

# Introducing GDC 2.0: A New Cohort-Centric Design

**23 February 2024**

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

## New Features in GDC 2.0

1. *Cohort-Centric Workflow Intro*
2. *Demo: Building a Cohort*
3. *Demo: Core Tools*
4. *Demo: Analysis Tools*
5. *Tutorials, Guides, and Support*
6. *Questions*



# Introduction to the Cohort-Centric Workflow

*GDC 2.0*

# GDC 2.0 – A new cohort-centric workflow

■ **Cohort – A group of cases that share a set of characteristics**

- The new GDC 2.0 workflow separates the cohort-building process from all other functions of the portal
- This allows for the analyses to be performed by on the same group of cases across all available GDC tools



# GDC 2.0 Workflow

## Build Cohort



### Cohort Builder

Build and define your custom cohorts using a variety of clinical and biospecimen features.



## Download Cohort Data



### Repository

Browse and download the files associated with your cohort for more sophisticated analysis.



## View Projects



### Projects

View the Projects available within the GDC and select them for further exploration and analysis.



## Analyze Cohort

### ANALYSIS TOOLS

**BAM Slicing Download** ▶  
1,121 Cases

**Clinical Data Analysis** ▶ Demo  
1,310 Cases

**Cohort Comparison** ▶ Demo  
1,310 Cases

**Gene Expression Clustering** ▶ Demo  
1,039 Cases

**Mutation Frequency** ▶ Demo  
1,039 Cases

**OncoMatrix** ▶ Demo  
1,039 Cases

**ProteinPaint** ▶ Demo  
1,039 Cases

**Sequence Reads** ▶  
1,121 Cases

**Set Operations** ▶ Demo

# GDC 2.0 Workflow: Step 1

## Build Cohort



### Cohort Builder

Build and define your custom cohorts using a variety of clinical and biospecimen features.



Step 1: Build a cohort based on clinical or biospecimen attributes

### Repository

Browse and download the files associated with your cohort for more sophisticated analysis.

## View Projects

### Projects

View the Projects available within the GDC and select them for further exploration and analysis.

## Analyze Cohort

### ANALYSIS TOOLS

BAM Slicing Download •  
1,201 Cases

Clinical Data Analysis •  
1,201 Cases

Cohort Comparison •  
1,201 Cases

Gene Expression Clustering •  
1,201 Cases

Mutation Frequency •  
1,201 Cases

OncoMatrix •  
1,201 Cases

ProteinPaint •  
1,201 Cases

Sequence Reads •  
1,201 Cases

Set Operations •  
1,201 Cases

# GDC 2.0 Workflow: Step 2

Step 2: Use the cohort with tools in the analysis center.

Tools will be automatically applied to the cohort.

## Download Cohort Data



### Repository

Browse and download the files associated with your cohort for more sophisticated analysis.



## View Projects



### Projects

View the Projects available within the GDC and select them for further exploration and analysis.



## Analyze Cohort

### ANALYSIS TOOLS

**BAM Slicing Download** ▶  
1,121 Cases

**Clinical Data Analysis** ▶ Demo  
1,310 Cases

**Cohort Comparison** ▶ Demo  
1,310 Cases

**Gene Expression Clustering** ▶ Demo  
1,039 Cases

**Mutation Frequency** ▶ Demo  
1,039 Cases

**OncoMatrix** ▶ Demo  
1,039 Cases

**ProteinPaint** ▶ Demo  
1,039 Cases

**Sequence Reads** ▶  
1,121 Cases

**Set Operations** ▶ Demo

# GDC 2.0 Workflow: Next Step

## Build Cohort



### Cohort Builder

Build and define your custom cohorts using a variety of clinical and biospecimen features.



## Download Cohort Data



### Repository

Browse and download the files associated with your cohort for more sophisticated analysis.



## View Projects



### Projects

View the Projects available within the GDC and select them for further exploration and analysis.



## Analyze Cohort

### ANALYSIS TOOLS

**BAM Slicing Download** ▶  
1,121 Cases

**Clinical Data Analysis** ▶ Demo  
1,310 Cases

**Cohort Comparison** ▶ Demo  
1,310 Cases

**Gene Expression Clustering** ▶ Demo  
1,039 Cases

**Mutation Frequency** ▶ Demo  
1,039 Cases

**OncoMatrix** ▶ Demo  
1,039 Cases

**ProteinPaint** ▶ Demo  
1,039 Cases

**Sequence Reads** ▶  
1,121 Cases

**Set Operations** ▶ Demo





# Demo: Building a Cohort

*GDC 2.0*

# Building a Cohort – GDC 2.0 Cohort Builder

- We will use the GDC 2.0 Cohort Builder to build a cohort with the following properties
  1. Primary tumor is from the kidney (General)
  2. Resection or biopsy came from the kidney (Diagnosis)
  3. Case's gender is male (Demographic)
  4. Case has WGS data (Available Data)
- Save and name cohort → This is our **Active Cohort**

# Building a Cohort – Other Features

- Import / Export Cohort
  - Importing and exporting cohort creates a list of cases
  - Cohort builder creates a set of filters
- Create New Cohort
  - One new cohort is allowed at a time
  - Save and name your cohort to create a new one

# Demo: GDC Core Tools

*GDC 2.0*

# Analysis Center – Central Hub for Data Analysis

- The core tools and analysis tools can be reached from the analysis center
- Click the name of each tool to see a description
- Click on the "play" button to launch the tool for your active cohort

# Repository – Data File Download

- **Goal:** Download the WGS BAM files for our active cohort
- The files in the Repository are the files associated with your active cohorts
- Further narrow them down with the facet filters on the left
- Add files to the cart to download files or a Data Transfer Tool manifest
- Biospecimen and clinical data is available for files in the cart

# Projects – Browse GDC Projects

- **Goal:** Create a new cohort with only projects that include:
  - Cases with primary site: **kidney**
  - Projects from the **TARGET** program
- Start with a new cohort, apply filters
- Select Projects → “Save New Cohort” Button → Name Cohort

# Demo: GDC Analysis Tools

*GDC 2.0*



# Mutation Frequency – Browse Genes and Mutations

- Displays mutations and genes from active cohort as well as
  - Most frequently mutated genes
  - Survival plot
- Switch between genes and mutations at the top
- Narrow down your genes and mutations using:
  - Filters on the left side of the portal
  - Custom gene/mutation sets

# Clinical Data Analysis – Visualize Clinical Data

- Clinical data fields are displayed based on toggle switches in left panel
- Histogram bins can be customized based on categories or ranges
- Survival plots can be customized to compare categories within your cohort
- Each graphic can be exported as an image (SVG/PNG) or as data (JSON)

# Set Operations – Compare Cohorts or Sets

- Up to three sets can be compared
- Each shared subset is visualized as a formula or graphic and selected
- Export set as new cohort or TSV

# ProteinPaint – Visualize Mutations on a Protein

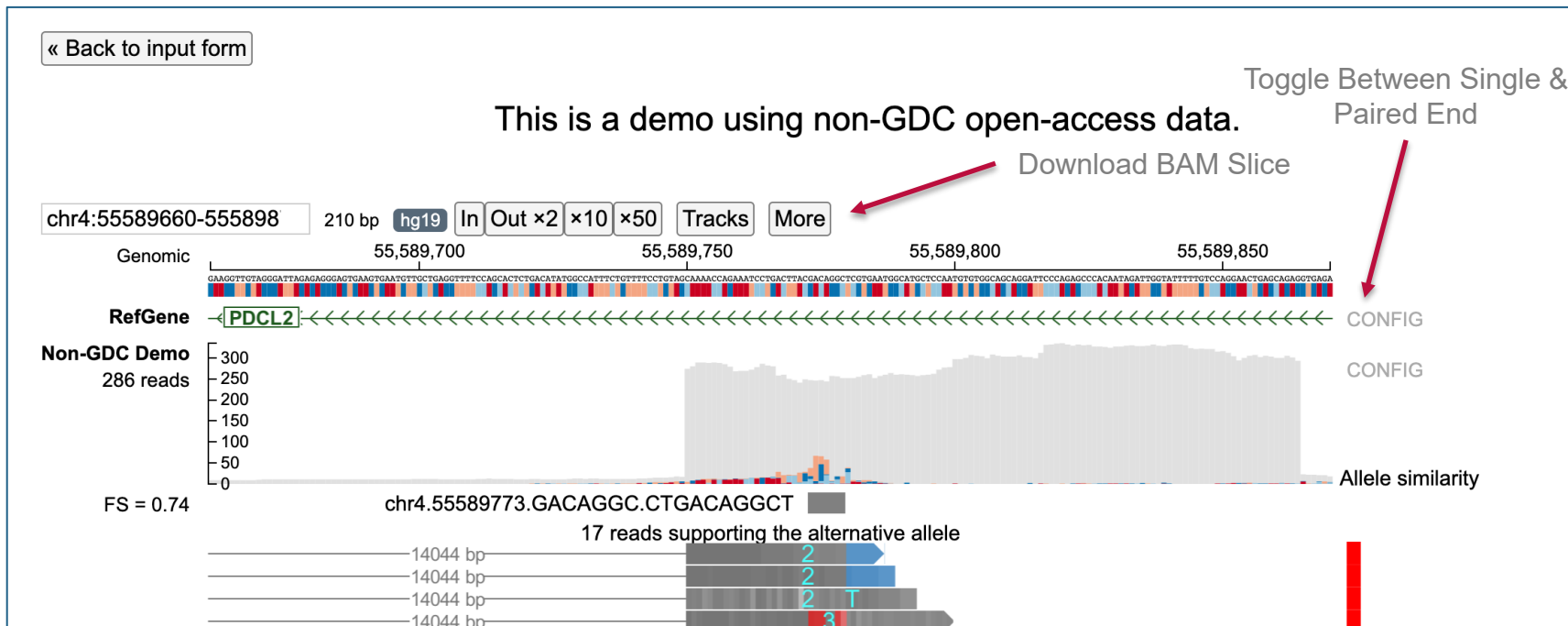
- Choose gene for ProteinPaint to visualize
- Each mutation is represented as a lollipop
  - Number represents number of cases with the mutation in the active cohort
  - Color represents mutation consequence
- Disco plot – representing the full set of mutations in the genome – can be visualized and exported
- Cohort can be created based on selected cases.

# OncoMatrix – Visualize Mutations in a Matrix

- Displays the top mutated genes in the active cohort
- Columns are cases, rows are genes
- Colored cells represent mutation occurrences
- Colors represent mutation consequences
- Appearance is customizable at the top



# GDC 2.0 Analysis Tools: Sequence Reads





# GDC 2.0 Analysis Tools: BAM Slicing Download Step 1

Step 1: Select a BAM file in your cohort. Use an identifier or select from the list

The screenshot shows the GDC 2.0 Analysis Tools interface. At the top, there is a search bar with the text "Kidney - Male" and a dropdown arrow. To the right of the search bar are icons for save, add, delete, upload, and download. Further right, there is a button that says "486 CASES" and a pin icon.

Below the search bar, there is a filter bar with the text "Kidney - Male" and a "Clear All" link. To the right of the filter bar are icons for expand and collapse. Below the filter bar, there are several filter tags: "TISSUE OR ORGAN OF ORIGIN" with a left arrow, "kidney, nos" with a right arrow and a close button, "GENDER" with a left arrow, "male" with a right arrow and a close button, "SITE OF RESECTION OR BIOPSY" with a left arrow, "kidney, nos" with a right arrow and a close button, "EXPERIMENTAL STRATEGY" with a left arrow, and "WGS" with a right arrow and a close button.

Below the filter bar, there is a section titled "BAM SLICING DOWNLOAD" with a close button. Underneath, there is a search bar with the text "Enter search string" and a "Submit" button. To the right of the search bar is a dropdown menu with the text "File Name / File UUID / Case ID / Case UUID". Below the search bar, there is a link that says "Or, browse 1000 available BAM files".

Below the link, there is a list of BAM files. The list is organized into two columns. The first column contains the case ID, and the second column contains the file name and size. The files are as follows:

CASE	BAM FILES, SELECT ONE TO VIEW
TCGA-AR-A1AX	Blood Derived Normal, WGS 161.43 GB
	Primary Tumor, RNA-Seq 86.52 MB
	Primary Tumor, miRNA-Seq 67.74 MB
	Primary Tumor, WGS 435.11 GB
	Blood Derived Normal, WXS 37.92 GB
	Primary Tumor, RNA-Seq 7.00 GB
TCGA-OL-A66N	Primary Tumor, WXS 36.21 GB
	Primary Tumor, miRNA-Seq 203.18 MB
	Blood Derived Normal, WGS 153.28 GB
	Primary Tumor, WGS 351.95 GB
	Primary Tumor, RNA-Seq 7.23 GB
	Primary Tumor, RNA-Seq 56.01 MB
AP	Blood Derived Normal, WXS 13.14 GB
	Primary Tumor, WXS 15.30 GB

At the bottom of the page, there are two columns of links. The first column is titled "MORE INFORMATION" and contains links for "Site Home", "Support", and "Listserv". The second column is titled "POLICIES" and contains links for "Accessibility", "Disclaimer", and "FOIA".



# GDC 2.0 Analysis Tools: BAM Slicing Download Step 2

Step 2: Select a region. Use a variant, gene, coordinates, or unmapped reads.

**BAM SLICING DOWNLOAD**

Enter search string  ✓  
Or, browse 1000 available BAM files

Entity ID   
Experimental Strategy   
Sample Type   
Size

24 variants [Gene or position](#) Unmapped reads

Enter gene, position, SNP, or variant  Press ENTER to search, ESC to cancel

- Enter gene, position, SNP, or variant.  file will be sliced at the given position and visualized.
- Position
  - Example: chr17:7676339-7676767
  - Coordinates are hg38 and 1-based.
- SNP example: rs28934574
- Variant:
  - Example: chr2.208248388.C.T
  - Fields are separated by periods. Coordinate is hg38 and 1-based. Reference and alternative alleles are on forward strand.
- Supported HGVS formats for variants:
  - SNV: chr2:g.208248388C>T
  - MNV: chr2:g.119955155\_119955159delinsTTTTT
  - Insertion: chr5:g.171410539\_171410540insTCTG
  - Deletion: chr10:g.8073734delTTTAGA





# GDC 2.0 Analysis Tools: BAM Slicing Download: Step 3

Step 3: Click “Submit”. BAM file will download.

**BAM SLICING DOWNLOAD**

Enter search string  ✓  
Or, browse 1000 available BAM files

Entity ID   
Experimental Strategy   
Sample Type   
Size

24 variants [Gene or position](#) Unmapped reads

Enter gene, position, SNP, or variant  ✓ VHL

- Enter gene, position, SNP, or variant. The BAM file will be sliced at the given position and visualized.
- Position
  - Example: chr17:7676339-7676767
  - Coordinates are hg38 and 1-based.
- SNP example: rs28934574
- Variant:
  - Example: chr2.208248388.C.T
  - Fields are separated by periods. Coordinate is hg38 and 1-based. Reference and alternative alleles are on forward strand.
- Supported HGVS formats for variants:
  - SNV: chr2:g.208248388C>T
  - MNV: chr2:g.119955155\_119955159delinsTTTTT
  - Insertion: chr5:g.171410539\_171410540insTCTG
  - Deletion: chr10:g.8073734delTTTAGA

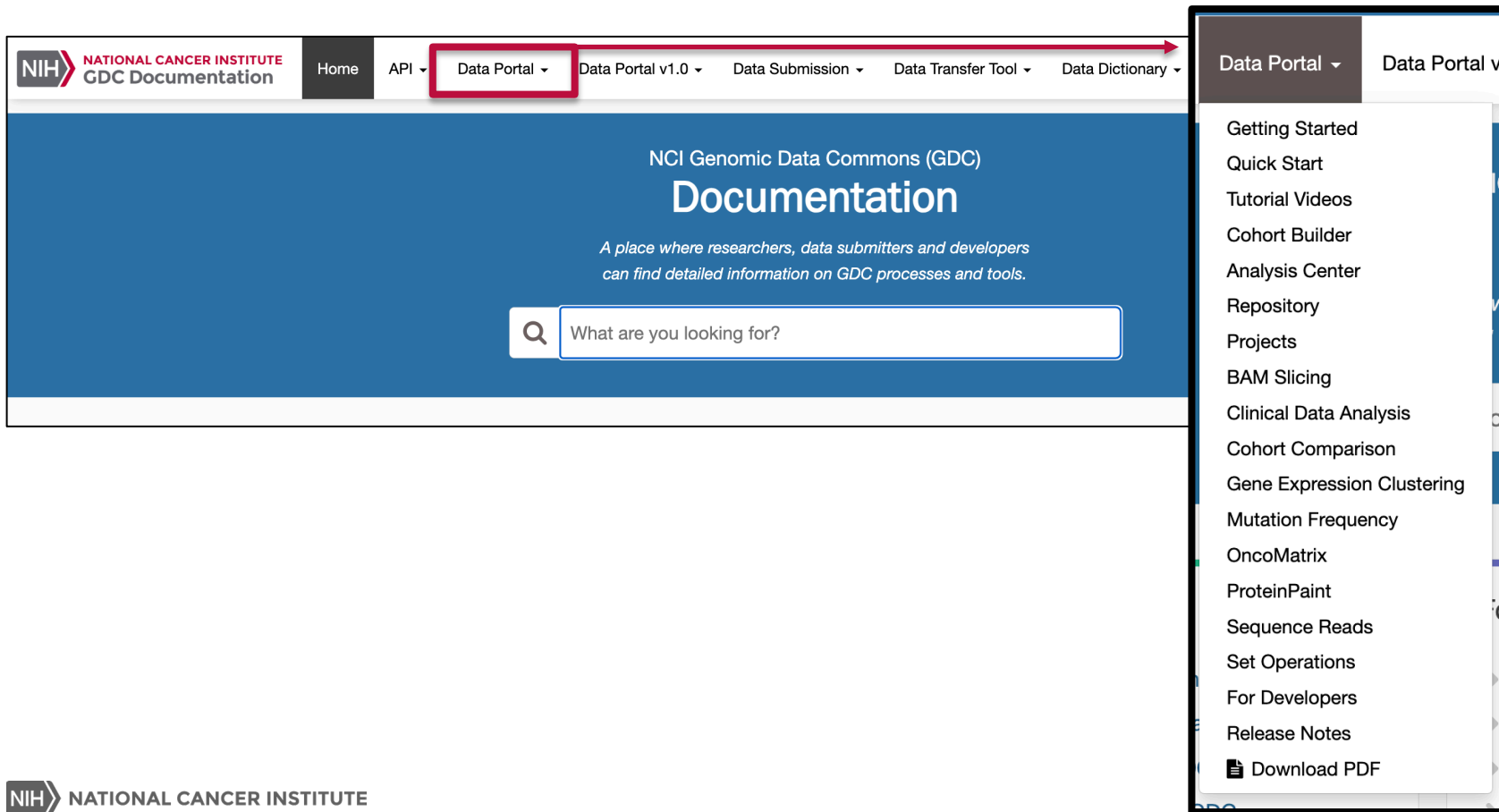
**Submit**



# Tutorials, Guides, and Support

*GDC 2.0*

# Users Guides for GDC 2.0 – <https://docs.gdc.cancer.gov>



NIH NATIONAL CANCER INSTITUTE GDC Documentation

Home API Data Portal Data Portal v1.0 Data Submission Data Transfer Tool Data Dictionary

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?

- Data Portal
- Getting Started
- Quick Start
- Tutorial Videos
- Cohort Builder
- Analysis Center
- Repository
- Projects
- BAM Slicing
- Clinical Data Analysis
- Cohort Comparison
- Gene Expression Clustering
- Mutation Frequency
- OncoMatrix
- ProteinPaint
- Sequence Reads
- Set Operations
- For Developers
- Release Notes
- Download PDF

NIH NATIONAL CANCER INSTITUTE

# Tutorial Videos for GDC 2.0

The screenshot shows the top navigation bar of the National Cancer Institute GDC Data Portal. The logo on the left reads "NIH NATIONAL CANCER INSTITUTE GDC Data Portal". A red box highlights the "Video Guides" link, which includes a play button icon. Other navigation options include "Send Feedback" (with a speech bubble icon), "Browse Annotations" (with a pencil icon), and "Manage Sets" (with a list icon). Below the navigation bar are four main menu items: "Analysis Center" (with a chart icon), "Projects" (with a group of people icon), "Cohort Builder" (with a person and globe icon), and "Repository" (with a database icon). A search bar on the right contains the text "e.g. BR". The main banner area features the text "Genomic Data Commons Data Portal" in large blue font. To the right of the text is an illustration of two human figures, one with internal organs highlighted in various colors. The text "Bone Ma" is partially visible at the bottom right of the banner.

# Questions or Feedback

NIH NATIONAL CANCER INSTITUTE  
GDC Data Portal

Video Guides

Send Feedback

Browse Annotations

Manage Sets

Analysis Center

Projects

Cohort Builder

Repository

Search: e.g. BR

## Genomic Data Commons Data Portal

Bone Ma

- Feedback is welcome and encouraged!
- Send to **support@nci-gdc.datacommons.io**

*Questions?*

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

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

1-800-4-CANCER

Produced February 2024