Hub Explore User Guide & Data Access Instructions

2/12/2025

Content

Introduction 2
Data Access 2
CCDI Hub Explore Dashboard and Cart 3
Finding Participants, Diagnoses, Studies, Samples, and Files 4
Downloading Metadata from the Studies tab 6
Creating an Exportable File Manifest from the Cart 7
Contact Information 9
Appendix A: database of Genotypes and Phenotypes (dbGaP) 11
Appendix B: NCI Data Commons Framework Services (DCFS): Controlled Data Access Instructions 12
File Download Procedure via User Interface 12
File Download Procedure via Command Line Interface (CLI) client 14
Appendix C: The Cancer Genomics Cloud (CGC) 15
Creating a CGC Account 15
Creating a DRS Manifest from downloaded Study metadata 19
Additional Resources on Working with Data at the CGC 19

Introduction

This document provides information about how to use the Childhood Cancer Data Initiative (CCDI) Hub Explore Dashboard, including how to find, download, and export manifests for open CCDI data, and submit requests for controlled access data.

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

Additional CCDI applications and other resources are described in informational pages and available through the various Hub menus. Related resources—internal and external to CCDI—are linked where relevant. Application and data update summaries and other announcements are available to browse from the News page, and user documentation details general site usage and data use terms and best practices.

A list of all CCDI studies released is available through the CCDI Childhood Cancer Data Catalog.

Data Access

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 participants' informed consent, and the confidentiality of the data and the privacy of participants are maintained.

For 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: CCDI Childhood Cancer Data Catalog, CCDI Molecular Targets Platform

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

Example: CCDI National Childhood Cancer Registry Explorer

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

Example: CCDI genomic data stored at the Cancer Data Service

To gain access to controlled data, researchers must:

1. Have an NIH eRA Commons account for authentication.
2. Request authorization by the NIH database of Genotypes and Phenotypes (dbGaP). Authorization is enforced by the Data Commons Framework Services (DCFS), whether accessing directly from DCFS or through the CGC. A step-by-step breakdown of the data access request process is located in this guide.

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

Platform	Data Types
CCDI Hub Explore Dashboard	Basic de-identified information on participant, samples, files, etc., to build cohorts
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
Cancer Genomics Cloud (CGC)	Tools, computing resources, etc., to analyze the data

Please note that additional information on using these resources for accessing or analyzing CCDI data is provided in the Appendices below; however, the most up-to-date documentation at any time will live with these external resources.

Reach out to the CCDI mailbox with any questions.

CCDI Hub Explore Dashboard and Cart

The CCDI Hub Explore Dashboard is a tool that allows for the exploration of participant-level, diagnoses, studies, samples, and files 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 search (i.e., demographics, diagnosis, samples, etc.). Upon interaction with these filters (Figure 1A), users can review the open-access information through visual summaries (Figure 1B) and browse the row level data in tabs organized by participants, diagnosis, studies, samples, and files (Figure 1C) to determine which data sets are applicable to their research questions. Users can then add desired files to the cart (Figure 1D), from which they can download a manifest for the selected data or take the manifest directly into the CGC. To access the controlled data, users must request them at the controlled-access login page on dbGaP.

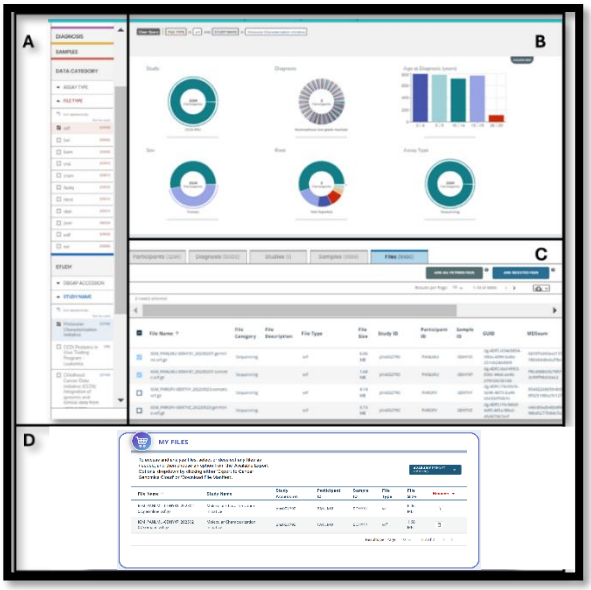


Figure 1: CCDI Hub Explore Dashboard and Cart features

Step-by-step instructions for finding and exporting data are included below.

Finding Participants, Diagnoses, Studies, 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. The CCDI Hub is located at ccdi.cancer.gov. From the CCDI Hub Home page, navigate to the Explore Dashboard by clicking "Explore" (Figure 2).

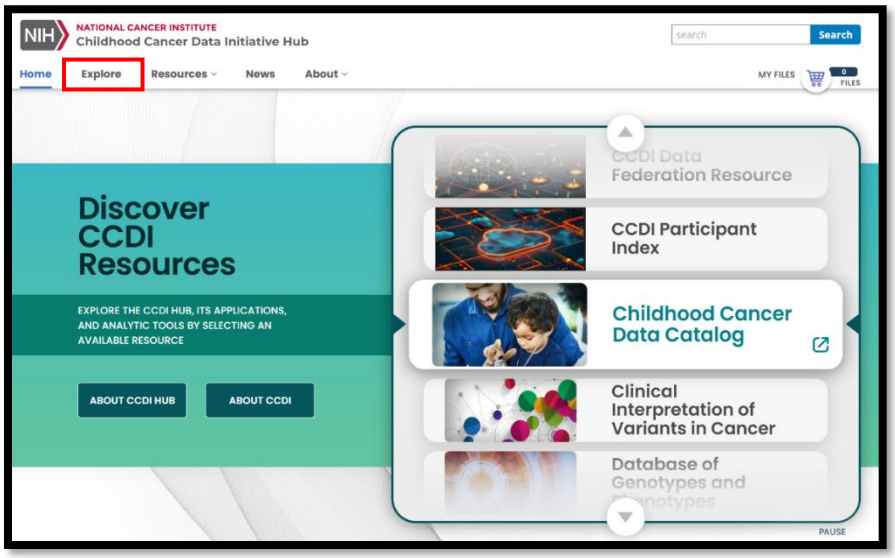


Figure 2: CCDI homepage with red box highlighting the "Explore" menu bar link

2. On the Explore Dashboard, you can filter row-level data and view them as visualizations (Figure 3). The Explore Dashboard is participant-centric, meaning that filtering criteria and results return de-identified information about a participant and their related studies, diagnoses, collected samples, or created file.

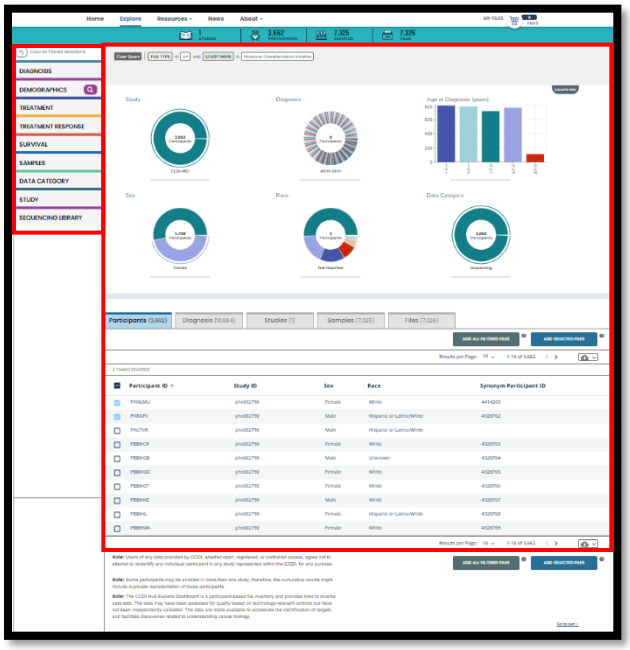


Figure 3: Explore Dashboard page with red boxes highlighting the search filters and results

3. Search criteria are displayed in the right panel (Figure 4A). Faceted filtering may be done by uploading a list of participant IDs (in “DEMOGRAPHICS” Figure 4B), text searches (“DIAGNOSIS,” “DIAGNOSIS ATOMIC SITE,” and “SAMPLE ANATOMIC SITE” Figure 4C), numerical sliders (“AGE AT DIAGNOSIS” and “AGE AT COLLECTION” Figure 4D), or checkbox selections for the remaining properties. You can apply multiple filtering criteria at the same time in a search. You can view and clear your current selection(s) in the query summary at the top of the widgets (Figure 4E).

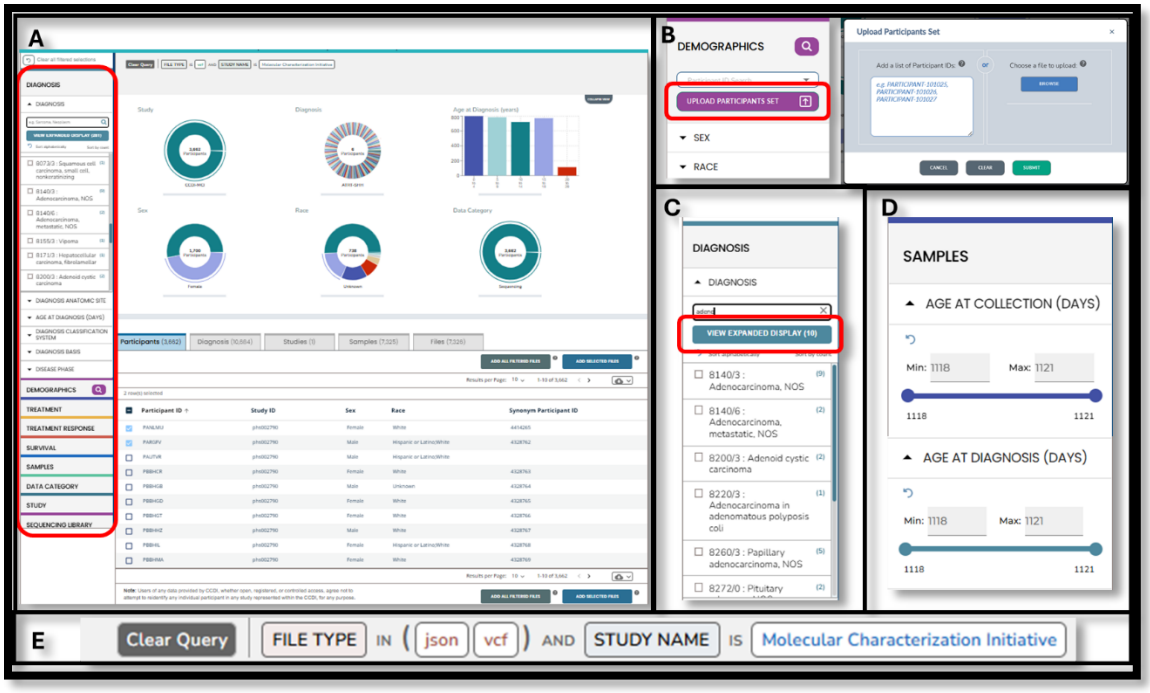


Figure 4: Full facet list in Explore Dashboard with highlights of various facet types and query display/clear function

4. Filtering your search will update the Explore Dashboard’s visualizations and the results tables (Figure 5). Each results table will be updated with information on the participants, samples, diagnoses, studies, or files that meet the filtered criteria. Information on each table is described below:

    a. “**Participants**”: Characteristics of a participant in the Explore Dashboard. Participants belong to a study, and they may have one or more samples, diagnoses, or files associated with them.
    b. “**Diagnosis**”: Diagnosis terms, codes, and other classification annotations are listed by participant. Diagnoses are associated with participants and samples, and different samples may have different diagnoses.
    c. “**Studies**”: Studies that are a part of the Explore Dashboard. Participants, diagnosis, samples, and files all belong to a CCDI study.
    d. “**Samples**”: Samples available from participants within the Explore Dashboard. Samples belong to a participant and can be associated with one or more files.
    e. “**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. Files may also be of many types, including sequencing, proteomics, imaging files, etc. DICOM imaging files are currently available for the Genomic Sequencing of Pediatric Rhabdomyosarcoma (phs000720) and Molecular Characterization Initiative (phs002790) studies and can be accessed directly from the Imaging Data Commons (IDC) Data Portal and

file paths to images are provided in the downloadable study manifest within Hub, described in the following section.

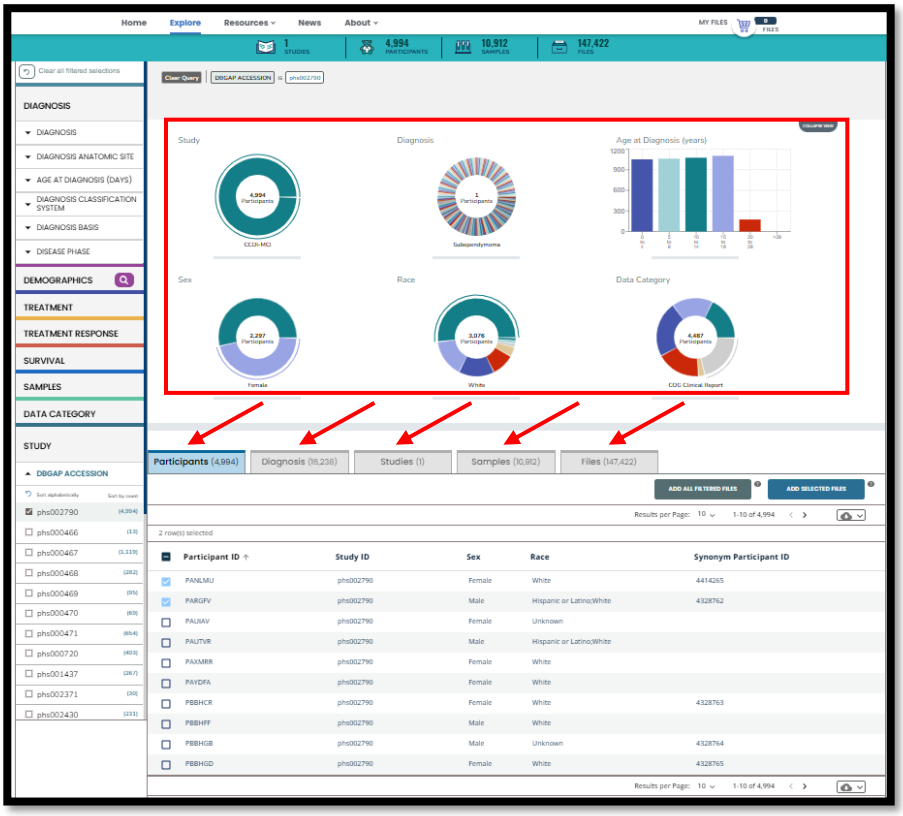


Figure 5: Explore Dashboard visualizations and results tables with arrows pointing to the available informational tables

Downloading Metadata from the Studies tab

From the CCDI Hub Explore Dashboard, you can download all open metadata for each study from the "Studies" tab to further filter data and build cohorts. For instance, additional filtering by diagnosis of interest can generate a set of participants and the resulting manifest can be uploaded into the CGC. As an example, the following steps guide you on how to download the metadata for the CCDI Molecular Characterization Initiative:

1. Using the process described above, open the "STUDY" set of filters from the lefthand menu, expand the "STUDY NAME" category, and scroll down to find "Molecular Characterization Initiative."
2. Select the checkbox corresponding to "Molecular Characterization Initiative" and see the Dashboard reload, filtered for this study's details.
3. Navigate to "Studies" in the results tables and locate the "Manifest" column.
4. Click the "Download study manifest" icon in the "Manifest" column to download the metadata for this study (Figure 6).
5. Open the resulting file on your local machine to browse the resulting metadata tables (Figure 7).

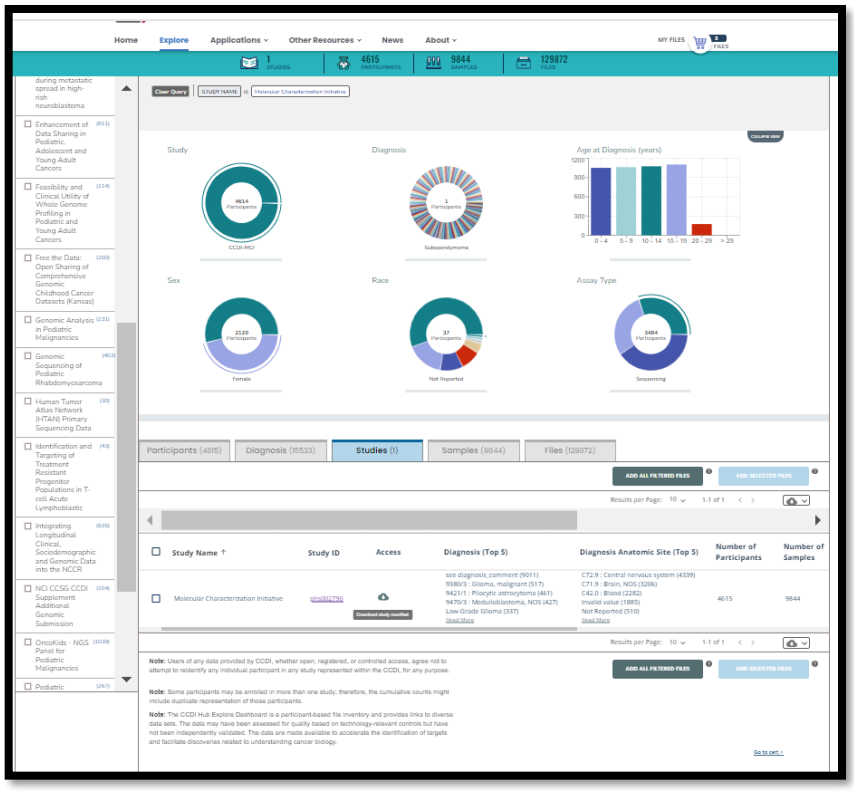
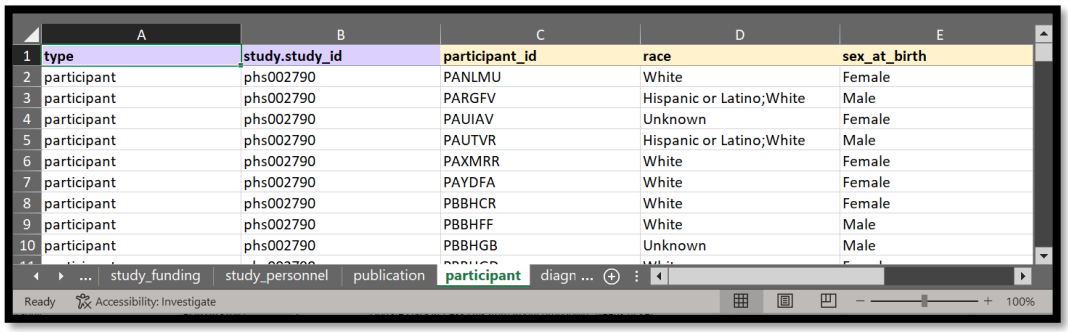


Figure 6: Download metadata manifest for a given study



	A	B	C	D	E
1	type	study.study_id	participant_id	race	sex_at_birth
2	participant	phs002790	PANLMU	White	Female
3	participant	phs002790	PARGFV	Hispanic or Latino;White	Male
4	participant	phs002790	PAUIAV	Unknown	Female
5	participant	phs002790	PAUTVR	Hispanic or Latino;White	Male
6	participant	phs002790	PAXMRR	White	Female
7	participant	phs002790	PAYDFA	White	Female
8	participant	phs002790	PBBHCR	White	Female
9	participant	phs002790	PBBHFF	White	Male
10	participant	phs002790	PBBHGB	Unknown	Male

Figure 7: Study metadata export file browsable on local machine

Appendix A details the process for generating a DRS manifest from the downloaded study metadata tables to be compatible with the CGC.

Creating an Exportable File Manifest from the Cart

In addition to the study-specific downloads, you can also export each row-level metadata element for CCDI participants, diagnoses, samples, or files based on your selections within the Explore Dashboard. 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.
2. After selecting desired rows, add files for that element to the My Files shopping cart (Figure 8) by clicking either the “ADD ALL FILTERED FILES” or “ADD SELECTED FILES” button. Note that selection of items in each tab depends on the specific content of that tab. For example, selecting an item in the “Participants” tab means every file associated with a participant will be added to the My Files shopping cart, whereas selecting an item in the “Files” tab will add that single selected file to the cart.

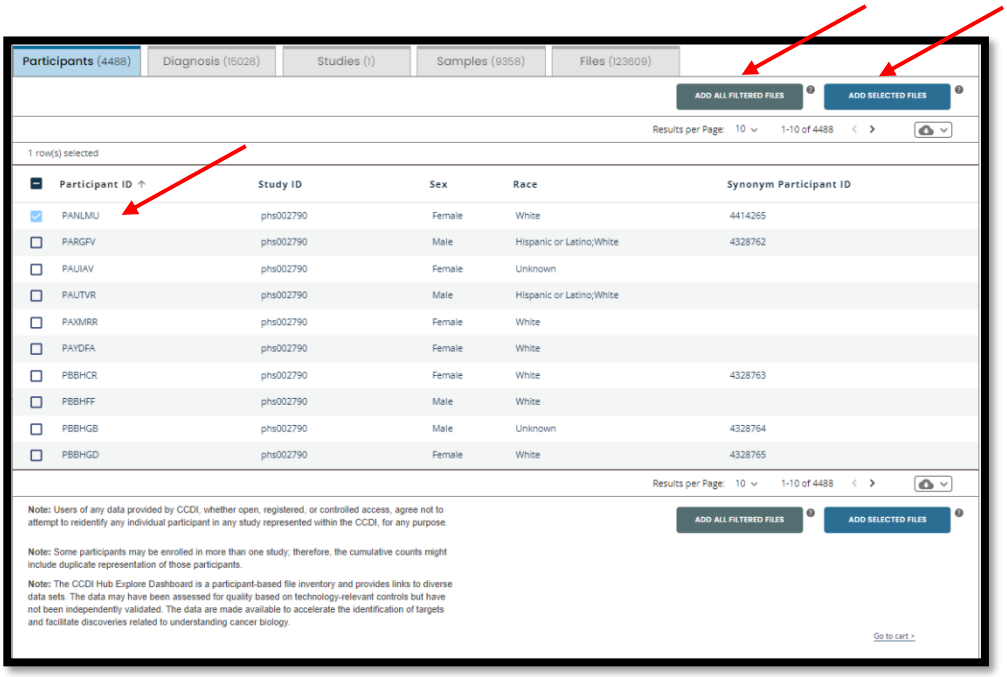


Figure 8: Selection checkboxes and buttons to add files to the cart for the “Participants” table

3. To navigate to the shopping cart, select “MY FILES” or the shopping cart icon on the menu bar (Figure 9).

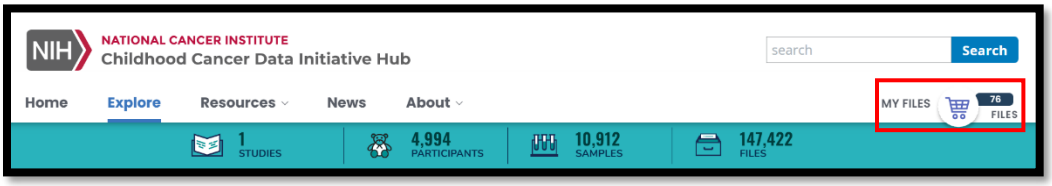


Figure 9: CCDI Hub menu bar with red box highlighting the My Files shopping cart

4. 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 from the “AVAILABLE EXPORT OPTIONS” dropdown menu (Figure 10) will produce a comma-separated values (CSV) file manifest of the items within the cart.

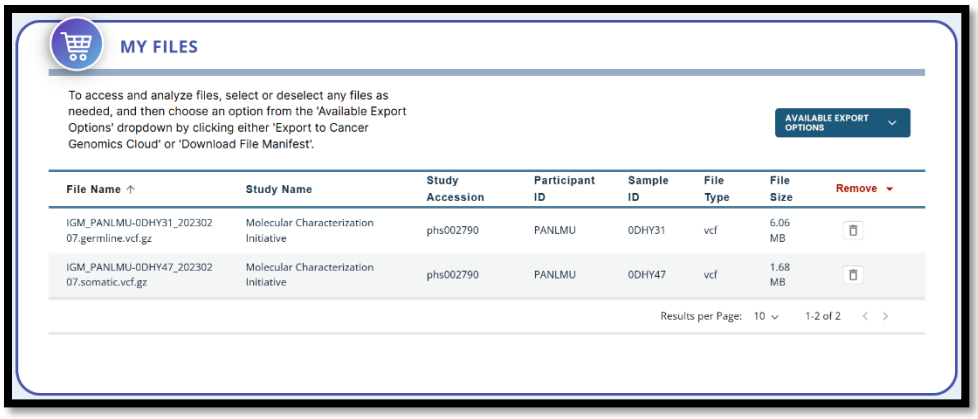


Figure 10: The Explore Dashboard Cart page with red box highlighting the "Available Export Options" button

5. You can then download this manifest file locally or upload it in the CGC (Appendix C). Similarly, you can instead select the "EXPORT TO CANCER GENOMICS CLOUD" button from the "AVAILABLE EXPORT OPTIONS" dropdown menu to load the resulting manifest directly into your CGC account.

Note that the Cart has a maximum capacity of 6,000 files, which limits the ability to create a single comprehensive manifest for use in the CGC. To work around this limitation, either create manifests from the cart in batches (containing up to 6,000 files) or use the comprehensive metadata downloads from the Explore page "Studies" tab to create a manifest that can take all data for a given study into the CGC. Longer term solutions are being researched.

Contact Information

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

THIS PAGE IS INTENTIONALLY LEFT BLANK

Appendix A: 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.

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

NCI 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 data is available for download using the 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 dbGaP.

Below are instructions for using the Data Commons Framework (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. Please navigate to nci-crdc.datacommons.io and click the "RAS Login" button. Login to the NCI DCF Services portal at nci-crdc.datacommons.io/login (Figure B1).

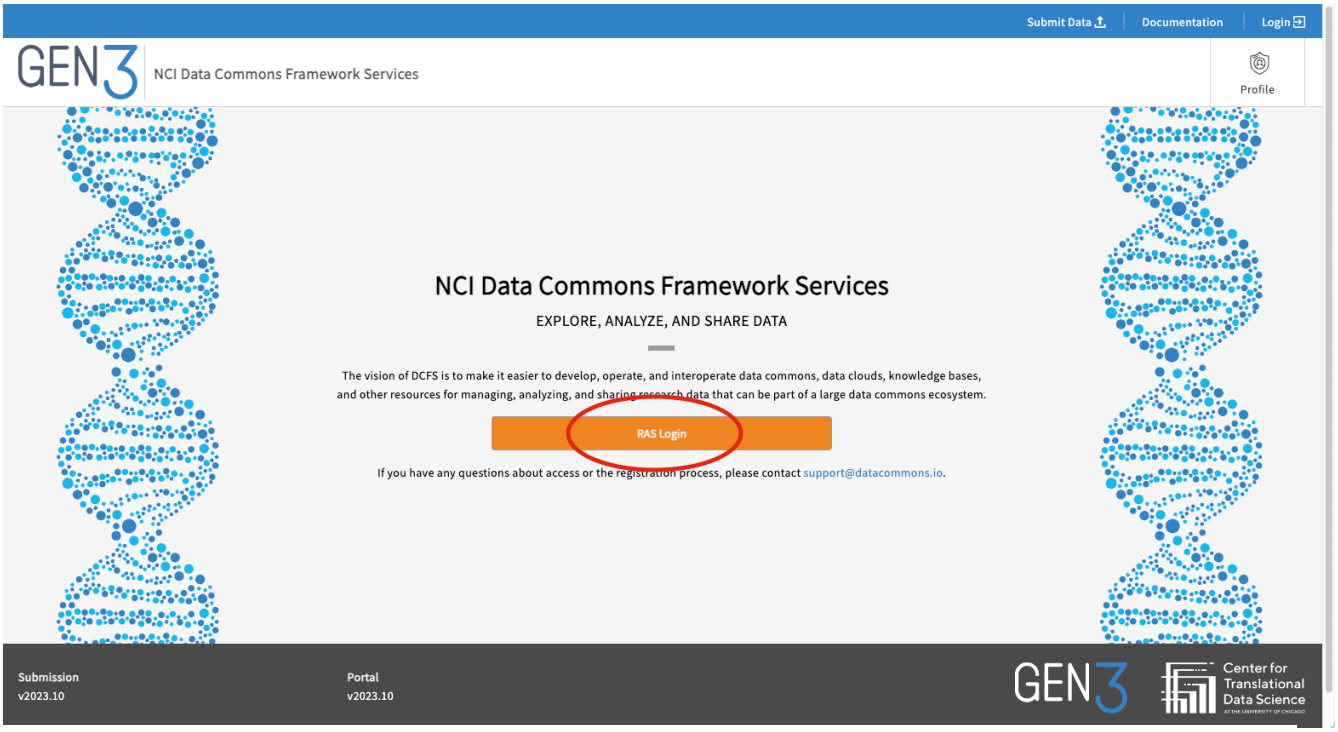


Figure B1: NIH DCF homepage with the NIH Researcher Auth Service (RAS) login highlighted

2. Once logged in, click the "Profile" section in the top right corner and review your project access to confirm study access (Figure B2).

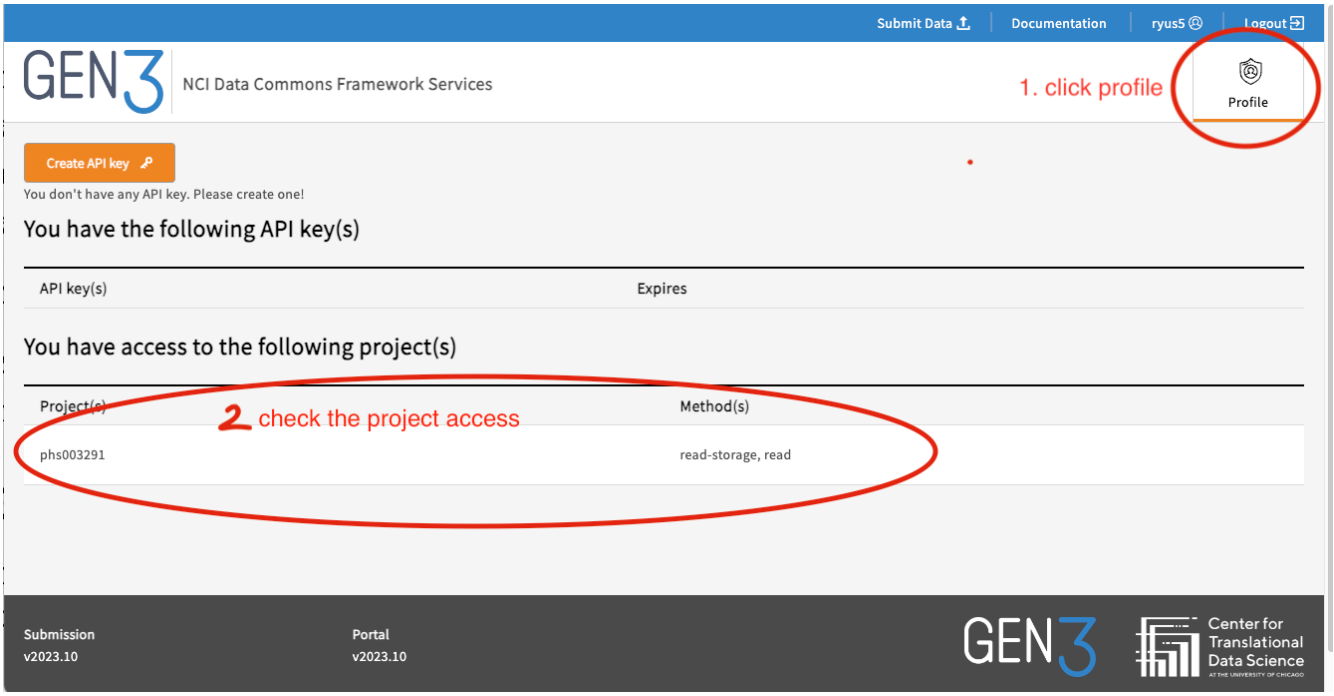


Figure B2: DCF Profile page highlighting "Profile" and the accessible projects.

3. Add the file GUID after the final backslash in this URL: https://nci-crdc.datacommons.io/user/data/download/. Paste the URL you created in a browser address field and press Enter or 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 into the browser address field and press Enter or Return (Figure B3).
    a. Note: Once issued, the signed URL provided is valid for a relatively short period of time.

6. The NCI DCF Services Portal will respond by displaying a URL. Click the URL to download the file (Figure B3).

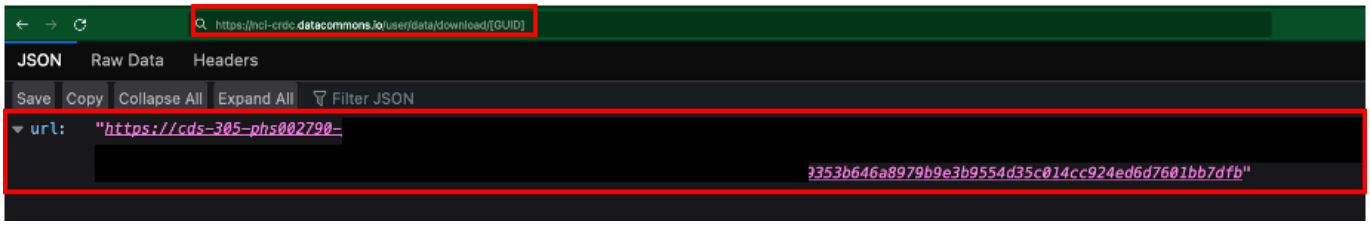


Figure B3: DCF Service Portal displaying the signed access URL and the file download URL

Note: If errors or problems are experienced during the file downloading process above, please contact the CCDI mailbox for assistance.

File Download Procedure via Command Line 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 the 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. Create a Gen3 structured manifest:

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

5. Download the file(s) using the Gen3 client (either the single or multiple download option).

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

Appendix C: The 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.

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, cancergenomicscloud.org, click “REGISTER” or “LAUNCH” in the center of the page (Figure C1).

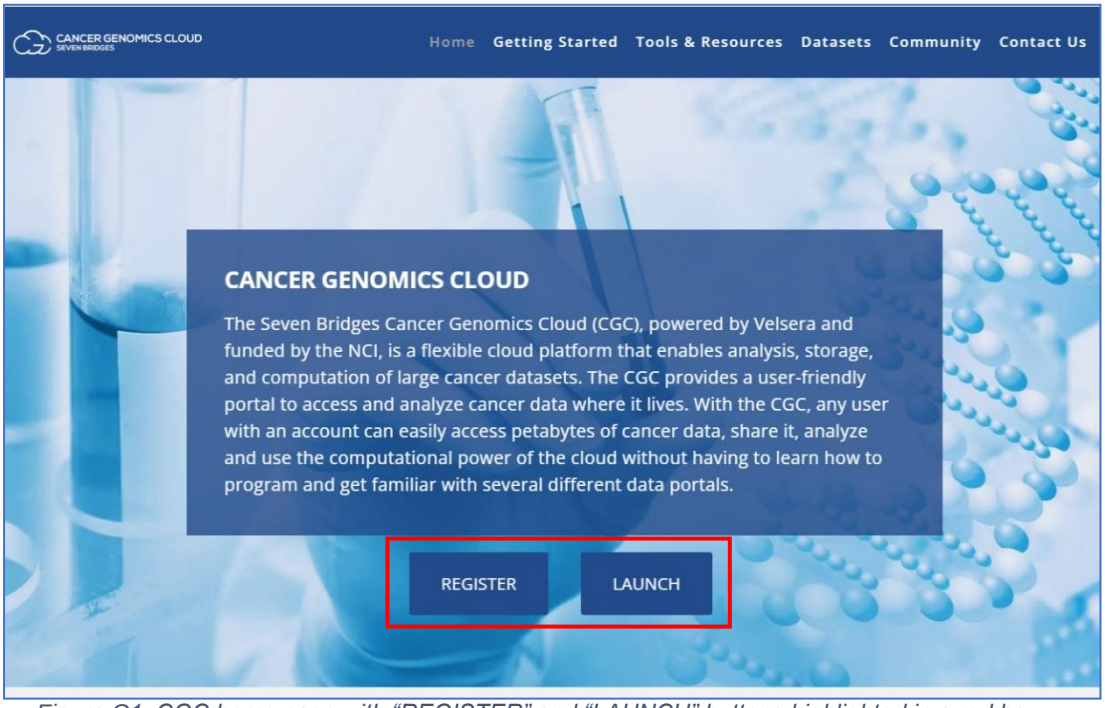


Figure C1: CGC home page with “REGISTER” and “LAUNCH” buttons highlighted in a red box.

2. On the login screen, click on “Log in with eRA Commons” (Figure C2).

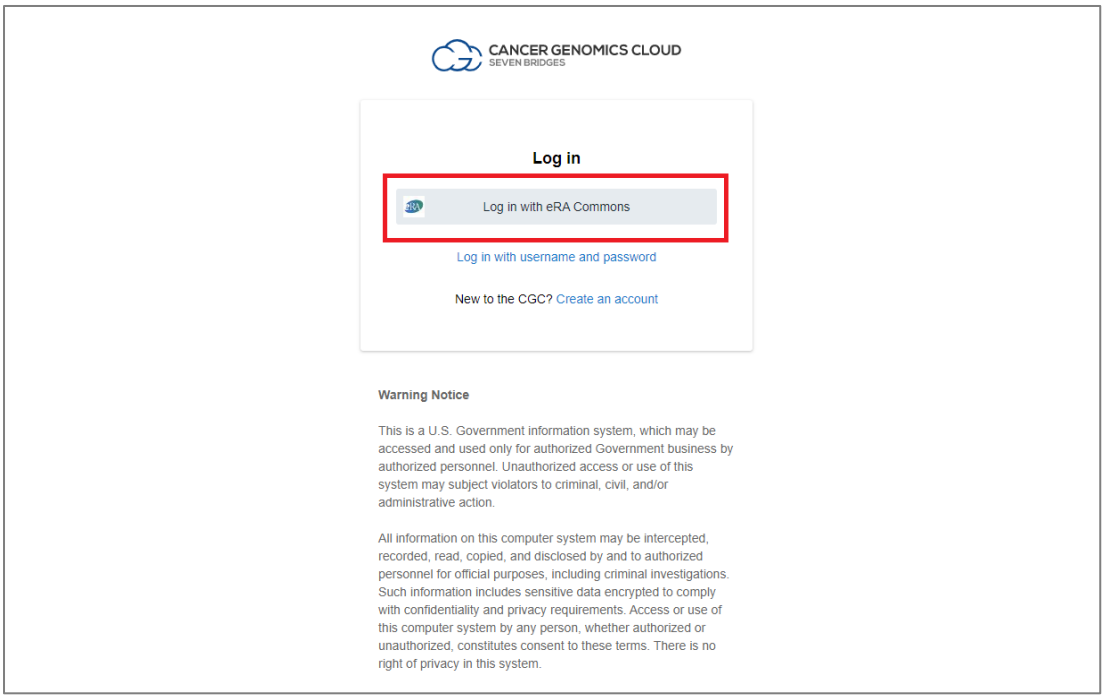


Figure C2: eRA Commons login screen for the CGC, with login button highlighted in a red box

3. On the login screen, enter the eRA Commons account credentials associated with your approved dbGaP study and click “Sign In” (Figure C3).

   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.

   For additional details about ensuring your Velsera eRA Commons account is correct and correctly configuring the download restriction and network access toggle at the project level, please refer to the following documentation:
   - docs.cancergenomicscloud.org/docs/modify-project-settings
   - docs.cancergenomicscloud.org/docs/define-download-restriction

   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.

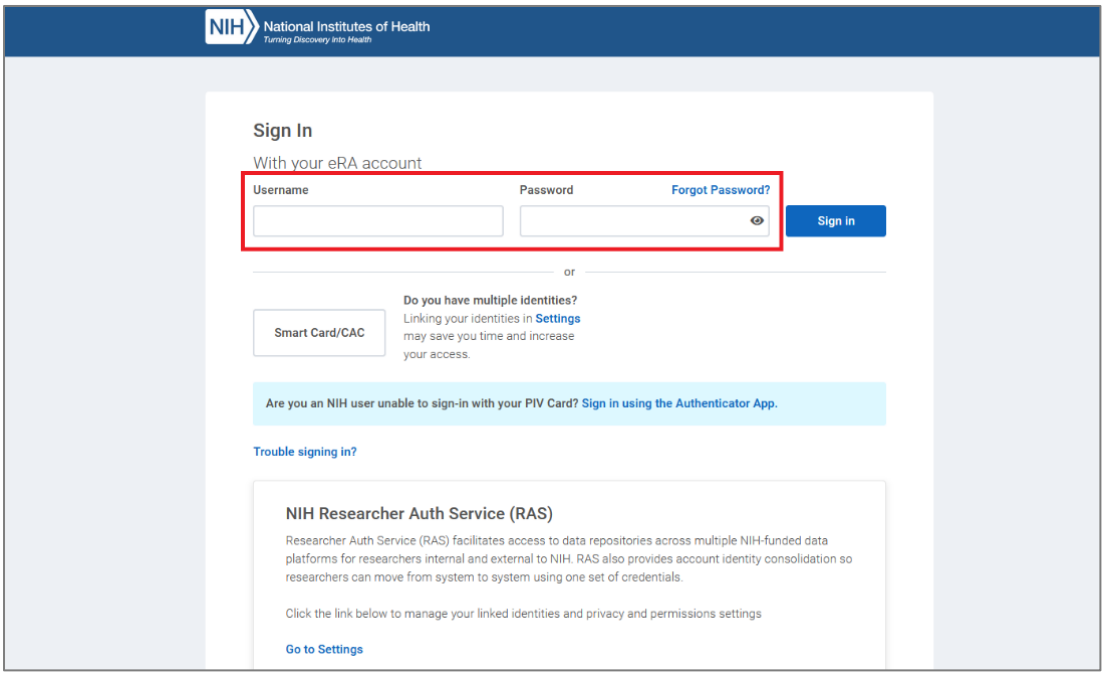


Figure C3: Login page for CGC, with username and password credentials sections highlighted in a red box.

4. If you agree to the “Consent to Share Information” on the following page, click the “Grant” button to continue (Figure C4).

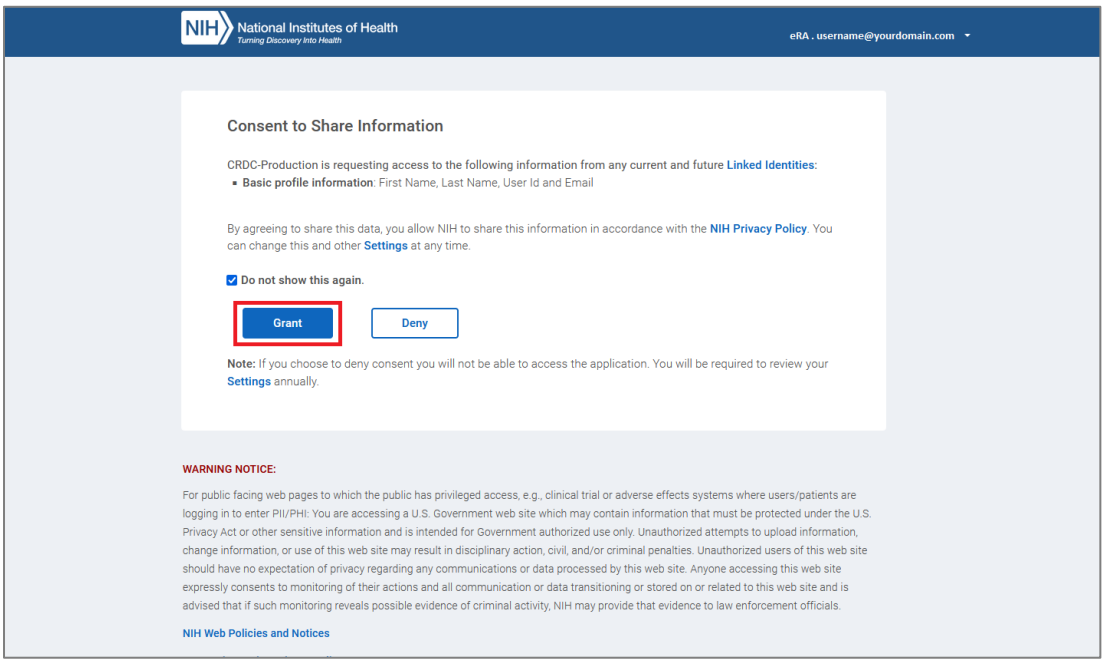


Figure C4: Consent to share information page with red box highlighting the “Grant” button to confirm consent to share information with the CGC

5. If you agree to authorize Gen3 DCFS 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 C5). Note

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

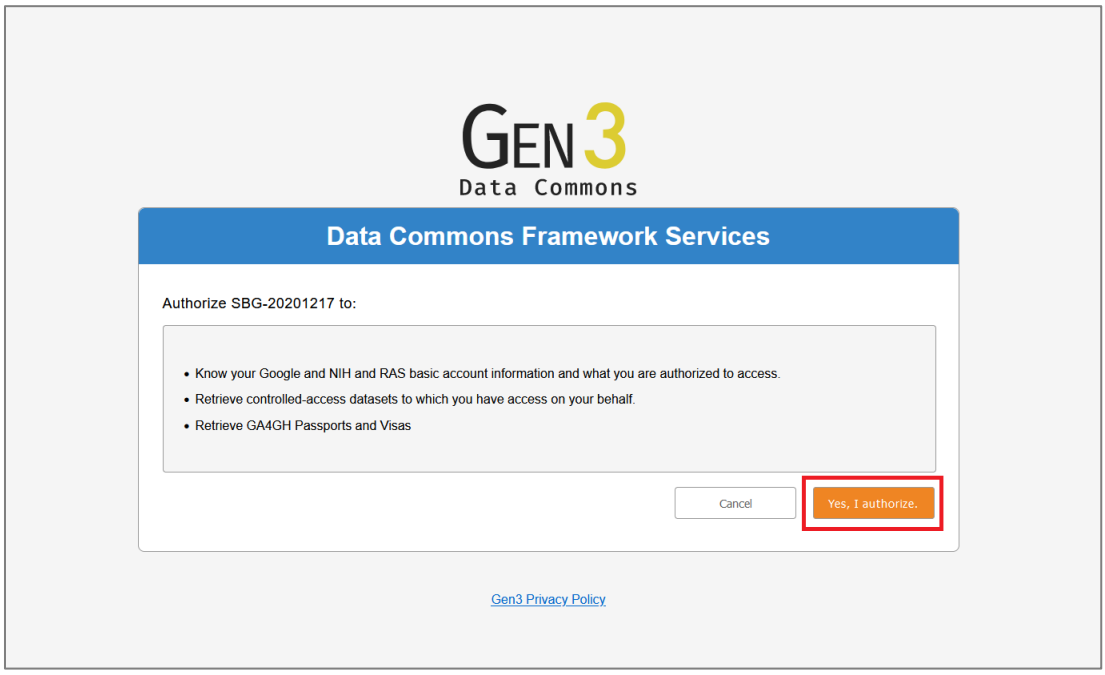


Figure C5: Gen3 DCFS authorization page with "Yes, I authorize" button highlighted in a red box

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 C6).

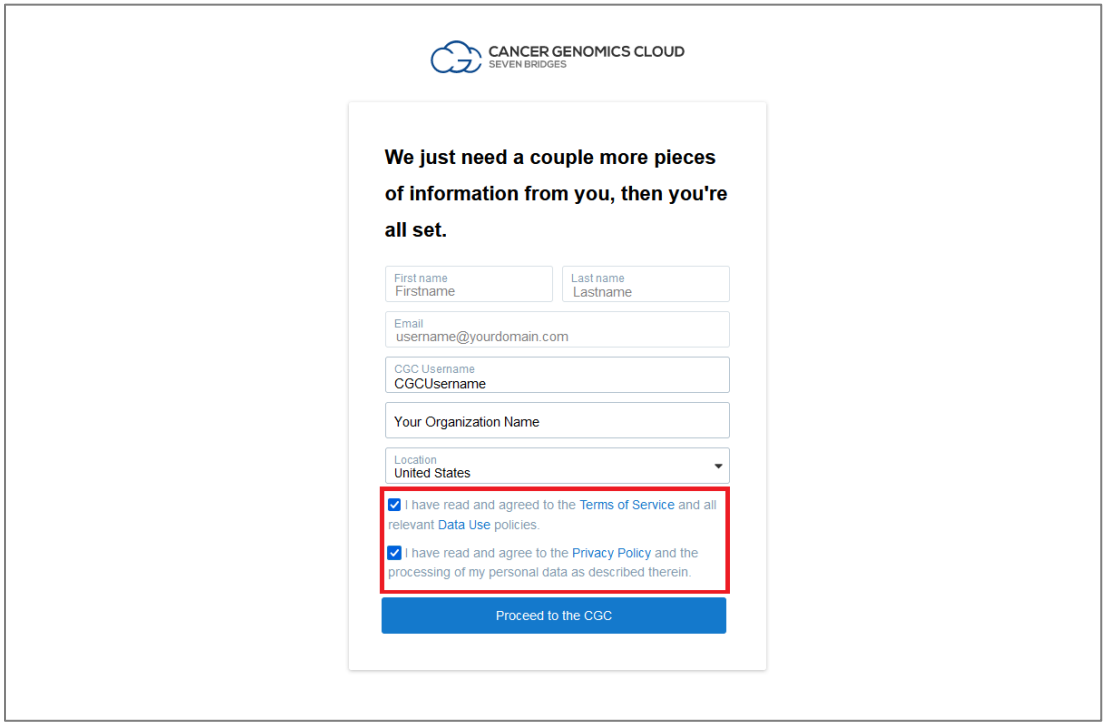


Figure C6: Confirmation of terms and policy for CGC registration highlighted in a red box

7. If the CGC questionnaire appears, complete it to continue (Figure C7).

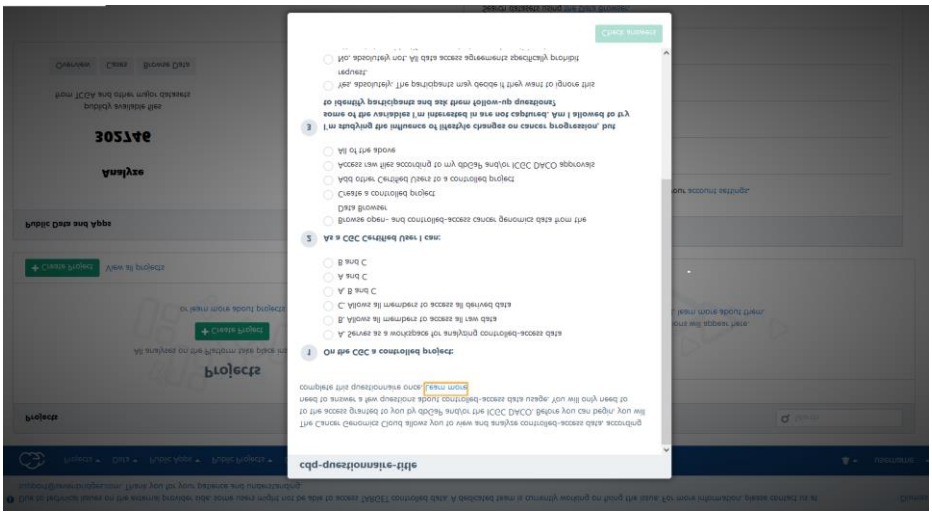


Figure C7: Screenshot of the CGC questionnaire

Creating a DRS Manifest from downloaded Study metadata

The Study Manifest file contains all metadata about a study available in the CCDI Hub. This file can be used to create a Data Repository Service (DRS) file that includes identifiers and metadata that allows users to locate and access specific files or data sets within a data repository for use on the CGC. To convert the Study Manifest file into a DRS manifest, please follow these steps:

1. Go to ccdi.cancer.gov/explore
2. Select "Molecular Characterization Initiative" under Study Name, or choose “phs002790” under Study dbGaP Accession.
3. Download the Study Manifest (contains metadata for the study).
4. Find the tabs ending with "_file" (look for FASTQ, JSON, etc.).
5. Create a DRS Manifest for Cloud Analysis
    - Extract and rename the column headers: file_name (name); dcf_indexd_guid (drs_uri); study_id (study_code); sample.sample_id/participant_id (case_id).
6. Add drs://nci-crdc.datacommons.io/ to all the beginning of all rows in the drs_uri column.Go to https://www.cancergenomicscloud.org/:
7. In CGC, go to your Project. Under Files, find Add Files and GA4GH Data Repository Service (DRS).
8. Upload your manifest.
9. On the next page, select the “From a manifest file” tab and drag and drop your manifest onto the page. Go through the next steps agreeing to the use of the system and ingest the manifest.
10. After being redirected back to the file page for the project, it will begin to load the files into the Files section of the CGC project.

Additional Resources on Working with Data at the CGC

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