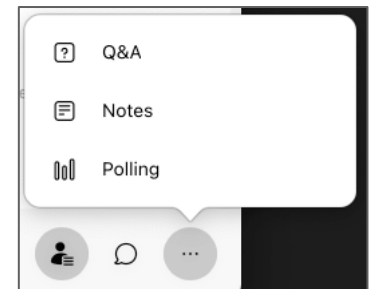# Integrating Analysis Tools with GDC 2.0: Harnessing the Power of the GDC SDK

16 September 2024

Bill Wysocki, Ph.D, Craig Barnes, Ph.D, Phil Tom – GDC Technical Leads
Center for Translational Data Science
University of Chicago

NIH ❯ NATIONAL CANCER INSTITUTE

Agenda

1. *Overview of GDC 2.0*

2. *The GDC SDK & API*

3. *GDC Analysis Tool Challenge*

4. *Resources for Support*

5. *Q&A*

Overview of GDC 2.0

Cohorts and Analysis Tools

GDC 2.0 Workflow

Step 1: Build and Save Cohort in Cohort-Builder

Step 2: Choose Analysis Tool in Analysis Center

Step 3: Cohort Applies to Analysis Tools Automatically
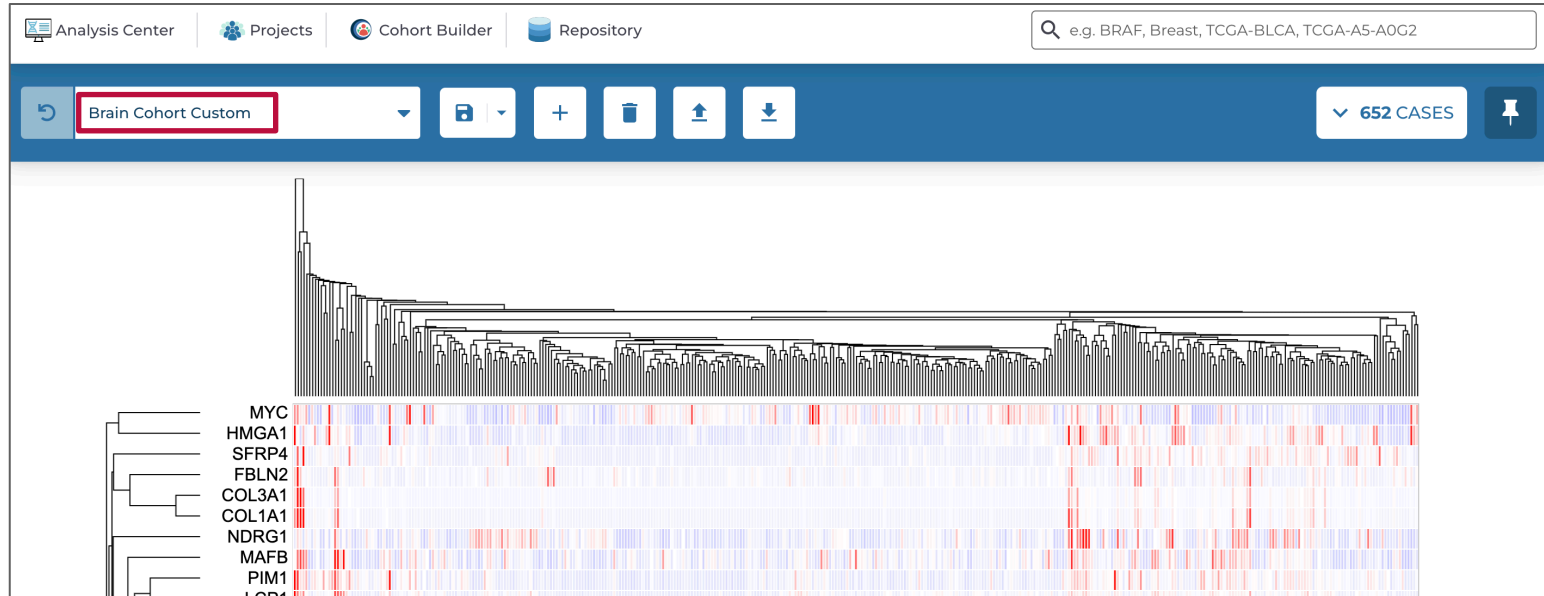
The GDC SDK and API

Development tools

GDC 2.0 Analysis Tool Software Development Kit (SDK)

GDC 2.0 Data Portal

UI Components

Core Context

Cohorts	Hooks
User Auth	Selectors
Apps	Actions

GDC API
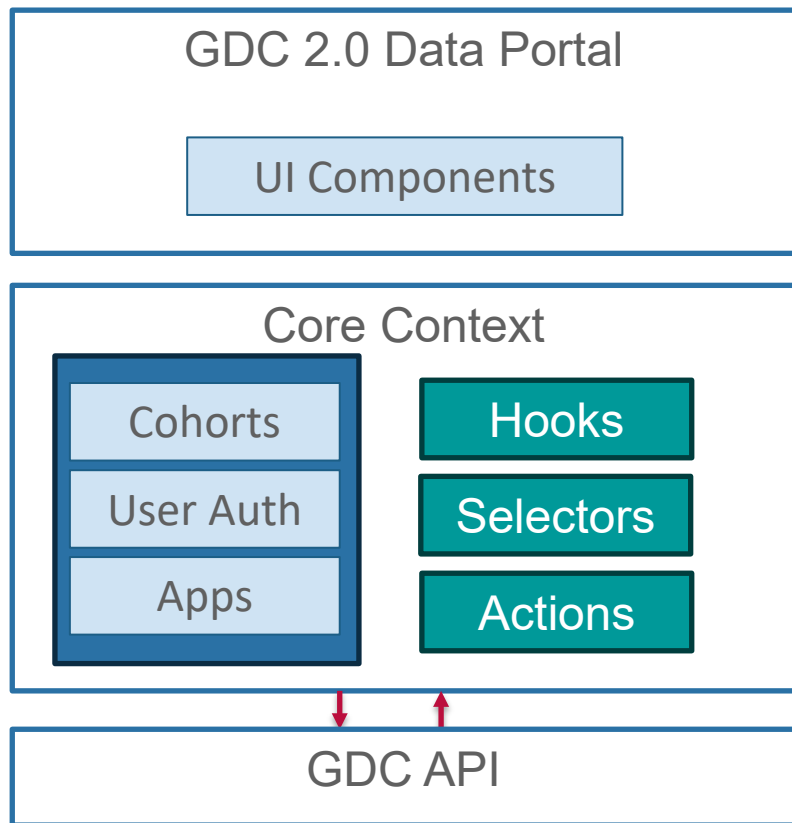
- The GDC 2.0 Portal is built on top of the **React** framework and uses the **Redux** library for state management

- UI Components are built using the React **Mantine** library, which provides a responsive design. Apps can use GDC UI Components

- UI Components and Apps interact with the GDC APIs via the SDK

  - Apps can access the core context and their own context using the SDK

  - Apps use *Hooks* to query for cohort information and fetch data from the GDC API

  - *Selectors* are used to access the state of the Portal's Redux store and get updates when a cohort changes

  - *Actions* are used to update the state of the App such as adding, removing, or clearing filters

App Structure

- Apps are higher-order components (HOC) that are rendered in the Analysis Center

- Each App handles a specific task and can be used to refine and analyze cohorts

- Local and Cohort filters are available to applications

  - Local filters are filters available from the GDC API that are specific to the application and are used to refine the data

  - Cohort filters are filters used to retrieve the active cohort, alter cohorts, or create new cohorts via the Cohort API

- Apps use *Hooks*, *Selectors*, and *Actions* for querying the GDC API and accessing and updating state, respectively

- A Component Library is provided for component re-use

Cohort Filters

Application

Local Filters

UI Elements:

State/Store:

Actions:

Alter Cohort   New Cohort
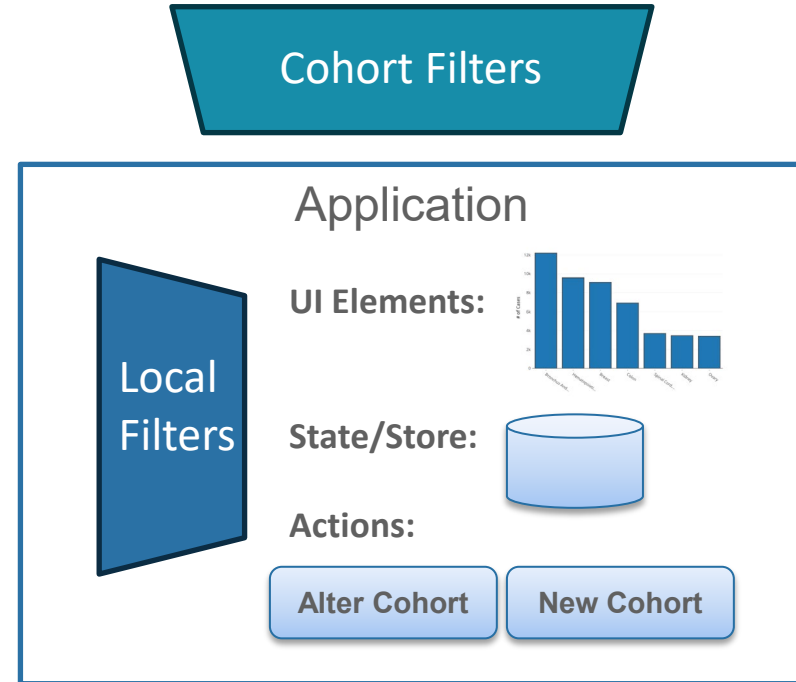
Hooks, Selectors, & Actions

REST/GraphQL
- *useGetGenesQuery*
- *useGetCasesQuery*
- *useGetSsmsQuery*
- *…*

Cohorts
- *selectCurrentCohort*
- *selectCurrentCohortFilters*
- *selectCurrentCohortCaseCount*
- *updateCohortFilter*
- *removeCohortFilter*
- *…*

Filters
- *useEnumFacets*
- *selectRangeFacets*
- *fetchFacetContinuousAggregation*
- *fetchEnumFacets*
- *…*

Projects, Files, Cart
- *useGetProjectsQuery*
- *useGetFilesQuery*
- *addFilesToCart*
- *removeFilesFromCart*
- *selectCart*
- *…*

Genomics
- *useGenesSummaryData*
- *selectGenesSummaryData*
- *useSSMS*
- *selectSsmsSummaryData*
- *…*

Cases
- *useCaseSummary*
- *selectCaseSummaryData*
- *useAllCases*
- *…*

GDC Component Library

https://github.com/NCI-GDC/gdc-frontend-framework

Buttons

Vertical Tables

Facets

Modals

Charts

- `EnumFacet` - a facet that is used to filter on an enum field
- `DateFacet` - a facet that is used to filter a date field
- `NumericRangeFacet` - a facet that is used to filter on a range field
- `PercentileFacet` - a facet that is used to filter on a percentile field
- `AgeRangeFacet` - a facet that is used to filter on an age range field
- `TextFacet` - a facet that is used to filter on a text field
- `BooleanFacet` - a facet that is used to filter on a boolean field

Developing an GDC Analysis Tool

- Clone GDC Portal repository

- Create a new directory in portal-proto/src/features.

- Incorporate hook/selectors/actions from gff/core

- Develop application specific hooks/selectors, app state, and UI components

- Add tests, documentation, and demo

- Register application by adding to the GDC app config:

    - Icon

    - Description

    - Cohort data count hook (if needed)

    - Metadata

Example: Mutation Frequency

Features:

- Local filtering of current cohort

- Interfaces to GDC API via SDK

- Add gene/ssm to current cohort

- Create new cohorts and sets

- Access Gene/SSM Summaries
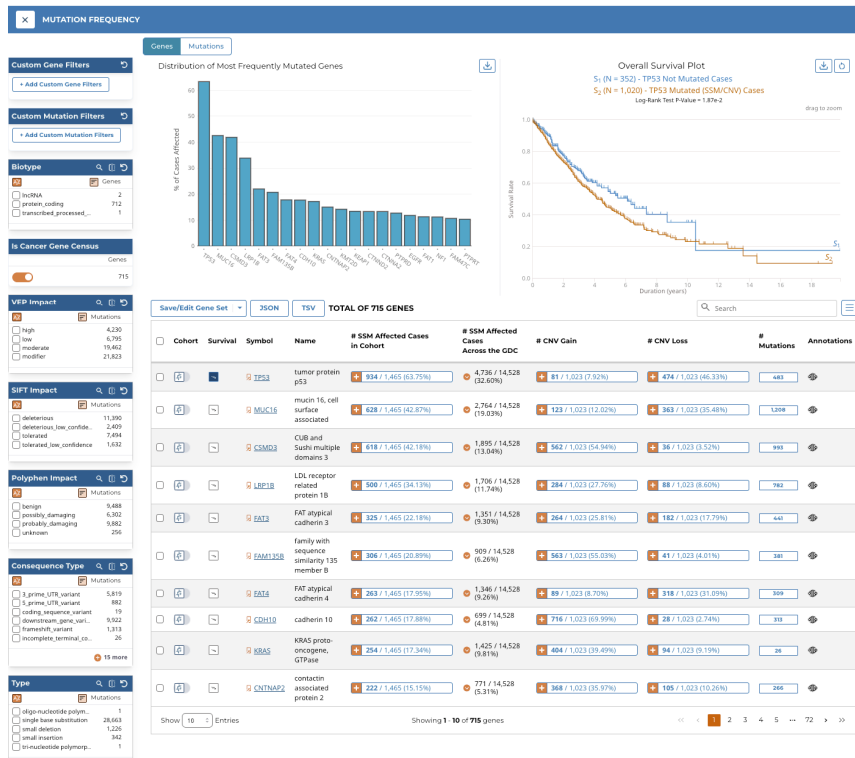
Mutation Frequency File Structure

```
∨ genomic
  > tests
  TS appApi.ts
  TS appFilters.ts
  TS Charts.tsx
  TS FilterPanel.tsx
  {} filters.json
  JS gene_mutation_facets.js
  TS geneAndSSMFiltersSlice.ts
  {} genes.json
  TS GenesAndMutationFrequencyAnalysisTool.tsx
  TS GenesPanel.tsx
  TS Genomic.tsx
  TS hooks.ts
  {} mutations.json
  TS registerApp.ts
  TS SSMSPanel.tsx
  TS types.ts
  TS utils.ts
```

Application state

Local filtering

Fetch from GDC APIs

Main app component

Application scoped hooks/slices

Application registration

Application local context/store

- If shared state information is needed this will require use of an Application State

- Enable creation of custom hooks based on redux-toolkit for custom data fetches

- SDK provides a functions to create application scoped redux objects

```
const persistConfig = {
  key: "MutationFrequency",
  version: 1,
  storage,
  whitelist: ["filters"],
};

const reducers = combineReducers({
  filters: geneFrequencyFiltersReducer,
});

export const { id, AppStore, AppContext,
useAppSelector, useAppDispatch } =
  createAppStore({
    reducers: persistReducer(persistConfig,
reducers),
    name: "MutationFrequency",
    version: "0.0.1",
  });

export type AppState = ReturnType<typeof reducers>;
```

Application registration

- Apps need to be registered with the GDC portal in order to be launched from the analysis center

- SDK provides a call to register the app using the root component for the application

```
import { createGdcAppWithOwnStore } from "@gff/core";
import { AppContext, AppStore, id } from
"@/features/genomic/appApi";
import GenesAndMutationFrequencyAnalysisTool from
"@/features/genomic/GenesAndMutationFrequencyAnalysisTool"
;

export default createGdcAppWithOwnStore({
    App: GenesAndMutationFrequencyAnalysisTool,
    id: id,
    name: "Genes and MutationFrequency",
    version: "v1.0.0",
    requiredEntityTypes: ["gene","ssm"],
    store: AppStore,
    context: AppContext,
});

export const GenesAndMutationFrequencyAppId: string = id;
```

Anatomy of a Tool Card

Icon

Title

Applicable number of cases in current cohort

Run App

Run App Demo

Detailed description
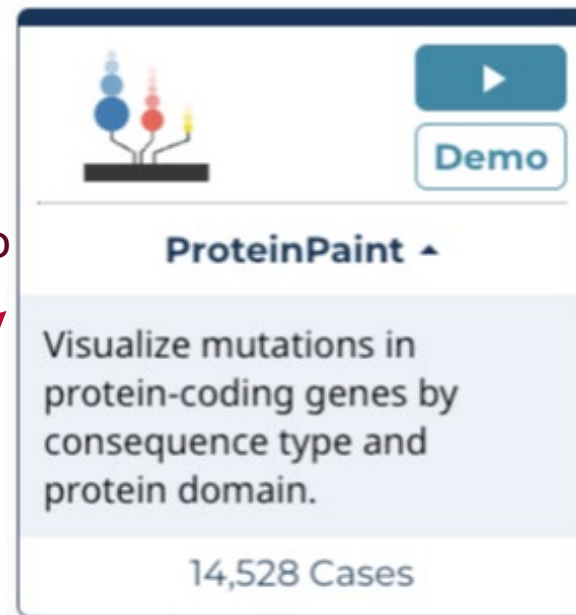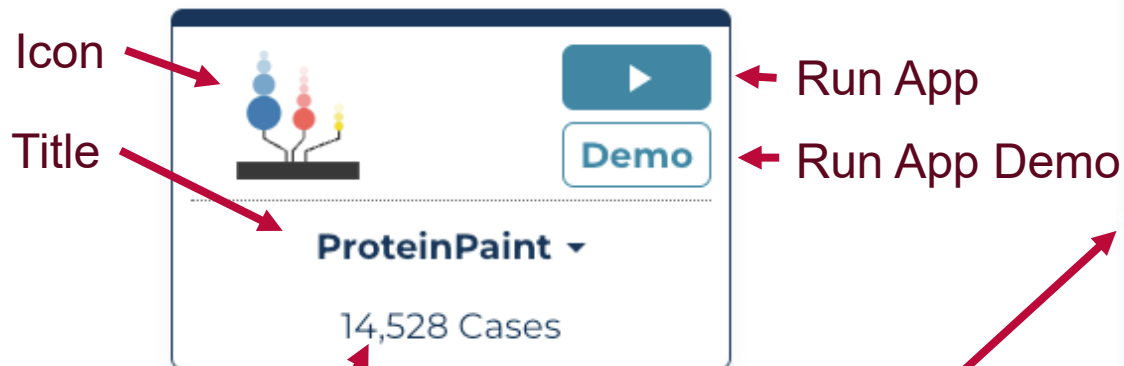
Tool Card registration

Register the tool card by adding an entry in:

packages/portal-proto/src/features/userflow/workflow/registeredApps.tsx

```
{

        name: "Mutation Frequency",
        icon: (<MutationFrequencyIcon role="img"
        aria-label="Mutation Frequency icon"
        />),
        tags: ["variantAnalysis", "ssm"],
        hasDemo: true,
        id:      "MutationFrequencyApp",
        countsField: "ssmCaseCount",
        description: "Visualize most frequently
        mutated genes and somatic mutations.",
        noDataTooltip:"Current cohort does not
        have SSM data available for
        visualization.",
}
```

Developer Resources

- [Developers Guide](#)

  - Overview, architecture, and application code walkthrough

- [Style Guide](#)

  - Guidelines for color, icons, typography

- API Documentation

  - Generated from the source code

  - Documents SDK functions and not smaller utility functions

- Code has a number of application to use as examples

What is the GDC API?

The GDC API is how we make harmonized clinical and genomic data available.

45,000 Cases

- 650 clinical properties
  - Demographics
  - Diagnosis
  - Treatments
- 200 biospecimen properties
  - Tissue Type (Tumor, Normal, etc.)
  - Tumor Descriptor (Primary, Metastatic, etc.)

1,000,000 Files

- 350 properties
- Experimental Strategy
  - WGS
  - WXS
- Data Category
  - Sequencing Reads
  - Simple Nucleotide Variation
  - Copy Number Variation

Getting Started with the GDC API

Documentation: https://docs.gdc.cancer.gov/API/Users_Guide/Getting_Started/

Basics

- Base URL: https://api.gdc.cancer.gov

- GraphQL endpoint

- RESTful endpoints

- Open-access data is available anonymously

- Controlled-access data requires authorization

Features

- Search and retrieval

- Analysis

- Download

Search and Retrieval

Query by any of the properties, and get properties you want.

- 45,000 cases

- 1,000,000 files

- 1,000 properties

Query Language

- Works with both GraphQL and RESTful endpoints

- Allows caller to define the search criteria

GraphQL

- `/graphql`

RESTful

- `/cases`

- `/files`

- `/projects`

Filter Language

Syntax

```
{
  "op": <OPERATOR>,
  "content": {
    "field": <FIELD NAME>,
    "value": <VALUE>
  }
}
```

Operators

- =, !=
- >, >=, <, <=

Example

```
{
  "op": "=",
  "content": {
    "field": "demographic.gender",
    "value": "female"
  }
}
```

Data Analysis

Somatic Mutations

- `/ssms`
  - 2,900,000 ssms
- `/ssm_occurrences`
  - 3,300,000 occurrences

Copy Number Variations

- `/cnvs`
  - 38,000 cnvs
- `/cnv_occurrences`
  - 75,000,000 occurrences

Genes

- `/genes`

Gene Expression

- `/gene_expression/availability`
- `/gene_expression/gene_selection`
- `/gene_expression/values`

Top Mutated Cases

- `/analysis/top_mutated_cases_by_gene`

Survival Analysis

- `/analysis/survival`

File Download

RESTful

- `/data`

Params

- One or more files ids

Single file download

- One file id is requested
- The response is an octet-stream of the file.

Multi-file download

- Multiple file ids are requested
- The response is a tarball of the files.

Example: Single file analysis

- Some files are small enough to fit in a browser
- Seurat analysis for scRNA-Seq

Example: BAM slicing

- `/slicing`
- Single file download
- Single part range request
  Range: bytes=1234-5678

GDC API Documentation

More information:

https://docs.gdc.cancer.gov/API/Users_Guide/Getting_Started/

GDC Analysis Tool Challenge

Add your tool to the GDC!

NIH NATIONAL CANCER INSTITUTE

Challenge Goal

Provide the research community with a novel analysis tool to analyze data in the GDC in support of cancer research

Analysis Tool Requirements

- ✓ The Analysis Tool shall demonstrate **scientific need and innovation** beyond the existing GDC Analysis Tools made available in the *Analysis Center*

- ✓ The Analysis Tool must **use GDC data** and be able to **integrate with the GDC** using the GDC Analysis Tool Software Development Kit

- ✓ The Analysis Tool must be **open-source**

Mutation Signature Analysis
- Tools to identify and interpret mutation signatures in cancer genomes, helping to understand the underlying mutational processes

Copy Number Variation (CNV) Analysis
- Tools for detecting and analyzing copy number variations in cancer genomes to identify amplifications or deletions

Structural Variant Analysis
- Tools for detecting and characterizing structural variations such as translocations, inversions, and duplications in cancer genomes

RNA Expression Analysis
- Tools for performing differential expression analysis, alternative splicing, alternative isoform analysis, and single cell RNA-Seq analysis

Regulatory Element Exploration
- Tools for exploring RNA expression profiles for gene regulatory elements, miRNA analysis, methylation data visualization, and functional enrichment

Pathway & Network Analysis
- Tools that analyze the impact of genetic alterations on biological pathways and gene networks relevant to cancer

Phylogenetic Analysis
- Tools to reconstruct the evolutionary history of cancer cells, helping researchers understand tumor heterogeneity and evolution

Integration of Multi-Omics Data
- Tools that integrate genomics data with other omics data (e.g., transcriptomics, proteomics) to provide a comprehensive view of cancer biology

Longitudinal Data Visualization
- Tools that plot clinical, diagnostic images, and genomic data on a time continuum, patient treatments and outcomes over time, and variation and expression plots in relation to time of other events

Machine Learning for Slide Images
- Machine Learning algorithms that use tissue or diagnostic slide images for predictive analysis (e.g., cancer prognosis, genomic correlations such as spatial transcriptomics)

Analysis Tool Data

GDC Data Types	GDC Data Subtypes
Whole Exome and Targeted Sequencing	• Alignments (BAM) • Raw somatic variants (variant caller level; VCF) • Annotated somatic variants (variant caller level; VCF + MAF) • Ensembled somatic variants (all callers; MAF) • Masked somatic variants (all callers; MAF)
Whole Genome Sequencing	• Alignments (BAM) • Raw somatic variants (indels + point mutations; VCF) • Copy number segments (TXT) • Copy number estimates (TSV) • Structural variants (BEDPE + VCF)
RNA-Seq	• Alignments (genomic, transcriptome, chimeric; BAM) • Augmented gene expression counts (counts + normalization; TSV) • Splice junctions (TSV) • Transcript fusions (BEDPE + TSV)
Single cell RNA-Seq	• Alignments (BAM) • Counts (raw + filtered; MEX) • Differential gene expression (TSV) • Single cell analysis (PCA, tSNE, UMAP; TSV) • Full analysis loom file (HDF5)
miRNA-Seq	• Alignments (BAM) • miRNA expression (TXT) • Isoform expression (TXT)
Methylation Array	• Masked methylation array (IDAT) • Methylation beta values (TXT)

GDC Data Types	GDC Data Subtypes
ATAC-Seq	• Alignments (BAM)
SNP6 Genotyping Array	• Copy number segment (TXT) • Copy number estimate (TXT/TSV) • In addition, SNP6 raw array intensities (CEL) files and SNP6 birdseed genotype files (TXT) are not processed, but made available to researchers as is.
Genomic Profile	• Raw somatic variants (variant caller level; VCF) • Annotated somatic variants (variant caller level; VCF + MAF) • Copy number estimates (TXT) • Structural variants (BEDPE)
RPPA	• Protein expression quantification (RPPA) (TSV) are not processed but provided to researchers as is
Clinical data	• Clinical supplement (XML, JSON, TSV, Biotab, XLSX) • Pathology report (PDF) • Clinical data as GDC node properties
Biospecimen data	• Biospecimen supplement (XML, JSON, TSV, Biotab, XLSX) • Biospecimen data as GDC node properties
Imaging data	• Tissue slides (SVS) • Diagnostic slides (SVS)

Whole Exome & Targeted Sequencing Analysis Tool Data

Whole Exome and Targeted Sequencing	- Alignments (BAM) - Raw somatic variants (variant caller level; VCF) - Annotated somatic variants (variant caller level; VCF + MAF) - Ensembled somatic variants (all callers; MAF) - Masked somatic variants (all callers; MAF)

RNA-Seq Analysis Tool Data

RNA-Seq	• Alignments (genomic, transcriptome, chimeric; BAM) • Augmented gene expression counts (counts + normalization; TSV) • Splice junctions (TSV) • Transcript fusions (BEDPE + TSV)

Analysis Tool Evaluation

- Analysis tools will be evaluated based on the following criteria:

  ✓ Scientific need

  ✓ Innovation

  ✓ Ability to integrate with the GDC Data Portal and use of GDC data

  ✓ Requirements compliance (e.g., open source)

Winning Tool Integration with the GDC

- Winning tools will be integrated with the GDC Data Portal and will be made available in the [Analysis Center](#)

Challenge Timeline

- **Phase 1: Challenge Registration**
  - 9/1/2024 – 10/31/2024
- **Phase 2: Finalists Tool Integration**
  - 1/1/2025 - 6/30/2025
- **Tool Awards & Recognition**
  - 10/1/2025 - 12/31/2025
- **How to Enter:**
  - https://www.challenge.gov/?challenge=nci-gdc-analysis-tool-challenge&tab=how%20to%20enter

NIH — NATIONAL CANCER INSTITUTE
Genomic Data Commons

ANALYSIS TOOL CHALLENGE

#NCIGDC
#GDCChallenge
#GDCAnalysisToolChallenge

Register Now!
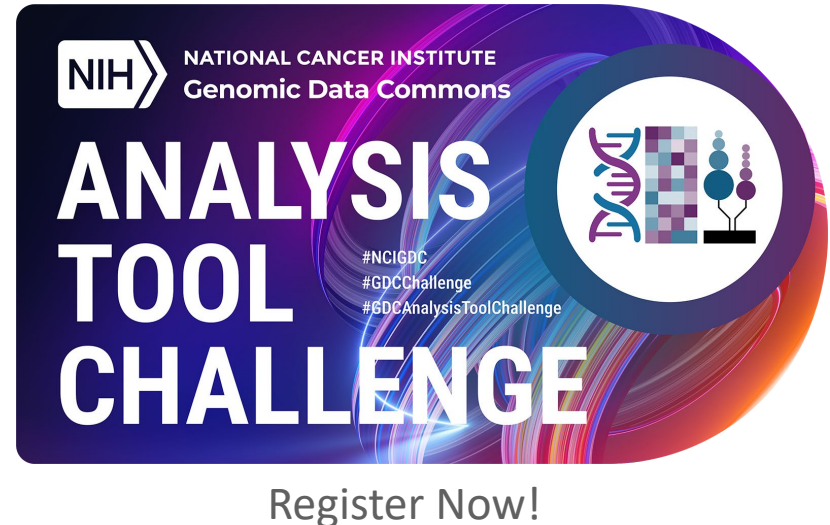
Resources for Support

Users Guides, Videos, and More!

Analysis Tool Integration

- GDC resources are available for analysis tool integration:

  - GDC Developer's Guide - Provides detailed instructions on using the GDC Analysis Tool Software Development Kit (SDK)

  - GDC Frontend Framework GitHub Repository - Allows developers to build a local instance of the GDC 2.0 Data Portal to test tool integration

  - ProteinPaint GitHub Repository - An example of tool integration

  - GDC API User's Guide - Provides examples of how to retrieve and analyze data from the GDC

  - GDC Component Library including a Style Guide - Provides guidance for UI components

GDC Developer's Documentation: docs.gdc.cancer.gov

NIH⟩ **NATIONAL CANCER INSTITUTE** **GDC Documentation**

Home   API   **Data Portal**   Data Submission   Data Transfer Tool   Data Dictionary   Data   Encyclopedia

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

GDC Analysis Tool Software Development Kit (SDK)

This guide will detail the process of developing applications for the GDC Data Portal 2.0. It describes the structure of the GDC Data Portal, how to use the GDC Data Portal API, and how to develop applications for the GDC Data Portal.

The GDC Data Portal is a repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy. The GDC Data Portal supports the development of applications that allow for analysis, visualization, and refinement of cohorts.

The GDC Data Portal is built on top of the GDC API, which provides access to the GDC data. The GDC Data Portal provides an Analysis Tool Framework (ATF) for developing applications that can be used to analyze, visualize, and download data from the GDC.

The GDC Data Portal is built with the React framework and the Redux library for state management. The GDC Data Portal uses NextJS as its application framework which provides server-side rendering of React components. Mantine.dev is the component library, and styling is through TailwindCSS. The GDC Data Portal is built on top of the GDC API, which provides access to the GDC data.

GDC API Search & Retrieval: docs.gdc.cancer.gov

Home | **API** | Data Portal | Data Submission | Data Transfer Tool | Data Dictionary | Data | Encyclopedia

API
Getting Started
Search and Retrieval
Downloading Files
Data Analysis
BAM Slicing
Submission
Python Examples
GraphQL Examples
System Information
Additional Examples
Appendix A: Available Fields
Appendix B: Key Terms
Appendix C: Format of Submission Queries and Responses
Release Notes
Download PDF

Search and Retrieval

Introducing Search and Retrieval Requests

The GDC API provides endpoints that search and retrieve information stored in the GDC according to the GDC Data Model. The general format of requests to search & retrieval endpoints is described below.

> **Note:** Queries described in this section work for datasets that have been released to the GDC Data Portal. Unreleased data that is in the process of being submitted to GDC cannot be queried using these methods. See Submission to learn how to query unreleased data using GraphQL.

Components of a Request

A typical search and retrieval API request specifies the following parameters:

- a `filters` parameter, that specifies the search terms for the query

- several parameters that specify the API response, such as:

GDC Website: gdc.cancer.gov

An official website of the United States government

NATIONAL CANCER INSTITUTE
Genomic Data Commons

Publications | Contact Us   Launch Data Portal   GDC Apps

Search

About the GDC ⌄   About the Data ⌄   Analyze Data ⌄   Access Data ⌄   Submit Data ⌄   For Developers ⌄   Support ⌃   News ⌄

Training

GDC Webinars

GDC Tutorial Videos

NCI GDC YouTube Playlist ⧉

Documentation

GDC Data Portal User's Guide

GDC Data Transfer Tool User's Guide

GDC Data Submission Portal User's Guide

GDC API User's Guide

GDC Application Developer Guide

GDC FAQs

Get Support

Contact GDC Help Desk ✉

Join the GDC User Listserv

𝕏 Follow us on @NCIGDC_Updates ⧉

in Subscribe to NCI Precision Oncology ⧉

→ More about Analyzing Data

to therapy.

Questions?

U.S. Department of Health & Human Services

National Institutes of Health | National Cancer Institute

gdc.cancer.gov

1-800-4-CANCER   Produced September 2024