

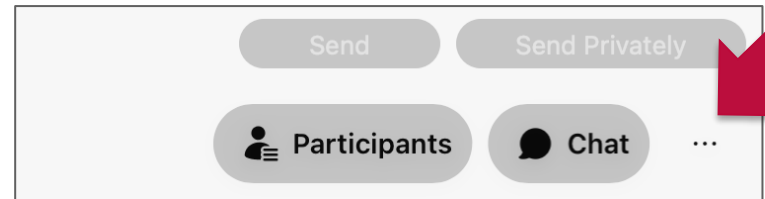
# Downloading Large-Scale Datasets at the GDC

**23 October 2023**

Bill Wysocki, Ph.D. – GDC User Services Lead  
Center for Translational Data Science  
University of Chicago

# Downloading Large-Scale Datasets at the GDC

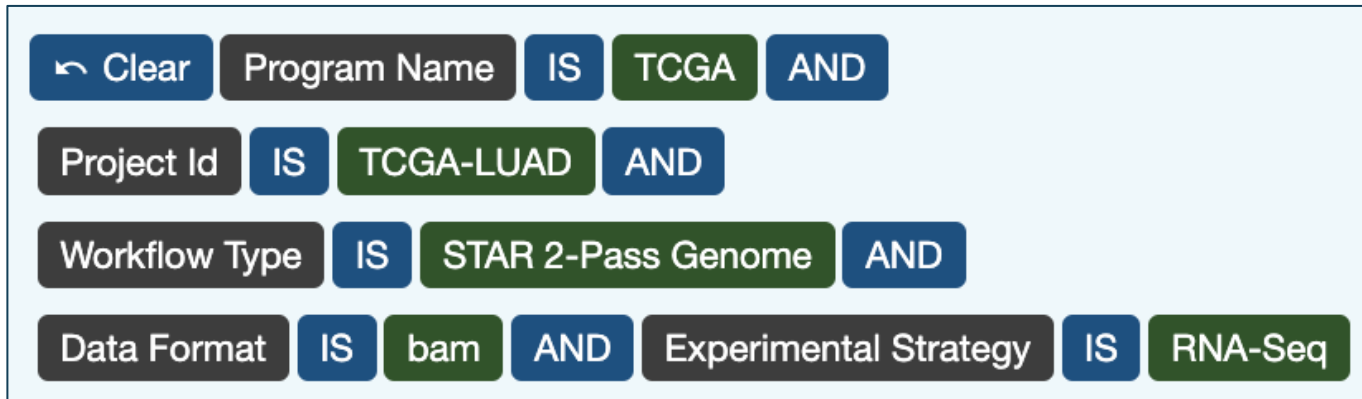
- *Brief Introduction*
- *Data Transfer Tool*
- *API Download*
- *Troubleshooting*
- *Q&A*



# Introduction to GDC File Downloads

# Genomic Data Commons File Download

- **The NCI's Genomic Data Commons (GDC)** provides the cancer research community with a unified repository and cancer knowledge base that enables data sharing across cancer genomic studies in support of precision medicine.
  - Large-scale downloads are focused on **Data Files over 5 GB**
  - Files can be browsed and filtered from the **GDC Data Repository**



# Options for Large File Download

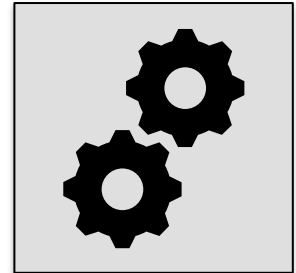
## Option 1: Data Transfer Tool

- Standalone tool using the command line
- Uses GDC API to download and applies settings automatically
- Download from:
  - <https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>



## Option 2: GDC API

- Download directly from GDC API
- Uses other software to access (**curl** in this presentation)
- More customizable in terms of settings, less automated



# Starting Point 1: One File UUID

One slide image from the TCGA-CESC project

File Properties	
<b>Name</b>	TCGA-XS-A8TJ-01Z-00-DX1.3CB10EF8-8A92-472B-8B5D-6CA88C8A70D5.svs
<b>Access</b>	open
<b>UUID</b>	216feaac-8b0c-468d-991f-0412215e7a02
<b>Data Format</b>	SVS
<b>Size</b>	5.12 GB
<b>MD5 Checksum</b>	cc40c34cb7639ae7f74e5dddba8225c7
<b>Archive</b>	--
<b>Project</b>	<a href="#">TCGA-CESC</a>

# Starting Point 2: Manifest with Many Files

All slide images from TCGA-CESC (open access)

Clear Project Id IS TCGA-CESC AND Data Type IS Slide Image

id	filename	md5	size	state	
a9e316b2-abcfc-4e40-870d-3e1d74abf8e4				TCGA-VS-A8EI-01A-01-TS1.64C2A4BF-CE1B-46CB-AE47-44D4BBF51ED6.svs	c9526a0e3df583efda8f0dc61bb21
040229b3-224c-4107-bd33-0854196b6423				TCGA-VS-A8EI-01Z-00-DX1.8DD9CBFB-C3B2-48D0-ADEE-046197F481A7.svs	d4ecae7c6f8f467afbcd060058ffd
73731492-4bc8-47b7-846b-d668bba76e77				TCGA-EA-A410-01A-01-TSA.C6985C2C-5532-43A0-A910-8171D63A07AA.svs	a347db27e95499197e9e0c8dcccfd8
7a0bd065-f980-4b17-9e49-d8bd3f7d4da1				TCGA-EA-A410-01Z-00-DX1.40217EF9-3F9A-4669-A78E-AC851F62E532.svs	4ed019748b6c798f696fb699a0ad7
973fd2d2-20aa-4c69-ab78-fbdce0c056ed				TCGA-JX-A5QV-01A-02-TSB.DEAD6C31-A859-4FE6-A78A-8F1A7993D622.svs	a3f276e4f7b74f8f2f35b86e2c885
92333ba0-aba2-481e-b0d4-2a25a308add8				TCGA-JX-A5QV-01Z-00-DX1.90769414-2C5C-4BAA-A432-9FC0A15EEF5A.svs	ddf5d47ae433cfad3249c65ca90dc
d97dccc5-faf8-40b7-a985-f8d8e060f9f9				TCGA-VS-A9UR-01A-01-TS1.329F5E80-EDFB-4C90-8374-06019A949E26.svs	73687c0037aff4fadd27895d5fdcf
837cb893-321f-45b8-a600-cff5e6a342a8				TCGA-C5-A8XI-01Z-00-DX1.7E8E0C2A-D3D4-4DCA-9B4E-70A8EBC48E67.svs	df7572482b006186108ef4abaa198
83ce77ff-cb47-4862-8da5-47383f005c03				TCGA-VS-A8EC-01A-01-TS1.A7457E86-58E0-4387-8B81-35E5F10F6AF5.svs	f8858cb8884f82f242a8a04db86ed
39c1e6be-e703-4de4-a25d-36ea6ffd3722				TCGA-FU-A57G-01Z-00-DX1.3007337B-A044-4831-B957-F8740E9ACB4E.svs	f2fe797b7878af87ce412a1499bf3
89e46249-a64d-4394-b13a-a7c918cf8438				TCGA-FU-A57G-01A-01-TS1.8552F1C4-20C7-4D3F-9451-C2F1881BF8BF.svs	16694587f403c4df6693858e131fc
e415e8ad-29f4-4004-827e-dce12944609d				TCGA-ZJ-AAXB-01A-01-TSA.AE149D7F-1291-4809-A2A8-EF41F5C5FECB.svs	406d62034c5e3d31283257df5d4ef
d8e4ebd1-a824-4022-8f51-bafdb8b5978f				TCGA-VS-A954-01A-01-TS1.33926C20-EA65-4ED1-B9DA-DACAB54FD8BF.svs	325ea922b2b922c7b813e58d2068f
b0f09af7-da56-4776-8adb-d97ad83e0935				TCGA-EA-A1QT-01A-01-TSA.5a517400-267d-4569-9c20-6f00a08e7ed7.svs	dbece92dac9e52ba6fff98531f3c3c
4f7c10c9-ccaa-46eb-b202-71b1216c61cd				TCGA-C5-A2M1-01Z-00-DX1.E03FE8EC-002B-4673-ACC5-A32F1CA94A98.svs	cdfefb5218b30ca7913bb6daab224
dc960749-7440-489c-8861-1fbf07323fd1				TCGA-Q1-A73S-01Z-00-DX1.D74CFF58-A032-45D2-98FA-4B4D9AB90069.svs	3b53fcd00658f3adfd3d3b511332h
2ca5c47d-120b-4f08-90e9-a9a345393bf1				TCGA-C5-A1M6-01Z-00-DX1.13F7405D-AD0E-4A1C-9DF4-00DC90756D28.svs	a9116ae8a7024a9330e1f2c54c54h
90911d5b-8307-439c-8bd7-ca27f08eefae				TCGA-VS-A8QH-01Z-00-DX1.FE72CE6D-0140-4A57-83CC-38F5EAD09FES.svs	140e94ecc24154e651a260c811979
733525ba-9ce1-43b2-965f-8e36c43d7bc5				TCGA-VS-A957-01Z-00-DX1.FE01C75F-EAF4-4421-A250-E083BC1AFB14.svs	f69032e4e043b3e983e637e918123

# Data Transfer Tool Demo



# Token Information



- The files we will be downloading today will be larger and open-access
- A simulated token will be used for demonstration purposes
  - Most large-scale download involves controlled data
  - This simulated token is not necessary but will not interfere

## Token File

```
sim_token.txt
```

## Token String (simulated)

```
aaabbbccdddeefffggg1112  
22333444555
```

# GDC Data Transfer Tool Commands (1/5)

One UUID:

```
./gdc-client download  
  
216feaac-8b0c-468d-991f-0412215e7a02  
  
-t sim_token.txt
```

a) Runs the Data Transfer Tool

# GDC Data Transfer Tool Commands (2/5)

One UUID:

```
./gdc-client download  
216feaac-8b0c-468d-991f-0412215e7a02  
  
-t sim_token.txt
```

- a) Runs the Data Transfer Tool
- b) Uses the download function

# GDC Data Transfer Tool Commands (3/5)

One UUID:

```
./gdc-client download  
216feaac-8b0c-468d-991f-0412215e7a02  
  
-t sim_token.txt
```

- a) Runs the Data Transfer Tool
- b) Uses the download function
- c) Specifies the file UUID

# GDC Data Transfer Tool Commands (4/5)

One UUID:

```
./gdc-client download  
216feaac-8b0c-468d-991f-0412215e7a02  
-t sim_token.txt
```

- a) Runs the Data Transfer Tool
- b) Uses the download function
- c) Specifies the file UUID
- d) **Specifies the token file**

# GDC Data Transfer Tool Commands (5/5)

Manifest with many UUIDs:

```
./gdc-client download  
-m gdc_manifest.2023-10-16.txt  
-t sim_token.txt
```

- a) Runs the Data Transfer Tool
- b) Uses the download function
- c) Specifies the manifest file
- d) Specifies the token file

# GDC API Demo

# GDC API Commands: Token Management

Store the token string as a variable for use with Curl

```
export MYTOKEN=$(cat sim_token.txt)
```

Verify that the token string was successfully stored

```
echo $MYTOKEN
```



# GDC API Commands (1/6)

One UUID:

```
curl (-X GET)
```

```
-H "x-auth-token: $MYTOKEN"
```

```
--remote-name --remote-header-name
```

```
"https://api.gdc.cancer.gov/data/  
216feaac-8b0c-468d-991f-0412215e7a02  
?related_files=true"
```

- a) Runs curl software, request type GET is default

# GDC API Commands (2/6)

One UUID:

```
curl (-X GET)
-H "x-auth-token: $MYTOKEN"

--remote-name --remote-header-name

"https://api.gdc.cancer.gov/data/
216feaac-8b0c-468d-991f-0412215e7a02
?related_files=true"
```

- a) Runs curl software, request type GET is default
- b) Specifies header with token string

# GDC API Commands (3/6)

One UUID:

```
curl (-X GET)

-H "x-auth-token: $MYTOKEN"

--remote-name --remote-header-name

"https://api.gdc.cancer.gov/data/
216feaac-8b0c-468d-991f-0412215e7a02
?related_files=true"
```

- a) Runs curl software, request type GET is default
- b) Specifies header with token string
- c) Downloads file name from API

# GDC API Commands (4/6)

One UUID:

```
curl (-X GET)

-H "x-auth-token: $MYTOKEN"

--remote-name --remote-header-name

"https://api.gdc.cancer.gov/data/
216feaac-8b0c-468d-991f-0412215e7a02
?related_files=true"
```

- a) Runs curl software, request type GET is default
- b) Specifies header with token string
- c) Downloads file name from API
- d) Main API URL with /data endpoint

# GDC API Commands (5/6)

One UUID:

```
curl (-X GET)

-H "x-auth-token: $MYTOKEN"

--remote-name --remote-header-name

"https://api.gdc.cancer.gov/data/
216feaac-8b0c-468d-991f-0412215e7a02
?related_files=true"
```

- a) Runs curl software, request type GET is default
- b) Specifies header with token string
- c) Downloads file name from API
- d) Main API URL with /data endpoint
- e) Specifies UUID

# GDC API Commands (6/6)

One UUID:

```
curl (-X GET)

-H "x-auth-token: $MYTOKEN"

--remote-name --remote-header-name

"https://api.gdc.cancer.gov/data/
216feaac-8b0c-468d-991f-0412215e7a02
?related_files=true"
```

- a) Runs curl software, request type GET is default
- b) Specifies header with token string
- c) Downloads file name from API
- d) Main API URL with /data endpoint
- e) Specifies UUID
- f) Allows for index files to be downloaded (BAM and VCF only)

# Downloading Multiple files using the API

## Option 1: Use API command and loop through list of UUIDs

- Can be performed using bash or Python scripts

## Option 2: Pass JSON formatted list of UUIDs

- Uses a POST request with header - “Content-Type: application/json”
- Requires conversion of list of UUIDs to JSON file

## Option 3: Use comma delimited list to specify multiple UUIDs in one line

- Same as GET request in demo
- Limited by URL length

```
{
  "ids": [
    "UUID1",
    "UUID2",
    ...
    "UUID3"
  ]
}
```

# Final Results: Downloaded Files

## Data Transfer Tool

- Files will be downloaded in folders named after their UUIDs
- The md5sum has been verified

## API Download

- Files will be downloaded under their respective filenames in your current directory unless otherwise specified
- We recommend checking the md5sum against the file's properties

The demonstrations in this webinar were based on MacOS or any other Unix-based terminal. These functions are all available on Windows.

- **Documentation and personalized assistance is available**



# Troubleshooting Data Download

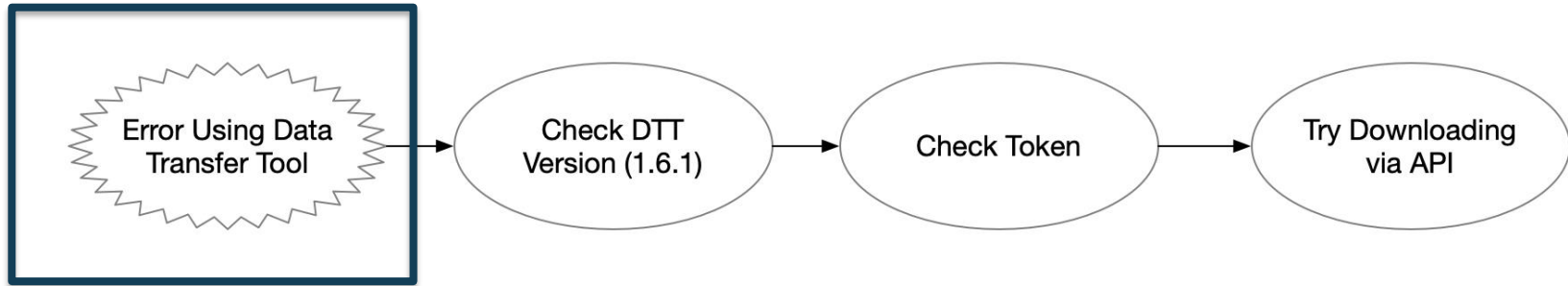
# Troubleshooting Data Transfer Tool Errors

- The GDC Data Transfer Tool can be used by researchers on a wide variety of operating systems. However, errors can arise due to security settings, connection issues, etc.
- Errors may be informative depending on the issue

Examples of informative error messages:

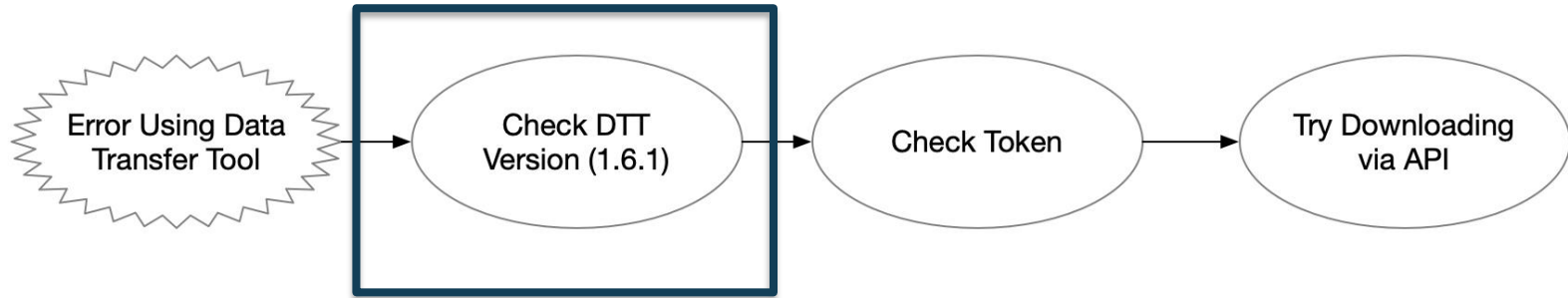
- `./gdc-client: No such file or directory`
  - **Solution:** The command needs to be pointed at a different directory
- `Your token is invalid or expired. Please get a new token from GDC Data Portal`
  - **Solution:** Investigate the token file

# DTT Error: Three Step Troubleshooting Flowchart



- Flowchart starts at a user receiving an error that doesn't specify the exact problem
- This series of checks will allow the user to either solve or narrow down the issue

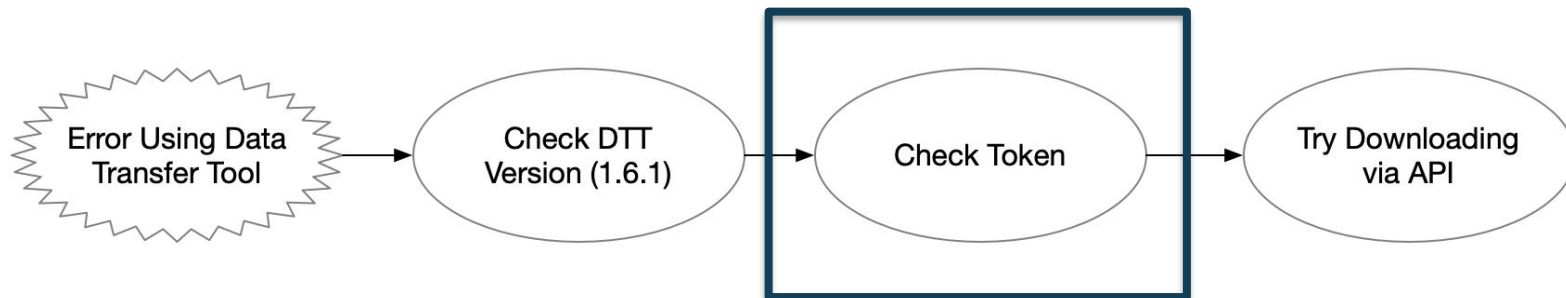
# Step 1: Check Data Transfer Tool Version



- The GDC has continuously released new versions of the data transfer tool to add new features and bug fixes
  - Based on user/developer feedback
  - Latest version is always available at [gdc.cancer.gov](http://gdc.cancer.gov)
- **Command:** `./gdc-client --version`

```
./gdc-client --version  
v1.6.1
```

## Step 2: Check Authentication Token

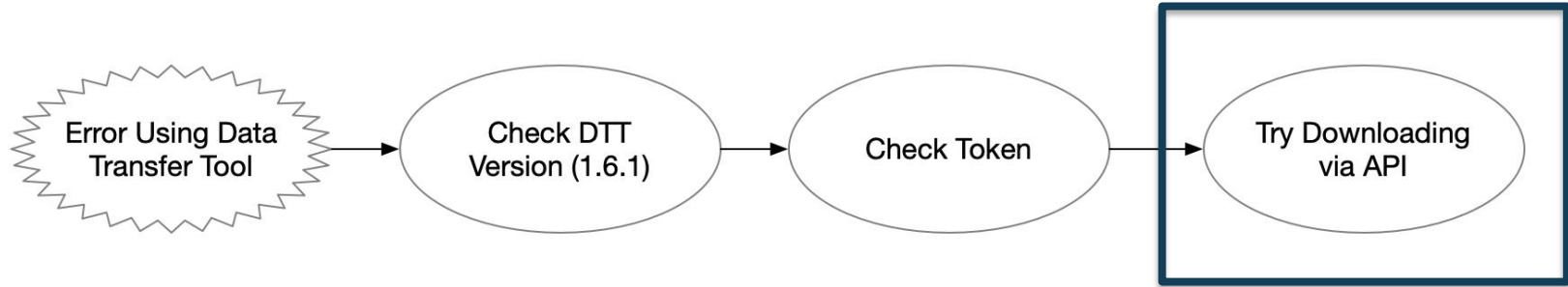


- The token is a common source of errors, because multiple issues can arise. The following criteria must be met to be a valid token.
  - The token must be current → **Reset token**
  - The token must be correctly parsed → **Check for spaces or truncated token**
  - The user must have dbGaP access to the project → **Check user profile**

The dropdown menu shows options: User Profile, Download Token, and Logout. An arrow points from the 'User Profile' option to the table below.

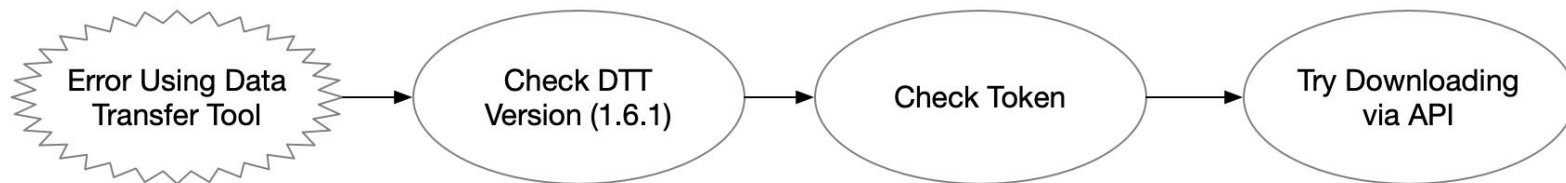
Project ID	create	read_report	read	release	update	_member_
GDC-INTERNAL	✓	✓	✓	✓	✓	✓

## Step 3: Download using the GDC API Directly



- Download errors with the Data Transfer Tool could arise from software incompatibility but could also stem from connection issues or security settings
- A successful download with the API rules out issues with your connection to the GDC
- This may also solve download issues if your downloads finish via API testing
- **Quick command:** `curl https://api.gdc.cancer.gov/status`

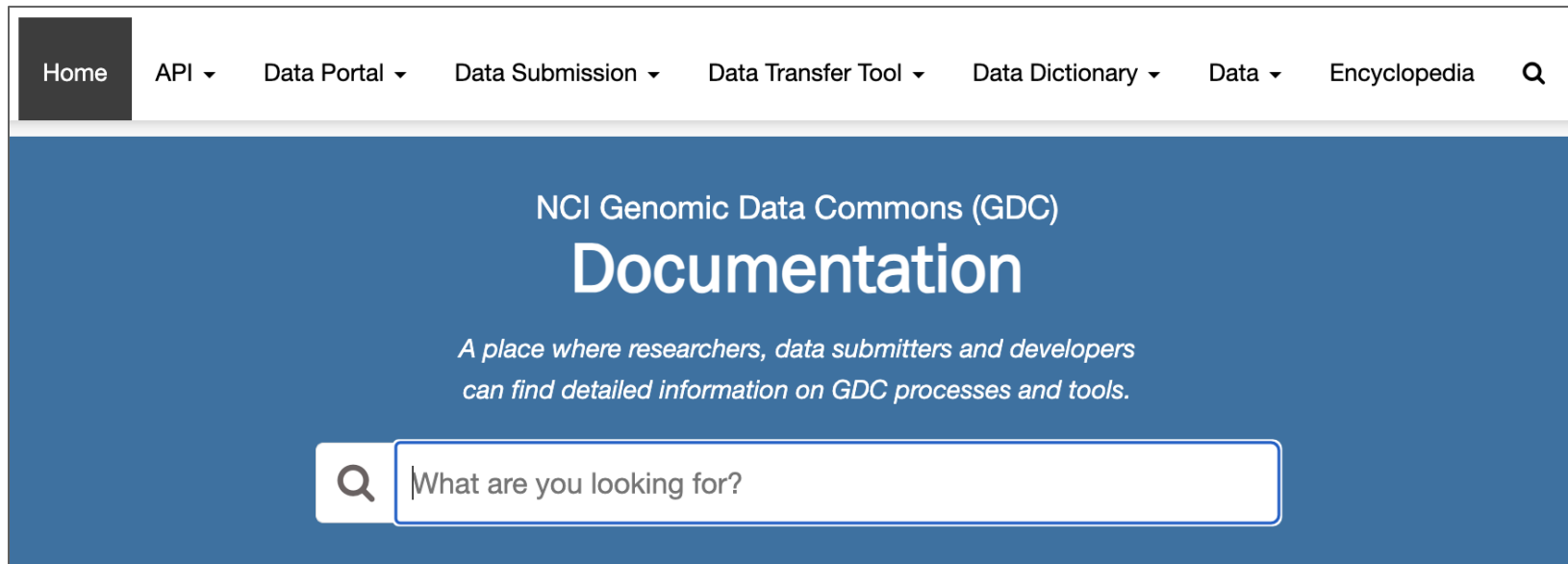
# GDC Help Desk



- Send an email to [support@nci-gdc.datacommons.io](mailto:support@nci-gdc.datacommons.io) for assistance with data download
- Provide information you gathered from the previous steps, and we can help you diagnose the issue
- The GDC Help Desk is also happy to help walk you through any of the previous steps outlined here
- We also recommend reaching out if you are using an operating system that isn't Windows, MacOS, or Ubuntu

# Useful Links – GDC Documentation

- <https://docs.gdc.cancer.gov>



Home API ▾ Data Portal ▾ Data Submission ▾ Data Transfer Tool ▾ Data Dictionary ▾ Data ▾ Encyclopedia 🔍

NCI Genomic Data Commons (GDC)  
**Documentation**

*A place where researchers, data submitters and developers can find detailed information on GDC processes and tools.*

🔍 What are you looking for?



# Useful Links – GDC Website

- <https://gdc.cancer.gov>

**NIH NATIONAL CANCER INSTITUTE Genomic Data Commons**

CCG Web Site | Contact Us | [Launch Data Portal](#) | GDC Apps

Search this website

/\*\*/

About the GDC | About the Data | Analyze Data | Access Data | Submit Data | For Developers | Support | News

## The Next Generation Cancer Knowledge Base

Cases by Major Primary Site

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified repository and cancer knowledge base that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and

### Analyze Data

The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

→ [More about Analyzing Data](#)

### Access Data

The **GDC Data Portal** provides a platform for efficiently querying and

# Useful Links – Additional Support

- [support@nci-gdc.datacommons.io](mailto:support@nci-gdc.datacommons.io)

The screenshot displays the National Cancer Institute Genomic Data Commons website. At the top, the NIH logo and 'NATIONAL CANCER INSTITUTE Genomic Data Commons' are visible. Navigation links include 'CCG Web Site', 'Contact Us', 'Launch Data Portal', and 'GDC Apps'. A search bar is present with the text 'Search this website'. A dark navigation bar contains links for 'About the GDC', 'About the Data', 'Analyze Data', 'Access Data', 'Submit Data', 'For Developers', 'Support', and 'News'. The 'Support' link is highlighted with a red box. Below the navigation bar, a section titled 'Support' features a question mark icon and the text: 'The GDC provides documentation and support resources to assist users. Explore our support resources »'. A 'TECHNICAL QUESTIONS' section is also present, containing three links: 'Monthly Support Webinar', 'Help Desk: support@nci-gdc.datacommons.io', and 'Join the GDC User Listserv'. These three links are enclosed in a red box. Below this, 'ADDITIONAL CONTACT INFORMATION' includes 'Twitter' and 'National Cancer Institute' links. The bottom of the page features a colorful pie chart, a section titled 'Studies in support of precision medicine' with text about CCG and TCGA programs, and an 'Access Data' section with a DNA helix icon and text about the GDC Data Portal.

Questions?

U.S. Department of Health & Human Services  
National Institutes of Health | National Cancer Institute

<https://www.cancer.gov/>

1-800-4-CANCER

Produced October 2023