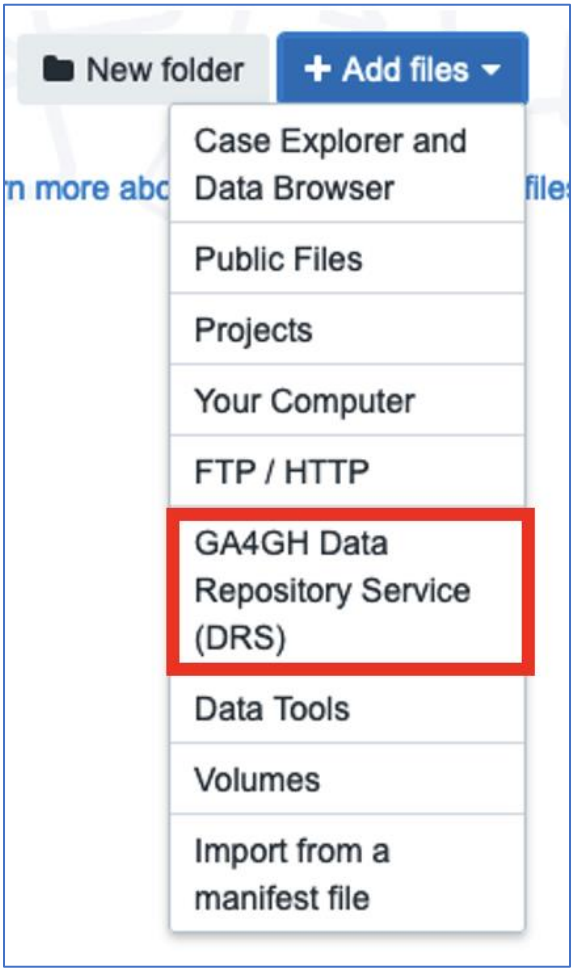

Figure 12: “Add files” menu highlighting the DRS import option

5. Select the “From a manifest file” tab in the toolbar and upload your manifest (Figure 13).

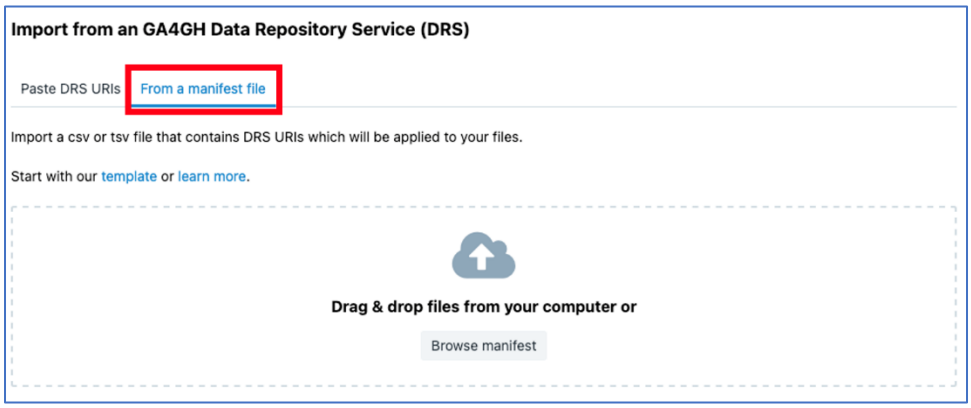

Figure 13: DRS Import prompt for importing with a manifest

6. After the file is read in and verified as a DRS manifest, you will have the option to add tags to the files and resolve possible naming conflicts (Figure 14). Finally, check the box that notes you understand and will follow the data use agreements, then click “Submit.”

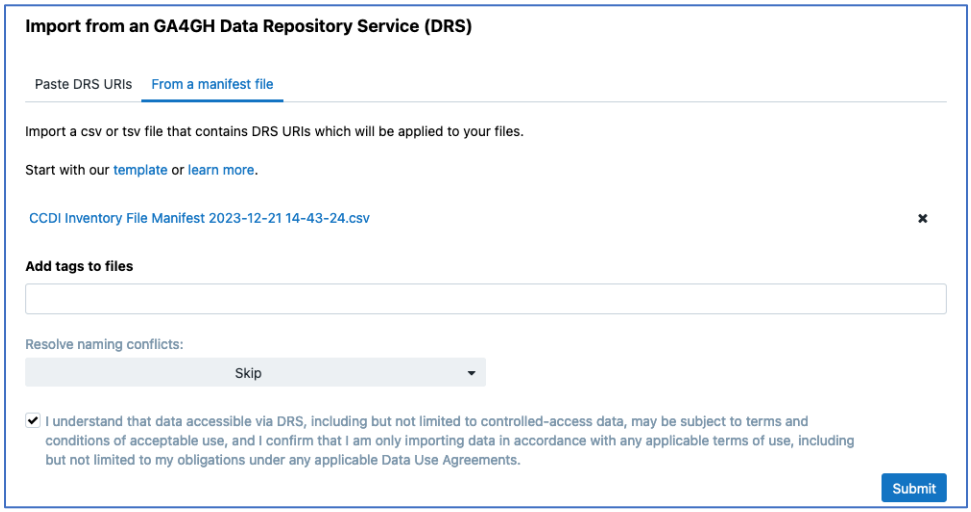

Figure 14: Import prompt for DRS for an uploaded manifest file with a field to add tags

7. Clicking submit will take you back to the project “Files” section, where you will now see the manifest in the list. The files within the manifest will soon populate within the “Files” section.

Using CGC Data Studio

After creating a CGC project, you can 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 15).

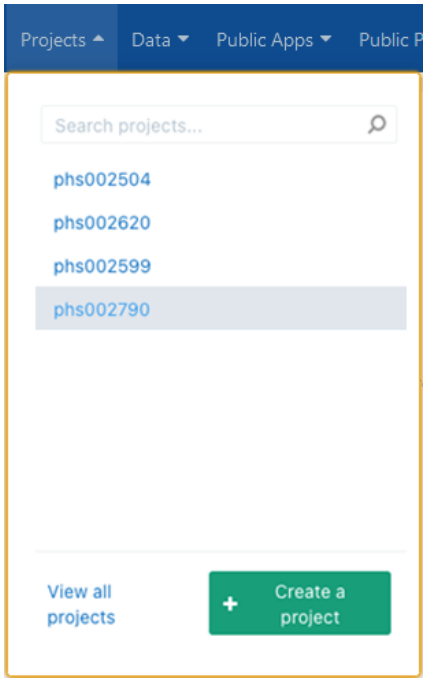

Figure 15: “Projects” drop-down with list of example projects with a highlighted selection

2. Once in the project, select “Data Studio” at the top of the screen and then click “Create new analysis” (Figure 16).

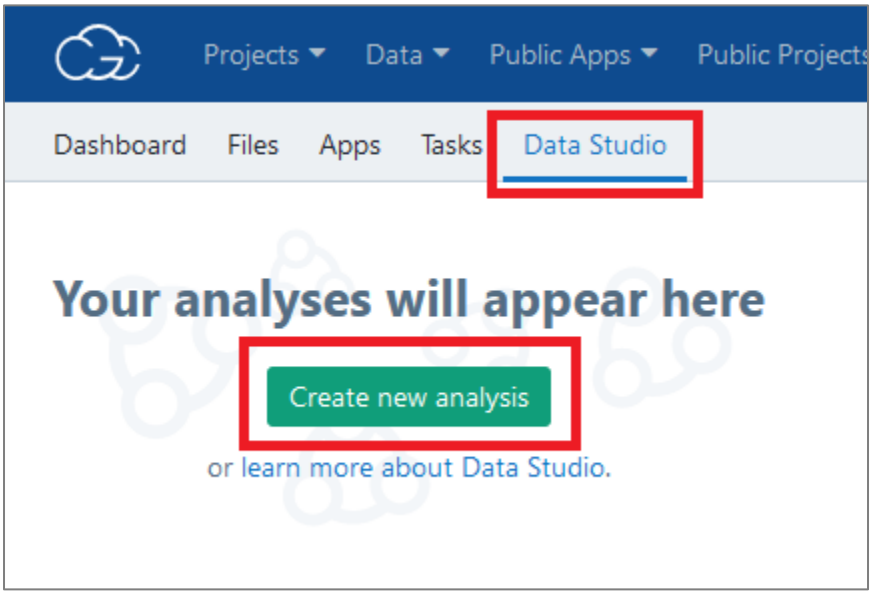

Figure 16: Data Studio menu page with “Create new analysis” button highlighted in red

3. Enter a name for the analysis, select “RStudio” or “JupyterLab” under “Environment,” and click “Start” at the bottom of the dialog box (Figure 17).

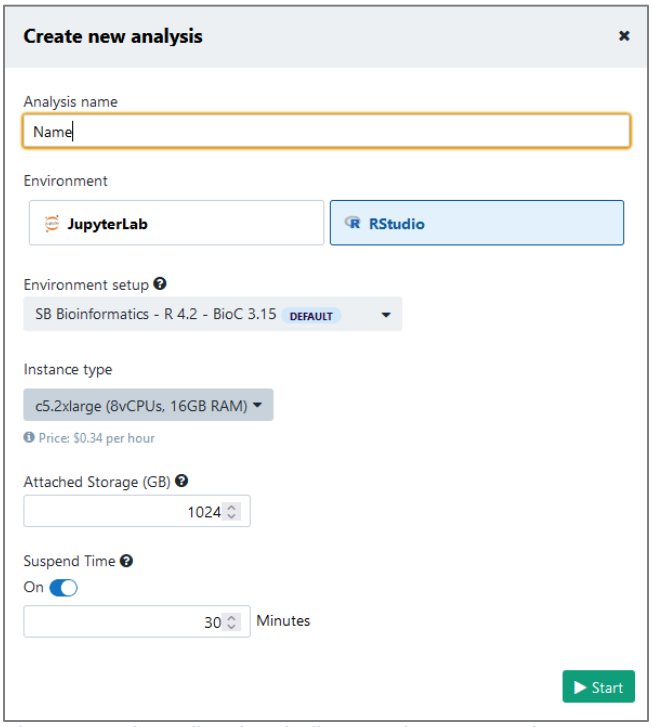

Figure 17: The yellow box indicates where to put the name of a new RStudio analysis.

4. A new workspace environment will be created (Figure 18).

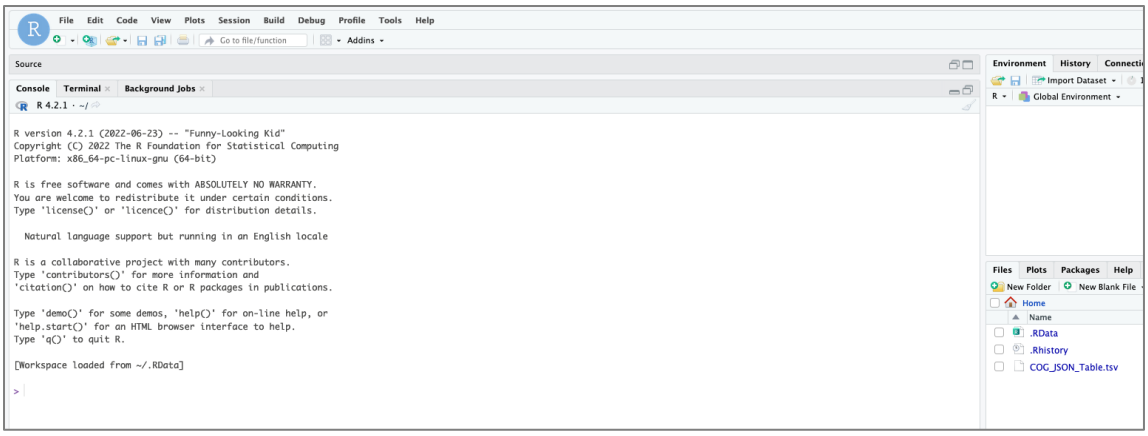

Figure 18: RStudio integrated development environment

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 to indicate when to stop the analysis or adjust its duration.

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 on Working with Data at the CGC

- CGC Documentation: https://docs.cancergenomicscloud.org/docs
- Importing CDS Data: https://docs.cancergenomicscloud.org/docs/import-cds-data
- Common Workflow Language Workflows and Apps: https://cgc.sbgenomics.com/public/apps
- Volumes: https://docs.cancergenomicscloud.org/docs/volumes-1
- Tool Editor Tutorial: https://docs.cancergenomicscloud.org/docs/tool-editor-tutorial
- About the Editor: https://docs.cancergenomicscloud.org/docs/about-the-editor

Contact Information

Please direct any questions or requests for further information to the CCDI mailbox.

THIS PAGE IS INTENTIONALLY LEFT BLANK

Appendix A: Creating a CGC Account

You need an account to access and analyze CCDI data on the CGC platform. Note that a list of all CCDI studies released is also available. The following instructions describe the process to create a CGC account.

1. From the CGC home page at cancergenomicscloud.org, click “REGISTER” or “LAUNCH” in the center of the page (Figure A1).

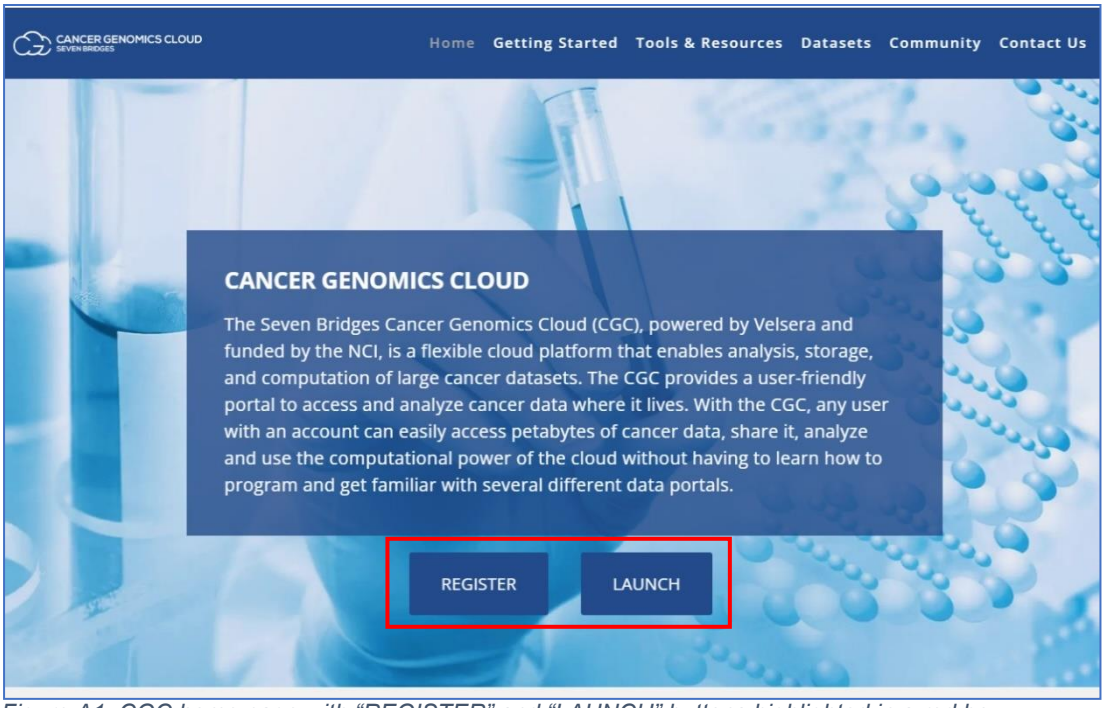

Figure A1: CGC home page with “REGISTER” and “LAUNCH” buttons highlighted in a red box.

2. On the login screen, click on “Log in with eRA Commons” (Figure A2).

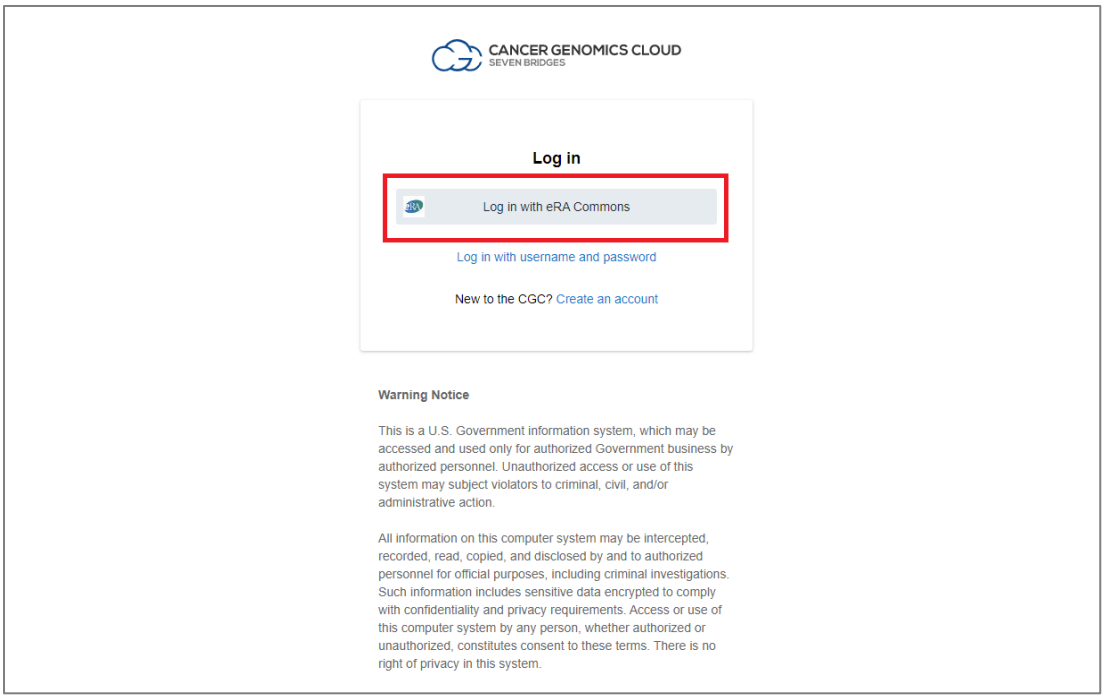

Figure A2: eRA Commons login screen for the CGC, with login button highlighted in a red box

3. On the login screen, enter the eRA Commons account credentials associated with your approved dbGaP study and click “Sign In” (Figure A3).

   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 “incognito” mode to ignore cached data and cookies so 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.

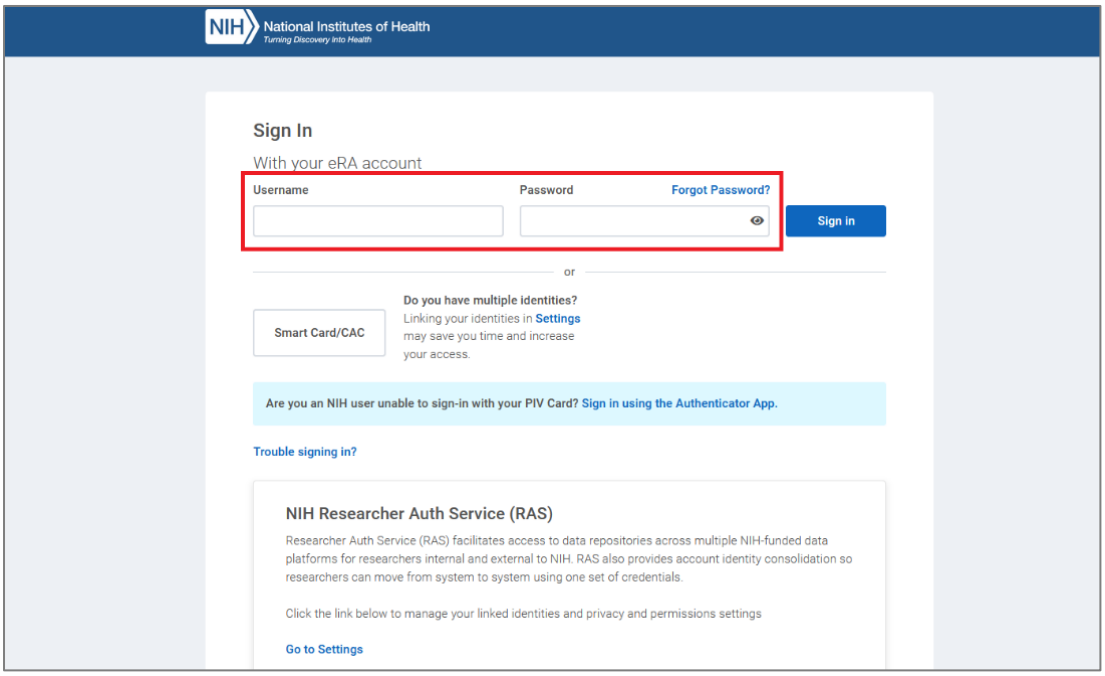

Figure A3: Login page for CGC, with username and password credentials sections highlighted in a red box.

4. If you agree to the "Consent to Share Information" on the following page, click the "Grant" button to continue (Figure A4).

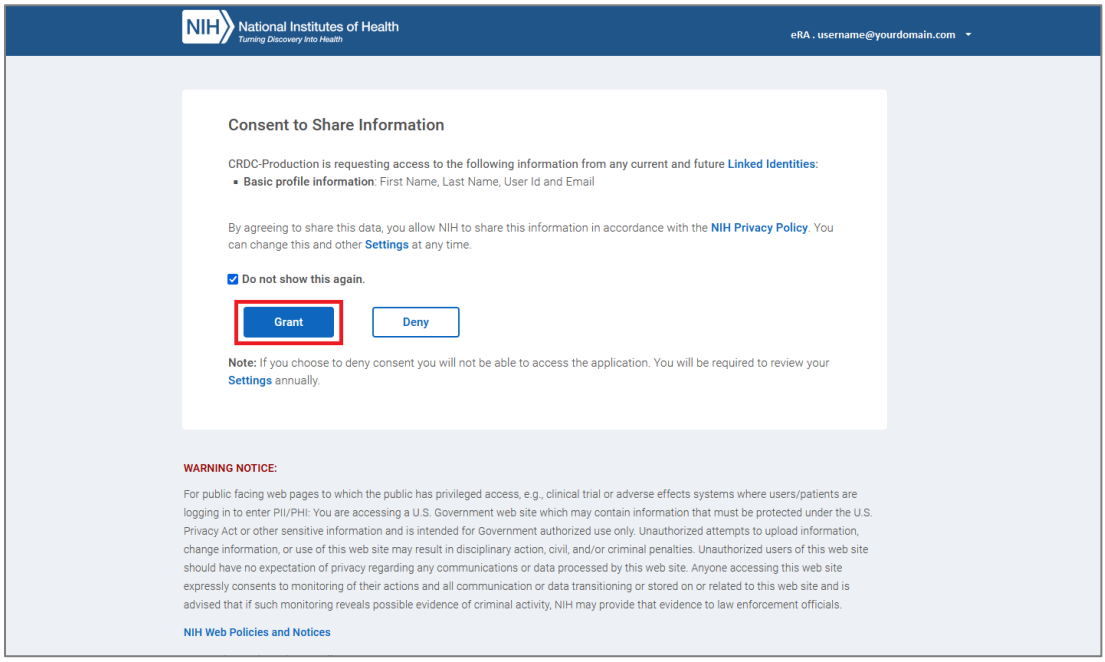

Figure A4: Consent to share information page with red box highlighting the "Grant" button to confirm consent to share information with the CGC

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 A5). Note that Gen3 is an authorization system that uses eRA Commons as an authentication tool and allows access to the CGC system.

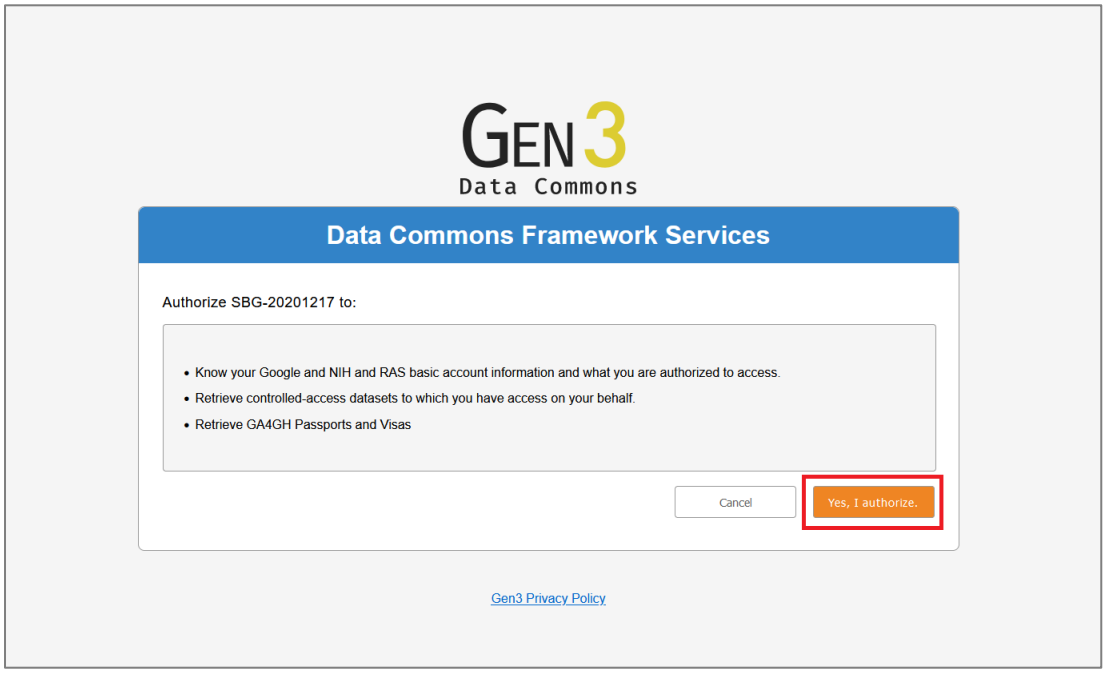

Figure A5: Gen3 Data Commons Framework Services authorization page with "Yes, I authorize" button highlighted in a red box

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 A6).

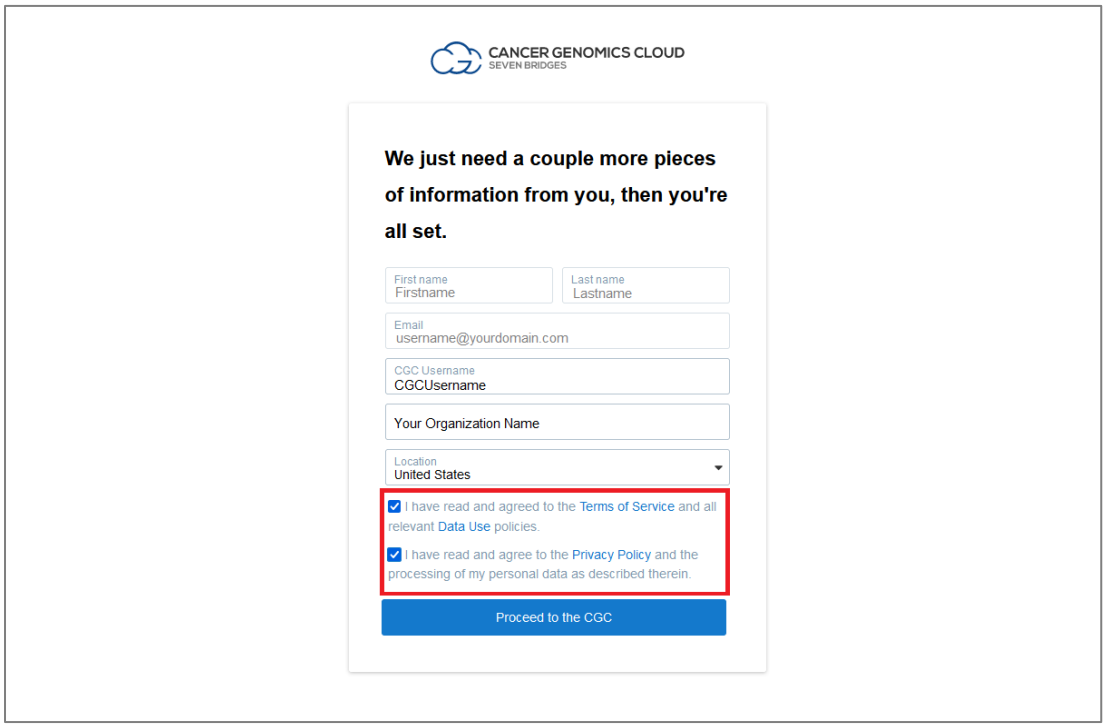

Figure A6: Confirmation of terms and policy for CGC registration highlighted in a red box

7. If the CGC questionnaire appears, complete it to continue (Figure A7).

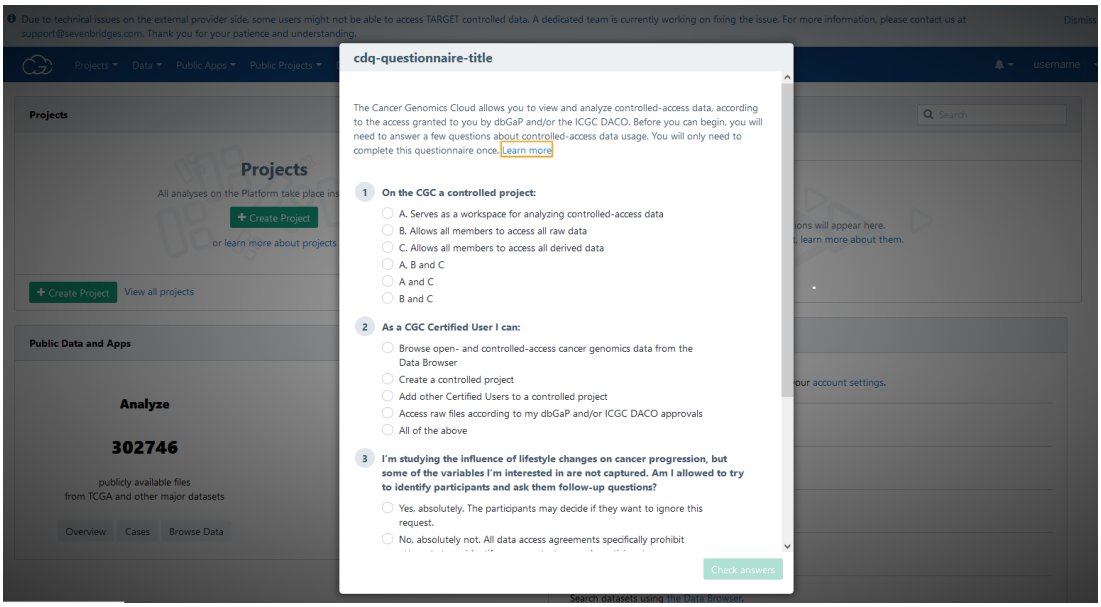

Figure A7: Screenshot of the CGC questionnaire

Appendix B: Using the CGC Cancer Data Service Explorer to Identify Files

As an alternative to the CCDI Hub Explore Dashboard, you can use the CGC Cancer Data Service (CDS) Explorer to identify files for analysis and exploration.

The below instructions for using CDS Explorer assume you have a CGC account and access permissions for the data and files. For information on creating a CGC account, see Appendix A.

1. Click on the “Data” drop-down at the top left of the CGC home page and then select “Cancer Data Service Explorer” (Figure B1).

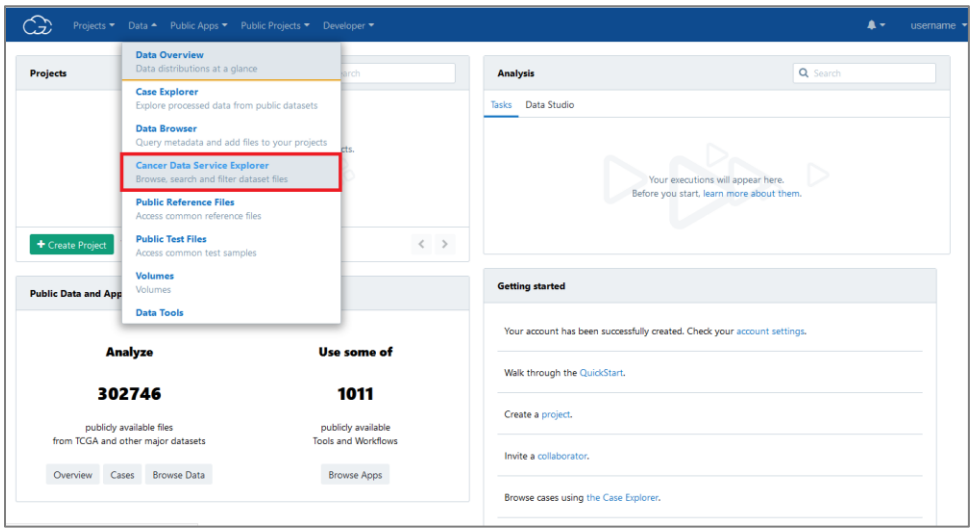

Figure B1: CGC data drop-down menu with red box around Cancer Data Services Explorer

2. CCDI studies are marked with “(CCDI)” at the end of the study name. Click on the “PHS002790” study link to view basic CCDI Molecular Characterization Initiative study information on its dbGaP study page (Figure B2).

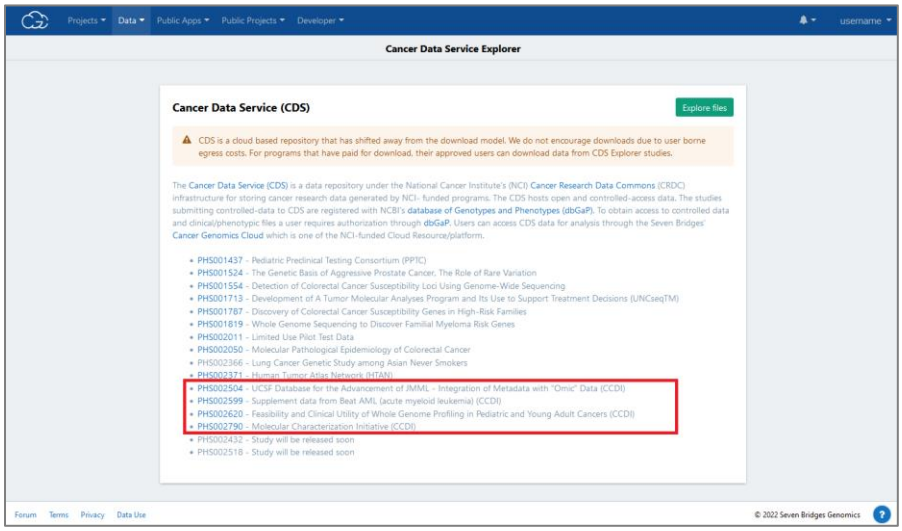

Figure B2: CDS Explorer study list page with red box around CCDI studies

3. 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 B3).

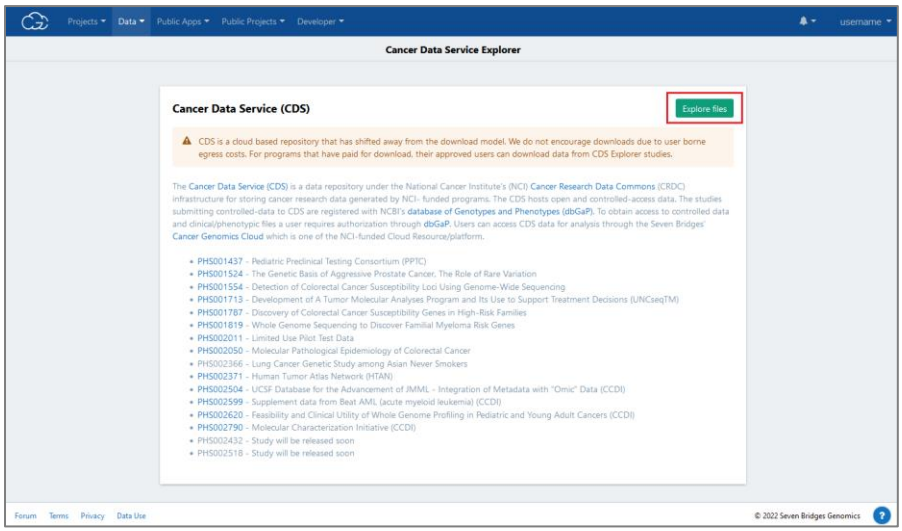

Figure B3: CDS Explorer study list page with red box highlighting “Explore files” button

4. Use the filters 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 B4).

   You may use any other data set filters 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.

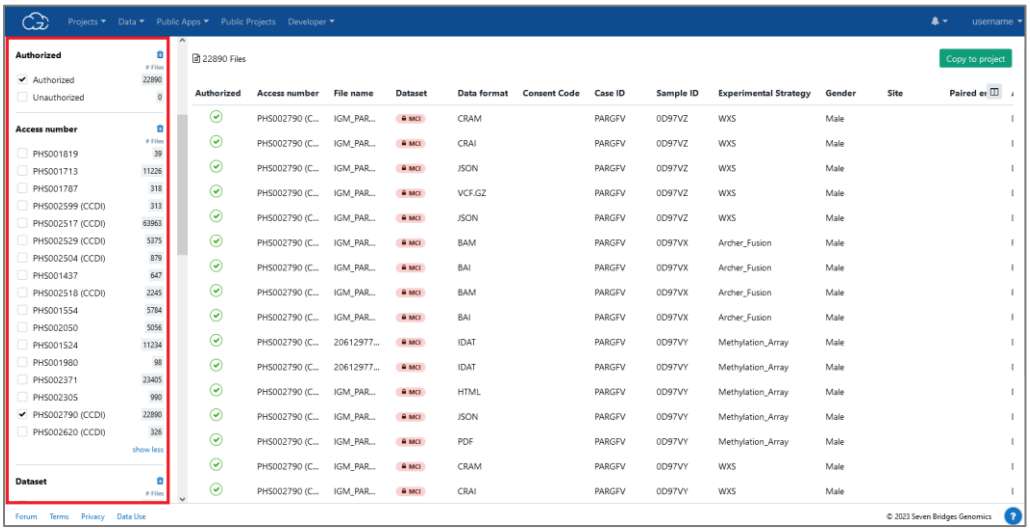

Figure B4: CDS Explorer page with red box around left column showing CDS Explorer filters

5. Once you’ve narrowed the data set based on your selections, you can click the “Copy to project” button at the top right of the page to add your data to a study (Figure B5).

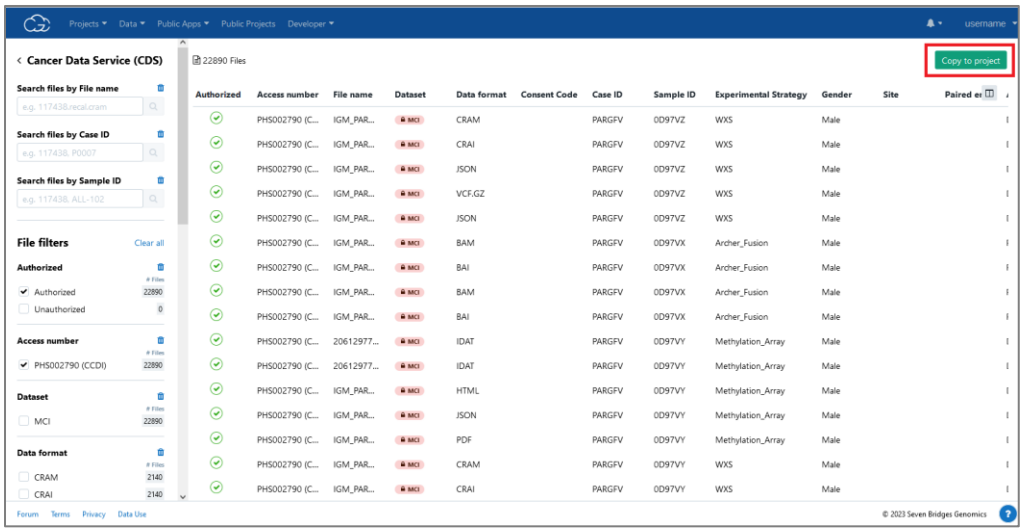

Figure B5: CDS Explorer page with red box highlighting "Copy to project" button (upper right) that will copy selected files to a CGC project

6. 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 B6).

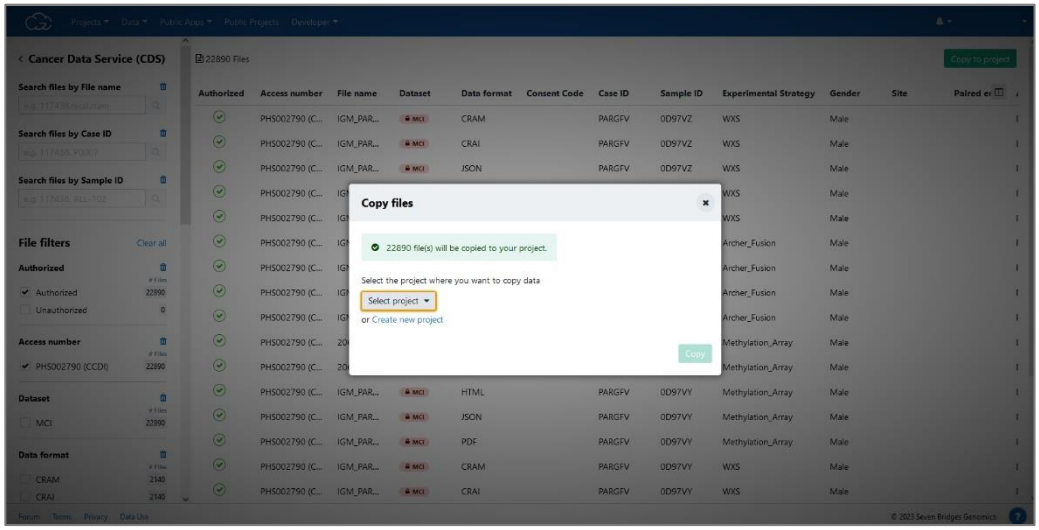

Figure B6: Pop-up showing a drop-down menu with option to copy selected files to a new or existing CGC project

Appendix C: Data Commons Framework Services (DCFS): Controlled Data Access Instructions

CCDI data is available for download using the Data Commons Framework Services (DCFS). To gain access to controlled data, researchers must first have an NIH eRA Commons account for authentication, after which they will need to obtain authorization (via an active DCFS login account) to access the data in the NIH dbGaP.

The below instructions are for using the DCF user interface or the DCF Gen3-client to access CCDI data.

File Download Procedure via User Interface

To download a study-specific research data distribution file with the DCF Services Portal interface, a researcher must execute the following steps:

1. Login to the NCI DCF Services portal - https://nci-crdc.datacommons.io/login (Figure C1).

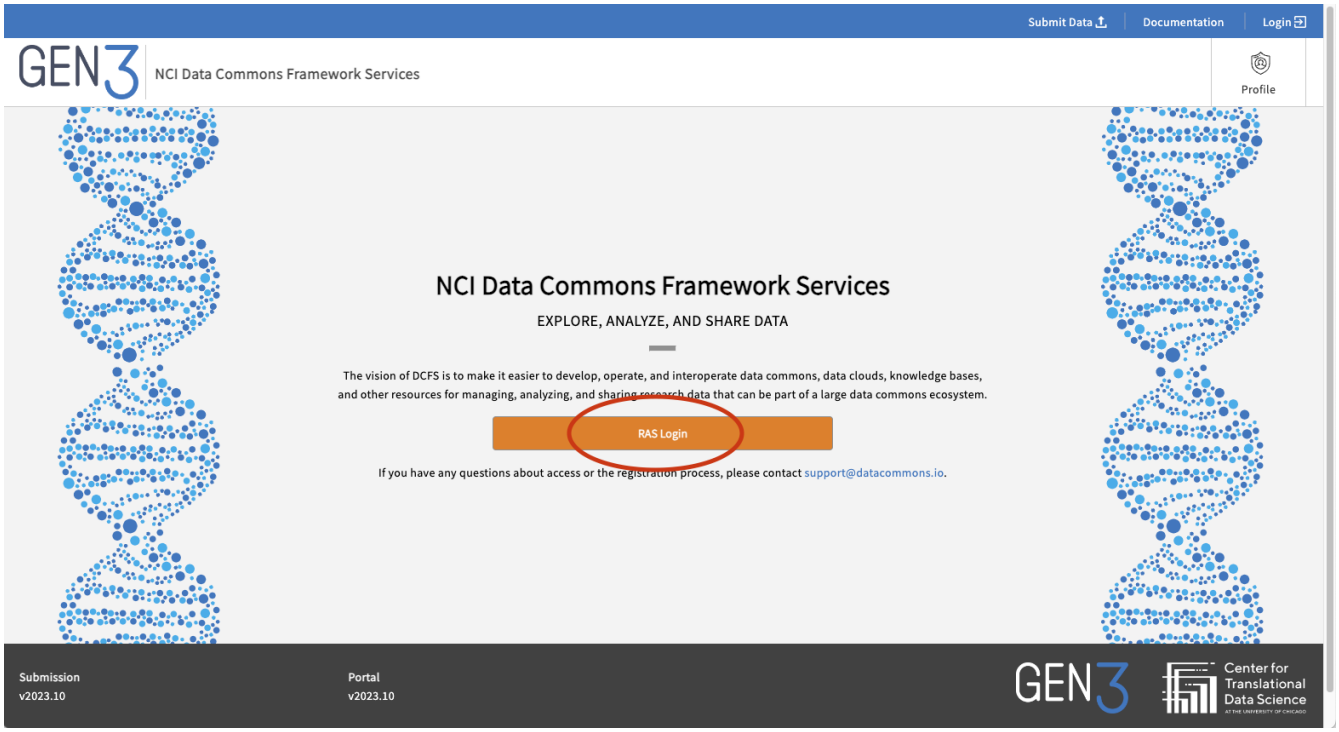

Figure C1: NIH Data Commons Framework (DCF) homepage with the NIH Researcher Auth Service (RAS) login highlighted

2. Once logged in, click the Profile section in the top right corner and review project access to confirm study access (Figure C2).

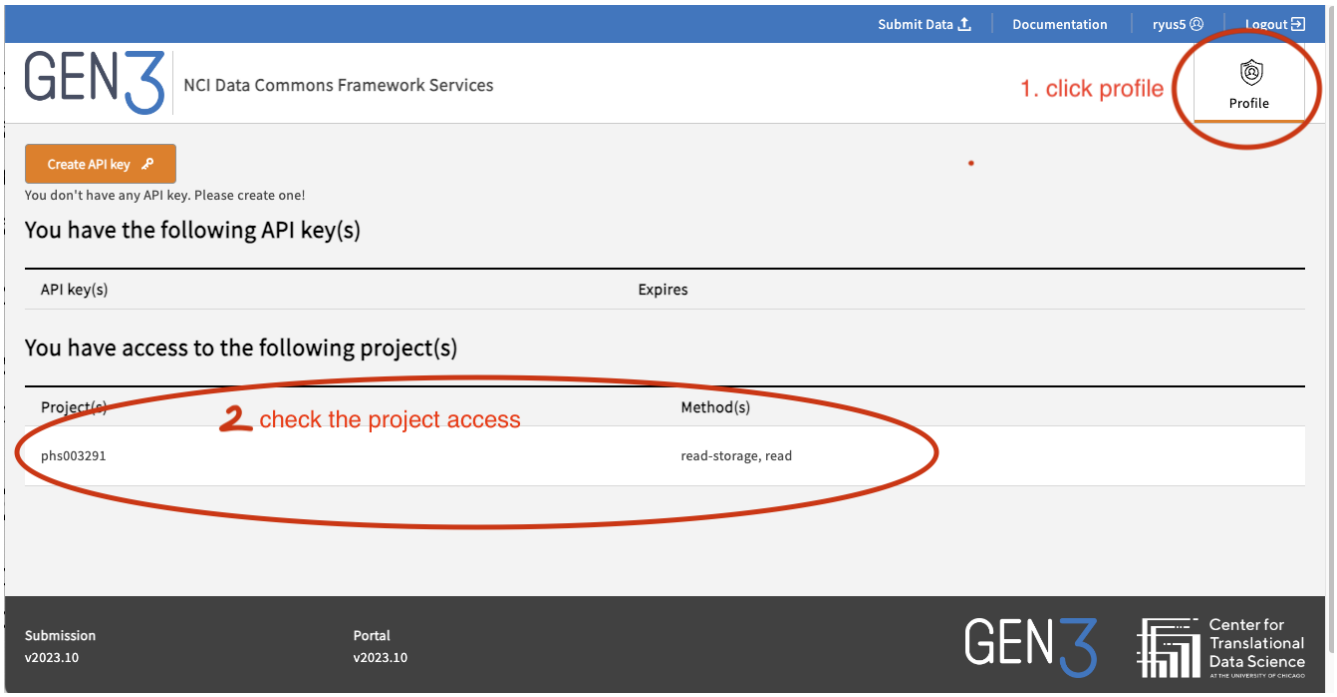

Figure C2: DCF Profile page highlighting Profile and the accessible projects.

3. Paste the study-specific research data distribution file URL from the [Explore Dashboard exportable manifest](), into the browser address field and press Return.

4. The NCI DCF Services Portal will respond by providing a JSON document with a new (signed) URL for the requested data file. Copy the signed URL.

5. Paste this new signed URL from into the browser address field and press Return (Figure C3).
    a. Note: the signed URL provided is valid for a relatively short period of time once issued by the CRDC portal

6. The NCI DCF Services Portal will respond by displaying a URL. Click the URL to download the file (Figure C3).

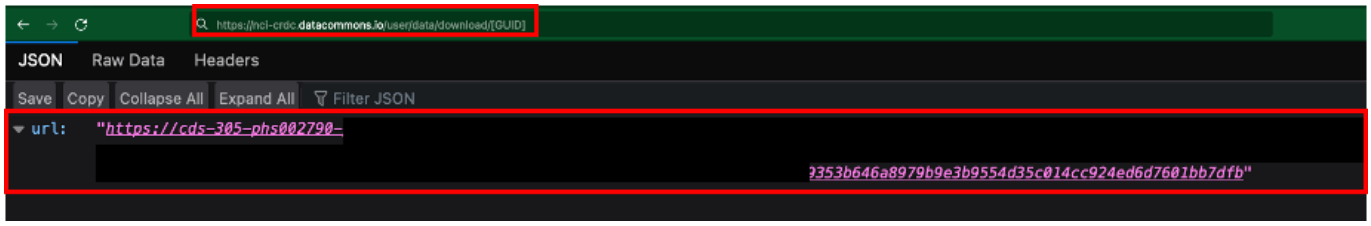

Figure C3: DCF Service Portal displaying the signed access URL and the file download URL

Note: If errors or problems are experienced during the file downloading process above, please contact ncichildhoodcancerdatainitiative@mail.nih.gov for assistance.

File Download Procedure via Call Level Interface (CLI) client

To download a study-specific research data distribution file with a CLI client, a researcher must execute the following steps:

1. Obtain the Gen3-client command-line tool from GitHub.
2. Install and configure the client based on the Gen3 instructions.
    a. These instructions include signing into DCF web client and obtaining a downloaded JSON API key, from the Profile page, and then configuring the client.
    b. The API endpoint that will be used for DCF configuration is 'https://nci-crdc.datacommons.io'.
3. Obtain either a GUID or manifest of GUIDs for the data files of interest from the CCDI Explore page or the Explore Dashboard exportable manifest
4. Created a gen3 structured manifest:

```
[
        {
                "object_id": "dg.4DFC/{guid_1}"
        },
        {
                "object_id": "dg.4DFC/{guid_2}"
        },
        …
        {
                "object_id": "dg.4DFC/{guid_n}"
        }
]
```

5. Using the Gen3 client, either the single or multiple download option, download the file(s).

For more information on this process, please visit the Gen3 Documentation page.