CHILDHOOD CANCER DATA INITIATIVE (CCDI)

Data Access Instructions

2/02/2024

Content

| Background | 2 |
|---|----|
| Introduction and Overview | 2 |
| database of Genotypes and Phenotypes (dbGaP) | 3 |
| NCI Data Commons Framework Services (DCFS) | 3 |
| CCDI Hub Explore Dashboard | 3 |
| Finding Participants, Samples, and Files | 3 |
| Creating an Exportable File Manifest | 7 |
| Cancer Genomics Cloud (CGC) | 9 |
| Creating a Project in CGC & Importing Explore Dashboard Manifest Files | 9 |
| Using CGC Data Studio | 13 |
| Additional Resources on Working with Data at the CGC | 15 |
| Contact Information | 15 |
| Appendix A: Creating a CGC Account | 17 |
| Appendix B: Using the CGC Cancer Data Service Explorer to Identify Files | 22 |
| Appendix C: Data Commons Framework Services (DCFS): Controlled Data Access Instructions | 25 |
| File Download Procedure via User Interface | 25 |
| File Download Procedure via Call Level Interface (CLI) client | 27 |

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 <u>CCDI mailbox</u> with any questions.

Introduction and Overview

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

CCDI studies are summarized and indexed in the <u>CCDI Hub Explore Dashboard</u>, 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 <u>Cancer Genomics Cloud</u> (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 <u>Cancer Research Data Commons</u> (CRDC), a cloud-based infrastructure where the data are hosted. For controlled-access studies, CRDC collaborates with <u>Data Commons Framework</u> <u>Services</u> (DCFS) to provide authentication and authorization services for the <u>database of Genotypes and</u> <u>Phenotypes (dbGaP)</u>. To gain access to controlled data, researchers must first have an <u>NIH eRA</u> <u>Commons account</u> for authentication, after which they will need to obtain authorization (via an active DCFS login access the data in the NIH <u>dbGaP</u>.

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 <u>instructional video</u> on applying for controlled access data and consult the <u>Tips on Preparing a Successful Data Access Request</u>. A step-by-step breakdown of the data access request process is located in this <u>guide</u>.

NCI Data Commons Framework Services (DCFS)

Data Commons Framework Services (DCFS), powered by <u>Gen3</u>, 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 <u>CCDI Hub Explore Dashboard</u> 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 <u>controlled-access login page</u>.

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).

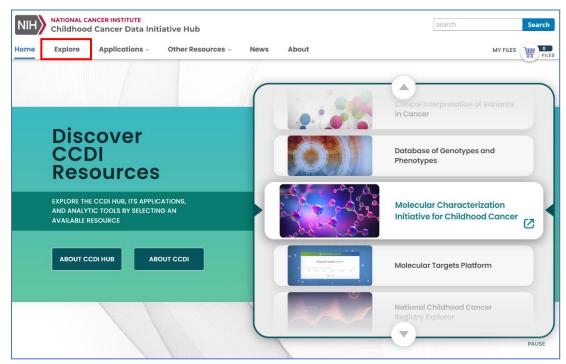


Figure 1: CCDI homepage with red box highlighting the Explore Dashboard menu bar link

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.

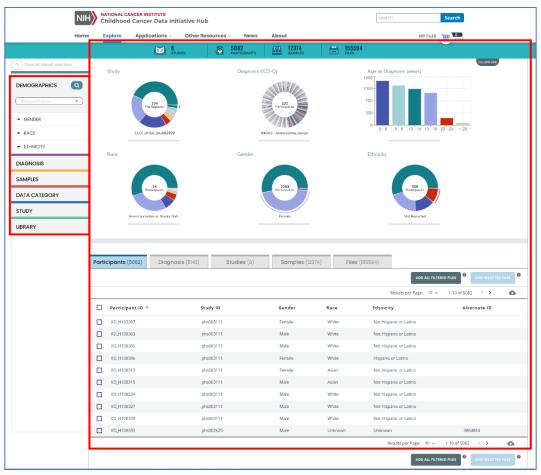


Figure 2: Explore Dashboard page with red boxes highlighting the search filters and results

| Clear all filtered selections | | Clear Query | DBGAP PHS ACCESSION IS phs00279 | • | _ | | | |
|--|------------|------------------|---------------------------------|---------------|------------------------|--------------|--------------------------|------------------|
| DEMOGRAPHICS | ٩ | | | | | | | |
| Participant ID Search | • | Study | | Diagnos | is (ICD-O) | | Age at Diagnosis (years) | COLLAPSE VIEW |
| ▼ GENDER | | | | | | | 150- | |
| ▼ RACE | | | 2297 | | | | 300- | |
| ETHNICITY | | | Participants | | Participants | | 150 | |
| DIAGNOSIS | | | | | A WWW | | 0-4 5-9 10-14 1 | 5 - 19 20 - 29 |
| SAMPLES | | | CTSMC | | 9370/3 : Chordoma, NOS | | | |
| | _ | Race | | Gender | | | Ethnicity | |
| DATA CATEGORY | _ | | | | | | | |
| STUDY | | | 1387 | | 1059 | | 110 | |
| DBGAP PHS ACCESSION | N | | Participants | | Participants | | Participants | 1 |
| Sort alphabetically Sort | t by count | | | | | | | |
| d phs002790 | (2297) | ┥ | White | | Female | | Unknown | |
| □ phs002504 | (188) | | | | | | | |
| phs002517 | (1011) | | | | | | | |
| phs002518 | (1039) | Participants (2 | 2297) Diagnosis (229 |) Studies (1) | Samples (524 | 9) Files (46 | 2400) | |
| phs002529 | (200) | Fullicipulits (2 | Diagnosis (228 | studies (i) | Sumples (524) | o) Filos (+(| 3422) | |
| phs002599 | (104) | | | | | | ADD ALL FILT | ERED FILES |
| phs002620 | (114) | | | | | | Results per Page: 10 🗸 | 1-10 of 2297 < > |
| □ phs003111 | (129) | Participa | nnt ID ↑ | Study ID | Gender | Race | Ethnicity | Alternate ID |
| ▼ FUNDING GRANT ID | | PANLMU | | phs002790 | Female | White | Not Hispanic or Latino | |
| | | PARGEV | | phs002790 | Male | White | Hispanic or Latino | |
| ▼ ACRONYM | | PBBHCR | | phs002790 | Female | White | Not Hispanic or Latino | |
| STUDY SHORT TITLE | | PBBHFF | | phs002790 | Male | White | Not Hispanic or Latino | |
| LIBRARY | | PBBHGB | | phs002790 | Male | Unknown | Not Hispanic or Latino | |
| and the second s | | PBBHGD | | phs002790 | Female | White | Not Hispanic or Latino | |
| | _ | PBBHGT | | phs002790 | Female | White | Not Hispanic or Latino | |

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.

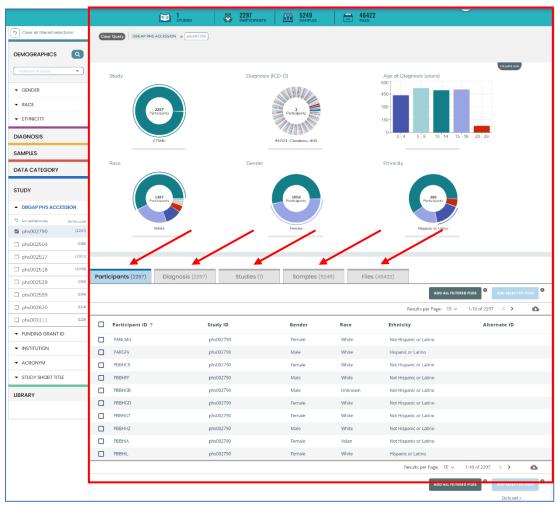


Figure 4: Explore Dashboard visualizations and results tables with arrows pointing to the available informational table

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:

- 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.

| articipants (2297) | Diagnosis (2297) | Studies (1) | Samples (5 | 249) Files | (46422) | |
|--------------------|------------------|-------------|------------|------------|------------------------|------------------------|
| | | | | | | |
| | | | | | | |
| row(s) selected | | | | | Results per Page: | 10 🗸 1-10 of 2297 🔇 🛆 |
| | | | | | | |
| Participant ID 🕆 | Study | ID | Gender | Race | Ethnicity | Alternate ID |
| PANLMU PANLMU | phs002 | 790 | Female | White | Not Hispanic or Latino | |
| PARGPV | phs002 | 790 | Male | White | Hispanic or Latino | |
| Ревнся | phs002 | 790 | Female | White | Not Hispanic or Latino | |
| PEOHIT | phs002 | 790 | Male | White | Not Hispanic or Latino | |
| пернов | phs002 | 790 | Male | Unknown | Not Hispanic or Latino | |
| PEBHGD | phs002 | 790 | Female | White | Not Hispanic or Latino | |
| PREHIGT | phs002 | 790 | Female | White | Not Hispanic or Latino | |
| P88HHZ | phs002 | 790 | Male | White | Not Hispanic or Latino | |
| PECHA | phs002 | 790 | Female | Asian | Not Hispanic or Latino | |
| P00HIL | phs002 | 790 | Female | White | Hispanic or Latino | |
| | | | | | Results per Page: 10 |) ↓ 1-10 of 2297 () 🖒 |
| | | | | | _ | |

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).

| NH NATIONAL CANCER INSTITUTE Childhood Cancer Data Initiative Hub | | | | | | | | |
|--|---------|------------------|------------------------|-----------------|----------------|--|----------|--|
| Home | Explore | Applications ~ 0 | Other Resources - News | About | | | MY FILES | |
| | | | PARTICIPANTS | 5249 SAMPLES | 46422 FILES | | | |

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 <u>Creating a Project in CGC & Importing Explore</u> <u>Dashboard Manifest Files</u>) or use it locally.

| Thank you for your interest in CCDI supported data. Selecting the "Download Manifest" button will produce a manifest of assay files for items within the cart. This manifest file can be ploaded in the <u>Concer connents</u> CLOU dD to access and analyze controlled access information. Additional help and nformation about the CGC use and access is available at the <u>CGC Knowledge Center</u> (2. | | | | | | | |
|--|--|-----------------|----------------|-----------|------------------|-------------------|-------------|
| File Name 🛧 | Study Short Title | Study Accession | Participant ID | Sample ID | File Type | File Size | Remove 👻 |
| 206917410085_R07C01_Grn.idat | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY47 | idat | 13.04 MB | Ō |
| 206917410085_R07C01_Red.idat | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY47 | idat | 13.04 MB | Ō |
| IGM_PANLMU-0DHY31_20230207.germlin e.vcf.gz | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY31 | vcf | 6.06 MB | Ō |
| IGM_PANLMU-0DHY31_20230207_Redact ed.pdf | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY47 | pdf | 128.39 KB | |
| IGM_PANLMU-0DHY31_20230207_cnv_ger mline.json | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY31 | json | 44.81 MB | Ō |
| IGM_PANLMU-0DHY31_20230207_normal .aln.cram | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY31 | cram | 21.42 GB | 0 |
| IGM_PANLMU-0DHY31_20230207_normal .aln.cram.crai | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY31 | crai | 241.39 KB | ۵ |
| IGM_PANLMU-0DHY3S_20230201.archer_ version.txt | Childhood Cancer Data initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY3S | txt | 46 Bytes | |
| IGM_PANLMU-0DHY3S_20230201.full_res ults.txt | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY3S | txt | 2 MB | ٥ |
| IGM_PANLMU-0DHY3S_20230201_Redact ed.pdf | Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative | phs002790 | PANLMU | 0DHY35 | pdf | 123.36 KB | Ō |
| | | | | | Results per Page | : 10 ∨ 1-1 | 0 of 26 < > |

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 <u>sign up</u> for the CGC. Please follow the procedure to "Register via an external account" to access dbGaP-regulated controlled data. Velsera also hosts <u>office hours</u> 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).

| Projects 🔶 Data 🖣 | Public Apps - | Publi |
|-------------------|--------------------------|-------|
| Search projects | | Q |
| phs003111 | | |
| phs002529 | | |
| phs002504 | | |
| phs002620 | | |
| phs002599 | | |
| phs002790 | | |
| | | |
| | | |
| | | |
| | | |
| View all projects | + Create a proje | _ |

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.

| Create a project × |
|---|
| Name |
| |
| https://cgc.sbgenomics.com/u/ |
| CONTROLLED This project will contain controlled data. 6 |
| General information Advanced settings |
| Billing group |
| Pilot Funds |
| Location 0 |
| AWS (us-east-1) - |
| Execution settings |
| Spot instances Spot instances can significantly reduce the cost of your task execution if results are not needed urgently. Learn more |
| Reuse Automatic reuse of precomputed results can significantly reduce the time and cost of your task execution. Learn more |
| Cancel Create |

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 11: Empty files page of a new project

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

| New for | older | + Add files - | | | |
|------------|---------------|--------------------------------|------|--|--|
| n more abc | | Explorer and Browser | file | | |
| | Public | c Files | | | |
| | Proje | cts | | | |
| | Your Computer | | | | |
| | FTP / HTTP | | | | |
| | | GH Data sitory Service) | | | |
| | Data | Tools | | | |
| | Volun | nes | | | |
| | | rt from a fest file | | | |

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).

| Import from an GA4GH Data Repository Service (DRS) |
|--|
| Paste DRS URIs From a manifest file |
| Import a csv or tsv file that contains DRS URIs which will be applied to your files. |
| Start with our template or learn more. |
| Drag & drop files from your computer or |
| Browse manifest |
| |

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."

| Import from an GA4GH Data Repository Service (DRS) | |
|---|--------|
| Paste DRS URIs From a manifest file | |
| Import a csv or tsv file that contains DRS URIs which will be applied to your files. | |
| Start with our template or learn more. | |
| CCDI Inventory File Manifest 2023-12-21 14-43-24.csv | × |
| Add tags to files | |
| Resolve naming conflicts: | |
| Skip 👻 | |
| I understand that data accessible via DRS, including but not limited to controlled-access data, may be subject to terms and conditions of acceptable use, and I confirm that I am only importing data in accordance with any applicable terms of use, including but not limited to my obligations under any applicable Data Use Agreements. | |
| | 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).

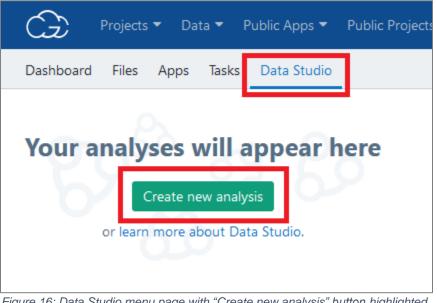


Figure 16: Data Studio menu page with "Create new analysis" button highlighted in red

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).

| Analysis name | | | | |
|--|------------|--------|-----|--|
| Name | | | | |
| Environment | | | | |
| 😇 JupyterLab | | 🗬 RStu | dio | |
| Environment setup 🕑 | | | | |
| SB Bioinformatics - R 4.2 - BioC | 3.15 DEFAU | u • | | |
| Instance type c5.2xlarge (8vCPUs, 16GB RAM) | - | | | |
| Price: \$0.34 per hour | | | | |
| Attached Storage (GB) 🕑 | | | | |
| 1024 🗘 | | | | |
| Suspend Time 😧 | | | | |
| On 🌔 | | | | |
| | | | | |

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).

| File Edit Code View Plots Session Build Debug Profile Tools Help | |
|--|-------------------------------|
| R • • • • • • • • • • • • • • • • • • • | |
| | |
| Source | Environment History Connec |
| Console Terminal × Background Jobs × | 🐨 🔒 📑 Import Dataset 👻 🖄 |
| | R 👻 🐴 Global Environment 👻 |
| AL P.4.6.1 | |
| R version 4.2.1 (2022-06-23) "Funny-Looking Kid" | |
| Copyright (C) 2022 The R Foundation for Statistical Computing | |
| Platform: x86_64-pc-linux-gnu (64-bit) | |
| | |
| R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. | |
| Type 'license()' or 'licence()' for distribution details. | |
| the rest of the rest of the rest for the rest of the r | |
| Natural language support but running in an English locale | |
| | |
| R is a collaborative project with many contributors. Type 'contributors()' for more information and | Files Plots Packages Help |
| 'station() on how to cite R or R packages in publications. | New Folder New Blank File |
| creation of the second of the | Home |
| Type 'demo()' for some demos, 'help()' for on-line help, or | A Name |
| 'help.start()' for an HTML browser interface to help. | RData |
| Type 'qO' to quit R. | Rhistory |
| [Workspace loaded from ~/.RData] | COG_JSON_Table.tsv |
| [nonspace could from -y notica] | COG_SON_TABLE.ISV |
| > | |
| | |
| | |

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 <u>documentation</u>.
- 7. (Optional) Modify the <u>suspend time settings</u> 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: <u>https://docs.cancergenomicscloud.org/docs</u>
- Importing CDS Data: <u>https://docs.cancergenomicscloud.org/docs/import-cds-data</u>
- Common Workflow Language Workflows and Apps: <u>https://cgc.sbgenomics.com/public/apps</u>
- Volumes: <u>https://docs.cancergenomicscloud.org/docs/volumes-1</u>
- Tool Editor Tutorial: <u>https://docs.cancergenomicscloud.org/docs/tool-editor-tutorial</u>
- About the Editor: <u>https://docs.cancergenomicscloud.org/docs/about-the-editor</u>

Contact Information

Please direct any questions or requests for further information to the <u>CCDI mailbox</u>.

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 <u>list of all CCDI</u> <u>studies</u> released is also available. The following instructions describe the process to create a CGC account.

1. From the CGC home page at <u>cancergenomicscloud.org</u>, click "REGISTER" or "LAUNCH" in the center of the page (Figure A1).

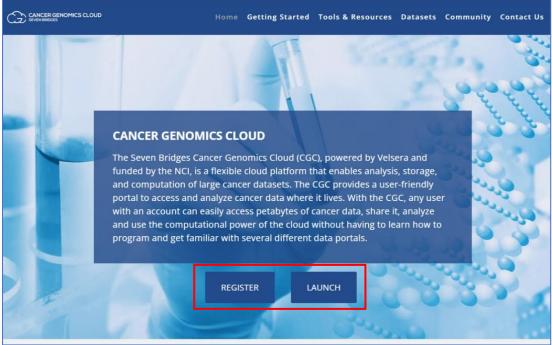


Figure A1: CGC home page with "REGISTER" and "LAUNCH" buttons highlighted in a red box.

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

| CANCER GENOMICS CLOUD |
|--|
| Log in Log in with eRA Commons Log in with username and password New to the CGC? Create an account |
| Warning Notice This is a U.S. Government information system, which may be accessed and used only for authorized Government business by authorized personnel. Unauthorized access or use of this system may subject violators to criminal, civil, and/or administrative action. All information on this computer system may be intercepted, recorded, read, copied, and disclosed by and to authorized personnel for official purposes, including criminal investigations. Such information includes sensitive data encrypted to comply with confidentiality and privacy requirements. Access or use of this computer system by any person, whether authorized or unauthorized, constitutes consent to these terms. There is no right of privacy in this system. |

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 <u>eRA Commons site</u>. 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.

| NIH National Institutes | of Health | | |
|---|---|---|---------|
| | | | |
| Sign In | | | |
| With your eRA ac | ccount | | |
| Username | Password | Forgot Password? | |
| | | • | Sign in |
| | or Do you have multiple identities? Linking your identities in Settings may save you time and increase your access. r unable to sign-in with your PIV Card? Sign in u | using the Authenticator App. | |
| Trouble signing in? | | | |
| Researcher Auth platforms for res researchers can r | cher Auth Service (RAS) I Service (RAS) facilitates access to data reposite searchers internal and external to NIH. RAS also i move from system to system using one set of c low to manage your linked identities and privacy | provides account identity consolida credentials. | |

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).

| NIH National Institutes of Health | eRA . username@yourdomain.com 👻 |
|--|--|
| | |
| Consent to Share Information | |
| CRDC-Production is requesting access to the following information from any current and f • Basic profile information: First Name, Last Name, User Id and Email | uture Linked Identities: |
| By agreeing to share this data, you allow NIH to share this information in accordance with th can change this and other Settings at any time. | ne NIH Privacy Policy. You |
| ✓ Do not show this again. | |
| Grant Deny | |
| Note: If you choose to deny consent you will not be able to access the application. You will I Settings annually. | be required to review your |
| WARNING NOTICE: | |
| For public facing web pages to which the public has privileged access, e.g., clinical trial or adverse effects s logging in to enter PII/PHL: You are accessing a U.S. Government web site which may contain information it Privacy Act or other sensitive information and is intended for Government authorized use only. Unauthorize change information, or use of this web site may result in disciplinary action, civil, and/or criminal penalities, should have no expectation of privacy regarding any communications or data processed by this web site. A expressly consents to monitoring of their actions and all communication or data transitioning or stored on advised that if such monitoring reveals possible evidence of criminal activity, NIH may provide that evidence | at must be protected under the U.S. d attempts to upload information, Unauthorized users of this web site nyone accessing this web site or related to this web site and is |
| NIH Web Policies and Notices | |

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.

| | GEN 3 Data Commons |
|----------------------------|--|
| Data | Commons Framework Services |
| Authorize SBG-20201217 to: | |
| | AS basic account information and what you are authorized to access. ts to which you have access on your behalf. isas |
| | Cancel Yes, Lauthorize. |

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).

| ouple more pieces om you, then you're |
|---|
| |
| |
| Last name Lastname |
| .com |
| |
| |
| • |
| to the Terms of Service and all |
| to the Privacy Policy and the data as described therein. |
| d to the CGC |
| |

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).

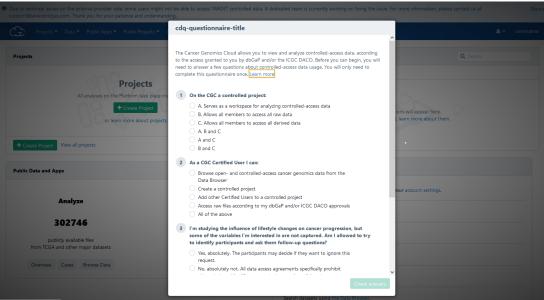


Figure A7: Screenshot of the CGC questionnaire

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

As an alternative to the <u>CCDI Hub Explore Dashboard</u>, 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 <u>Appendix A</u>.

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).

| Projects * | Data 🍝 Public Apps 🔻 Public Projects 👻 De | eveloper 🔻 | | 🜲 🔻 username |
|---------------------|---|-----------------------------------|---------------------------------------|--------------------------------------|
| Projects | Data Overview Data distributions at a glance | arch | Analysis | Q Search |
| | Case Explorer Explore processed data from public datasets | | Tasks Data Studio | |
| | Data Browser Query metadata and add files to your projects | | | |
| | Cancer Data Service Explorer Browse, search and filter dataset files | 6 | | our executions will appear here. |
| | Public Reference Files Access common reference files | 1 | Before | you start, learn more about them. |
| + Create Project | Public Test Files Access common test samples | < > | | |
| Public Data and App | Volumes Volumes | | Getting started | |
| | Data Tools | | Your account has been successfully cr | reated. Check your account settings. |
| An | alyze U: | e some of | Walk through the QuickStart. | |
| 302 | 2746 | 1011 | | |
| | | blicly available and Workflows | Create a project. | |
| | | sand worknows | Invite a collaborator. | |
| Overview Cas | ses prowse para t | rrowse Apps | Browse cases using the Case Explorer | |
| | | | | |

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

 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).

| Projects • Data • | Public Apps * Public Projects * Developer * | 4- | use |
|-------------------|--|----|-----|
| | Cancer Data Service Explorer | | |
| | | | |
| | | | |
| | Cancer Data Service (CDS) | | |
| | CDS is a cloud based repository that has shifted away from the download model. We do not encourage downloads due to user borne egress costs. For programs that have paid for download, their approved users can download data from CDS Epiporer studies. | | |
| | The Cancer Data Service (CDB) is a data repository under the National Cancer Institute's (HCI) Cancer Research Data Commons (CRDC) infrattructure for storing cancer research data generated by HCI- Intelde programs. The CDS hosts open and controlled-acoss data. The Hudles submitting controlled-data to CDS we registred with NCIS (database of Centrolypes and Phenetypes) (bGba?). To colour acoss data, that and dinical/phenotypic files a user requires authin NCIS (database of Centrolypes and Phenetypes) (bGba?). To colour acoss to controlled data Cancer Generatic Cloud which is one of the NCI-Indee Cloud BisocurcityDatom. | | |
| | PHS201437 - Reduitic Precinical Testing Consortium (PPIC) PHS201524 - The Genric Basis of Agenesise Prostate Cancer. The Role of Rare Variation PHS201524 - Decide of Conference Lance Successibility Los Using Genome-Web Sequencing | | |
| | PH501713 - Development of A Turnor Mellecular Analyses Program and Its Uter Support Treatment Decisions (UNCseqTM) PH501773 - Discovery of Contend Curnor Susceptibility Generic in Public Public Public PH501819 - Whele Genome Sequencing to Discover Familial Mysloma Risk Genes | | |
| | PH5022011 - Limited Use Pilot Test Data PH502205 - Melocular Philopolicy af Editioniogy of Colorectal Cancer PH5022566 - Lung Cancer Genetic Study among Asian News Snokers PH502271 - Limitan Timor Callio Philotophil (PMI) | | |
| | HH000354 - LICE Database for the Advancement of IAML - Integration of Mendata with "One" Data (ICED) HH000259 - Supeliment data from Biol AML calcular myoloid inkernili (ICED) HH0002502 - Realishily and Clinical IURI) of Whole Genome Pulling in Pediatric and Young Adult Cancers (ICED) HH000259 - Melocidic Characteristics Individue ICED) | | |
| | PHS002432 - Study will be released soon PHS002518 - Study will be released soon | | |
| | | | |

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).

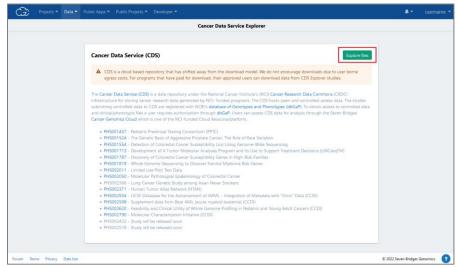


Figure B3: CDS Explorer study list page with red box highlighting "Explore files" button

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.

| uthorized | | 🖹 22890 Files | | | | | | | | | | | Copy to proje |
|--------------------------------------|------------------|---------------|---------------|-----------|---------|-------------|--------------|---------|-----------|-----------------------|--------|------|---------------|
| Authorized | # Files 22890 | | | | | | | | | | | | _ |
| Unauthorized | 0 | Authorized | Access number | File name | Dataset | Data format | Consent Code | Case ID | Sample ID | Experimental Strategy | Gender | Site | Paired er |
| ccess number | | • | PHS002790 (C | IGM_PAR | A MCI | CRAM | | PARGEV | 0D97VZ | WXS | Male | | |
| PH5001819 | # Files 39 | | PH5002790 (C | IGM_PAR | A MCI | CRAI | | PARGEV | 0D97VZ | WXS | Male | | |
| PHS001713 | 11226 | \odot | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGFV | 0D97VZ | WXS | Male | | |
| PHS001787 | 318 | \odot | PHS002790 (C | IGM_PAR | A MCI | VCF.GZ | | PARGFV | 0D97VZ | WXS | Male | | |
| PHS002599 (CCDI) PHS002517 (CCDI) | 313 63963 | | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGEV | 0D97VZ | WXS | Male | | |
| PHS002529 (CCDI) | 5375 | \odot | PHS002790 (C | IGM_PAR | A MCI | BAM | | PARGFV | 0D97VX | Archer_Fusion | Male | | |
| PHS002504 (CCDI) PHS001437 | 879 647 | ۲ | PHS002790 (C | IGM_PAR | A MCI | BAI | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| PHS002518 (CCDI) | 2245 | \odot | PHS002790 (C | IGM_PAR | A MCI | BAM | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| PHS001554 PHS002050 | 5784 5056 | ۲ | PHS002790 (C | IGM_PAR | R MCI | BAI | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| PHS001524 | 11234 | \bigcirc | PHS002790 (C | 20612977 | A MCI | IDAT | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| PHS001980 | 98 | \odot | PH5002790 (C | 20612977 | A MCI | IDAT | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| PHS002371 PHS002305 | 23405 990 | | PHS002790 (C | IGM_PAR | ê MCI | HTML | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| PHS002790 (CCDI) | 22890 | \odot | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| PHS002620 (CCDI) | 326 show less | | PHS002790 (C | IGM_PAR | A MCI | PDF | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| | | \bigcirc | PH5002790 (C | IGM_PAR | A MCI | CRAM | | PARGEV | 0D97VY | WXS | Male | | |
| ataset | E Files | \odot | PHS002790 (C | IGM_PAR | A MO | CRAI | | PARGEV | 0D97VY | WXS | Male | | |

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).

| Cancer Data Service | e (CDS) | 22890 Files | | | | | | | | | | | Copy to projec |
|--------------------------|------------------|-------------|---------------|-----------|---------|-------------|--------------|---------|-----------|-----------------------|--------|------|----------------|
| earch files by File name | ۵ | Authorized | Access number | File name | Dataset | Data format | Consent Code | Case ID | Sample ID | Experimental Strategy | Gender | Site | Paired er 🖽 |
| | | | PHS002790 (C | IGM_PAR | A MCI | CRAM | | PARGEV | 0D97VZ | wxs | Male | | |
| arch files by Case ID | ۵ | | PHS002790 (C | IGM_PAR | B MCI | CRAI | | PARGEV | 0D97VZ | WXS | Male | | |
| | | | | - | | | | | | | | | |
| arch files by Sample ID | | | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGEV | 0D97VZ | WXS | Male | | |
| | | \odot | PHS002790 (C | IGM_PAR | ₿ MCI | VCF.GZ | | PARGEV | 0D97VZ | WXS | Male | | |
| | | \odot | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGEV | 0D97VZ | WXS | Male | | |
| e filters | Clear all | \odot | PHS002790 (C | IGM_PAR | A MCI | BAM | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| thorized | | \odot | PHS002790 (C | IGM_PAR | ê MCI | BAI | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| Authorized | # Film 22890 | \odot | PHS002790 (C | IGM_PAR | A MCI | BAM | | PARGEV | 0D97VX | Archer_Fusion | Male | | |
| Unauthorized | 0 | \odot | PHS002790 (C | IGM_PAR | ê MCI | BAI | | PARGFV | 0D97VX | Archer_Fusion | Male | | |
| cess number | | \odot | PHS002790 (C | 20612977 | A MCI | IDAT | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| PH5002790 (CCDI) | # Files 22890 | \odot | PHS002790 (C | 20612977 | A MCI | IDAT | | PARGEV | 0D97VY | Methylation_Array | Male | | |
| taset | | \odot | PHS002790 (C | IGM_PAR | A MCI | HTML | | PARGFV | 0D97VY | Methylation_Array | Male | | |
| MCI | # Files 22890 | \odot | PHS002790 (C | IGM_PAR | A MCI | JSON | | PARGFV | 0D97VY | Methylation_Array | Male | | |
| | | \odot | PHS002790 (C | IGM_PAR | ê MCI | PDF | | PARGFV | 0D97VY | Methylation_Array | Male | | |
| ta format | 1 Files | \bigcirc | PHS002790 (C | IGM_PAR | A MCI | CRAM | | PARGEV | 0D97VY | WXS | Male | | |
| CRAM | 2140 2140 | \bigcirc | PHS002790 (C | IGM_PAR | B MCI | CRAI | | PARGEV | 0D97VY | wxs | Male | | |

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).

| | _ | A | | | _ | | | | | | | | | |
|---------------------------|-------------------|------------|---------------|-----------|-------------------|------------------------------|---------|-----------|-----------------------|--------|------|-----------------|--|--|
| Cancer Data Service (CDS) | | | | | | | | | | | | Copy to project | | |
| earch files by File name | | Authorized | Access number | File name | Dataset | Data format Consent Code | Case ID | Sample ID | Experimental Strategy | Gender | Site | Paired er | | |
| | | ۲ | PHS002790 (C | IGM_PAR | A MCI | CRAM | PARGEV | 0D97VZ | wxs | Male | | | | |
| earch files by Case ID | 8 | 0 | PH5002790 (C | IGM_PAR | # MCI | CRAI | PARGEV | 0D97VZ | WXS | Male | | | | |
| | | 0 | PHS002790 (C | IGM_PAR | A MCI | JSON | PARGEV | 0D97VZ | WXS | Male | | | | |
| earch files by Sample ID | 0 | 0 | PHS002790 (C | IG | files | | _ | × | wxs | Male | | | | |
| | | 0 | PHS002790 (C | | Thes | | | • | wxs | Male | | | | |
| ile filters | Clear all | 0 | PH\$002790 (C | IGh 💿 | 22890 file(s) wil | I be copied to your project. | | | Archer_Fusion | Male | | | | |
| uthorized | ŵ | \odot | PHS002790 (C | | | | | | Archer_Fusion | Male | | | | |
| Authorized | # Film 22890 | \odot | PH5002790 (C | IGN | the project whe | re you want to copy data | | | Archer_Fusion | Male | | | | |
| Unauthorized | 0 | | PHS002790 (C | | ate new project | | | | Archer_Fusion | Male | | | | |
| ccess number | | 0 | PHS002790 (C | 20 | | | | | Methylation_Array | Male | | | | |
| PH5002790 (CCDI) | # # Hink 22890 | I | PHS002790 (C | 20 | | | | Сору | Methylation_Array | Male | | | | |
| lataset | 0 | 0 | PHS002790 (C | IGM_PAR | A MCI | HTML | PARGEV | 0D97VY | Methylation_Array | Male | | | | |
| | # Files 22890 | 0 | PH5002790 (C | IGM_PAR | A MCI | JSON | PARGEV | 0D97VY | Methylation_Array | Male | | | | |
| | | ۲ | PHS002790 (C | IGM_PAR | A MCI | PDF | PARGEV | 0D97VY | Methylation_Array | Male | | | | |
| ata format | 0 + 21m | 0 | PH5002790 (C | IGM_PAR | A MCI | CRAM | PARGEV | 0D97VY | wxs | Male | | | | |
| CRAM | 2140 | . 0 | PHS002790 (C | IGM PAR | A MO | CRAI | PARGEV | 0D97VY | WXS | Male | | | | |

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 <u>NIH eRA Commons account</u> for authentication, after which they will need to obtain authorization (via an active DCFS <u>login account</u>) to access the data in the NIH <u>dbGaP</u>.

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 <u>NCI DCF Services</u> portal - <u>https://nci-crdc.datacommons.io/login</u> (Figure C1).

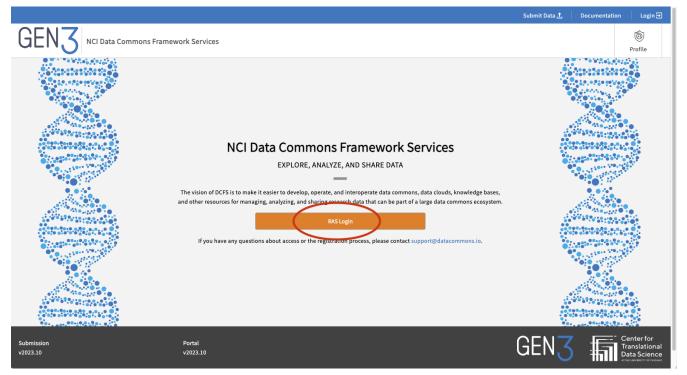


Figure C1: NIH Data Commons Framework (DCF) homepage with the NIH Researcher Auth Service (RAS) login highlighted

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

| | Submit Da | ata 📩 🕴 Documentation | ryus5 ® Logout € |
|--|--------------------|-----------------------|---|
| GEN3 NCI Data Commons Framework Services | | 1. click p | rofile |
| Create API key P You don't have any API key. Please create one! | | • | \smile |
| You have the following API key(s) | | | |
| API key(s) | Expires | | |
| You have access to the following project(s) | | | |
| Project(a) 2 check the project access | Method(s) | | |
| phs003291 | read-storage, read | | |
| | | | |
| | | | |
| | | | |
| Submission Portal v2023.10 v2023.10 | | GENZ | Center for Translational Data Science |

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).

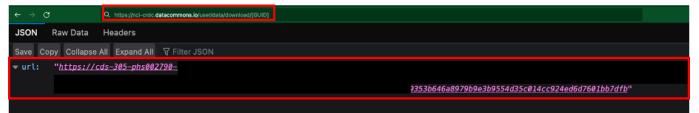


Figure C3: DCF Service Portal displaying the signed access URL and the file download URL

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 <u>Gen3-client command-line tool</u> 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://ncicrdc.datacommons.io'.
- 3. Obtain either a GUID or manifest of GUIDs for the data files of interest from the <u>CCDI Explore page</u> or the <u>Explore Dashboard exportable manifest</u>
- 4. Created a gen3 structured manifest:

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

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