GDC Data Portal User's Guide

NCI Genomic Data Commons (GDC)

Contents

Chapter 1

Getting Started

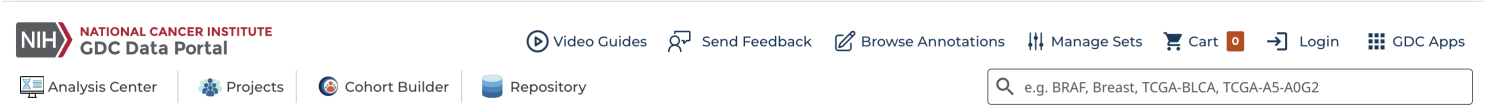
Getting Started with the GDC Data Portal

Accessing the GDC Data Portal

First, go to https://portal.gdc.cancer.gov/.

GDC Data Portal Header

The header of the GDC Data Portal contains frequently used links and features.



On the upper-left is the GDC Data Portal logo, which links to the home page of the GDC Data Portal. Below the logo are links in the following order:

- Analysis Center: the central hub for accessing all the tools in the GDC Data Portal
- Projects: allows exploration of all the projects within the GDC Data Portal
- Cohort Builder: the Cohort Builder tool consists of a variety of clinical and biospecimen filters for building custom cohorts for analysis
- Repository: allows exploration of files associated with a cohort

On the right are the following features:

- **Video Guides**: videos that demonstrate the various features of the Data Portal
- **Send Feedback**: send feedback to the GDC team
- **Browse Annotations**: the Annotations Browser, where the user can view and search for annotations that may provide additional context when analyzing GDC data
- **Manage Sets**: review gene and mutation sets that have been saved, upload new sets, and delete existing sets
- **Cart**: the data files that are ready for download
- **Login**: allows authentication for access to controlled access datasets. Once authentication is successful, the eRA Commons username will be displayed in place of the "Login" button. Clicking on the username will then allow users to see which projects they have access to and to download an authentication token and the API.
- **GDC Applications**: contains links to other GDC sites and applications
- **Search**: search for projects, cases, files, genes, mutations, and annotations within the GDC Data Portal

Cohorts

The GDC Data Portal 2.0 is a cohort-centric cancer research platform. Users can create custom cohorts based on specific projects, primary sites, disease types, or any combination of clinical, biospecimen, and molecular features. Custom cohorts can then be

used with various tools in the Analysis Center to perform further analysis. Files from custom cohorts can also be downloaded for further analysis with other research tools.

If the user does not already have a custom cohort when they are in the Analysis Center, a custom cohort ("Unsaved_Cohort") containing all cases in the GDC will be automatically created. This allows the user to explore the Analysis Center without first needing to create a cohort.

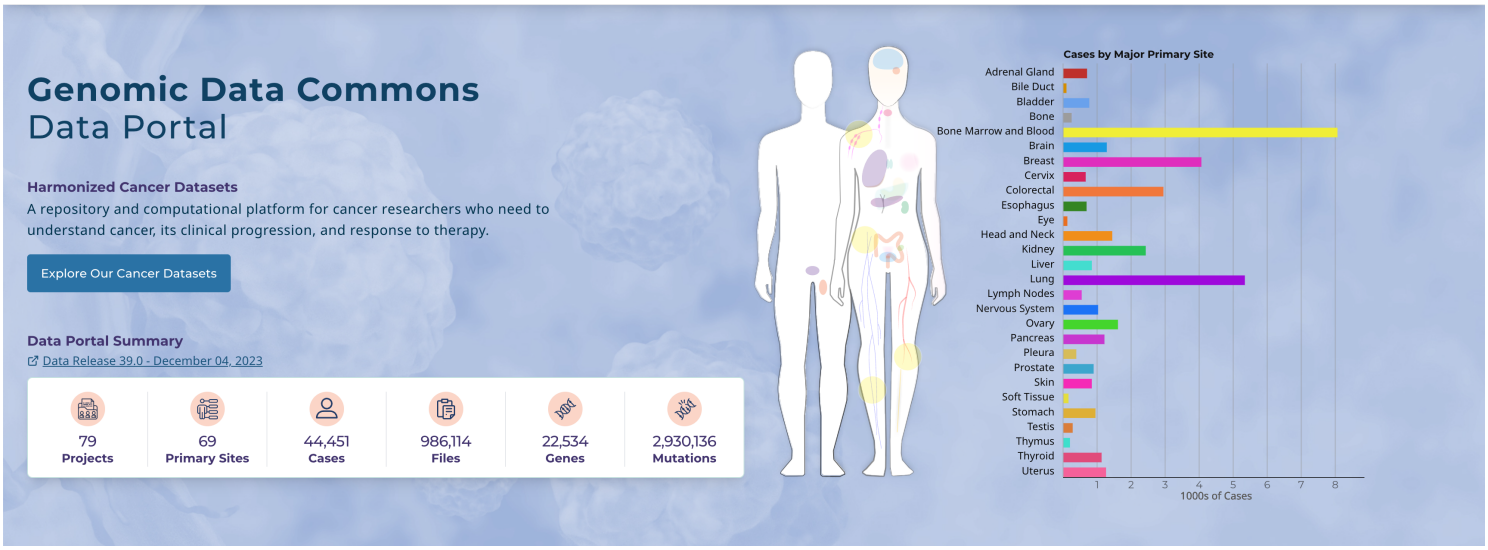
Additional cohorts can be created using the main toolbar in the Analysis Center. Cohorts can also be saved or deleted using the main toolbar. See the section below on the Analysis Center for more information on the main toolbar.

Unsaved cohorts are not retained once the browser tab is closed. Saved cohorts continue to be accessible as long as the same browser is used and should be available through data releases.

Home Page

The Analysis Center can be accessed by clicking on the "Explore Our Cancer Datasets" button on the left side of the home page.

On the right side of the home page are a human anatomical outline and a bar graph. Choosing a site on the outline or graph will lead the user to the Analysis Center and automatically create a custom cohort consisting of cases corresponding to that site.



Analysis Center

The Analysis Center can be accessed by clicking on the corresponding link in the GDC Data Portal header, on the "Explore Our Cancer Datasets" button on the home page, or on one of the sites in the human anatomical outline or bar graph.

The Analysis Center has the following sections:

- **Main Toolbar**: manage and create custom cohorts
- **Query Expressions**: displays the filters applied to the current cohort
- **Analysis Tools**: all analysis tools available are located in the Analysis Center as individual cards. When individual analysis tools are launched, they are displayed in this section of the Analysis Center.

Main Toolbar

By default, the main toolbar is always visible in the Analysis Center and Tools. Users can use the main toolbar to view information and perform a number of actions on their cohorts.

The name of the current cohort is displayed in a field on the left. Previously created cohorts can be accessed by choosing this field and selecting their names from the dropdown menu.

The main toolbar also contains a set of buttons that are used to manage or create new cohorts. To the left of the cohort name is the "Discard Changes" button, which discards unsaved changes that have been made to the current cohort.

To the right of the cohort name are the following buttons:

- **Save Cohort:** Two options for saving cohorts are available in the dropdown menu. Select the "Save" option to save the active cohort and any changes made to it. Select the "Save As" option to save the active cohort, along with any changes made to it, as a new cohort. Cohorts with unsaved changes have a yellow exclamation mark icon displayed next to their names. Custom cohorts that are saved should persist through releases and continue to be accessible if the same browser is used. When the GDC releases new data, saved cohorts will be updated to include the newly released cases matching the filters applied to the cohort. **It is recommended that users export and securely store any cohort that cannot be easily recreated in case the browser session is cleared.**
- **Create New Unsaved Cohort:** Adds a new unsaved cohort with all the cases in the GDC and changes the active cohort to this new cohort
- **Delete Cohort:** Deletes the current active cohort. This action cannot be undone.
- **Import New Cohort:** Imports a set of cases as a cohort. These can be imported as a plain text list of UUIDs or submitter_ids (barcodes).
- **Export Cohort:** Exports the active cohort to a file. A cohort will be exported as a list of UUIDs. Exporting a cohort allows users to obtain a static list of the cases which are in their cohort at the time of export.

Two other buttons are located on the far right of the toolbar:

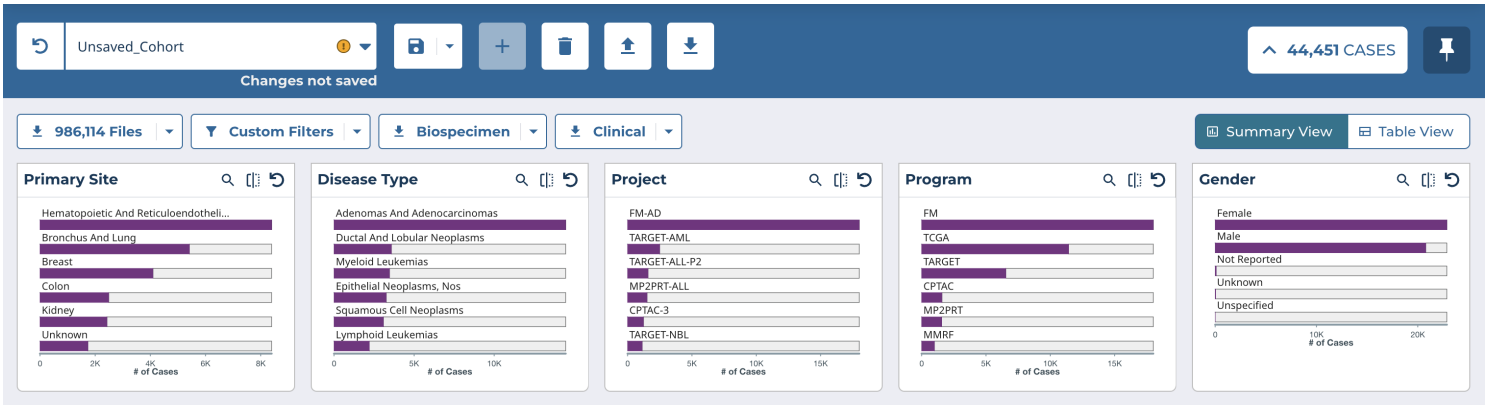
- **Expand/Collapse:** Displays the number of cases associated with the current cohort. Displays summary charts of the current cohort, as well as a table of the cases in the current cohort. The Summary View and Table View buttons can be used to toggle between a display of the summary charts and the table.
- **Pin/Unpin Cohort Bar:** Toggles between pinning the main toolbar to the top of the Analysis Center so that it is always in view, and unpinning it from the top of the Analysis Center.

Cohort Summary Charts

Cohort summary charts display graphics that show the number of cases with each value of a set of commonly used properties. The following buttons are available at the top of the summary charts:

- **Files:** Displays the number of files associated with the active cohort and provides the ability to add these files to the cart, download a manifest, or download metadata associated with the files
- **Custom Filters:** Allows for the cohort to be filtered by a custom set of cases, mutations, or genes
- **Biospecimen/Clinical:** Downloads the biospecimen/clinical metadata for the cohort in JSON or TSV format

In the default view, the number of cases for the top five most common values are displayed. Other values can be searched by choosing the magnifying glass button at the top right of each card.

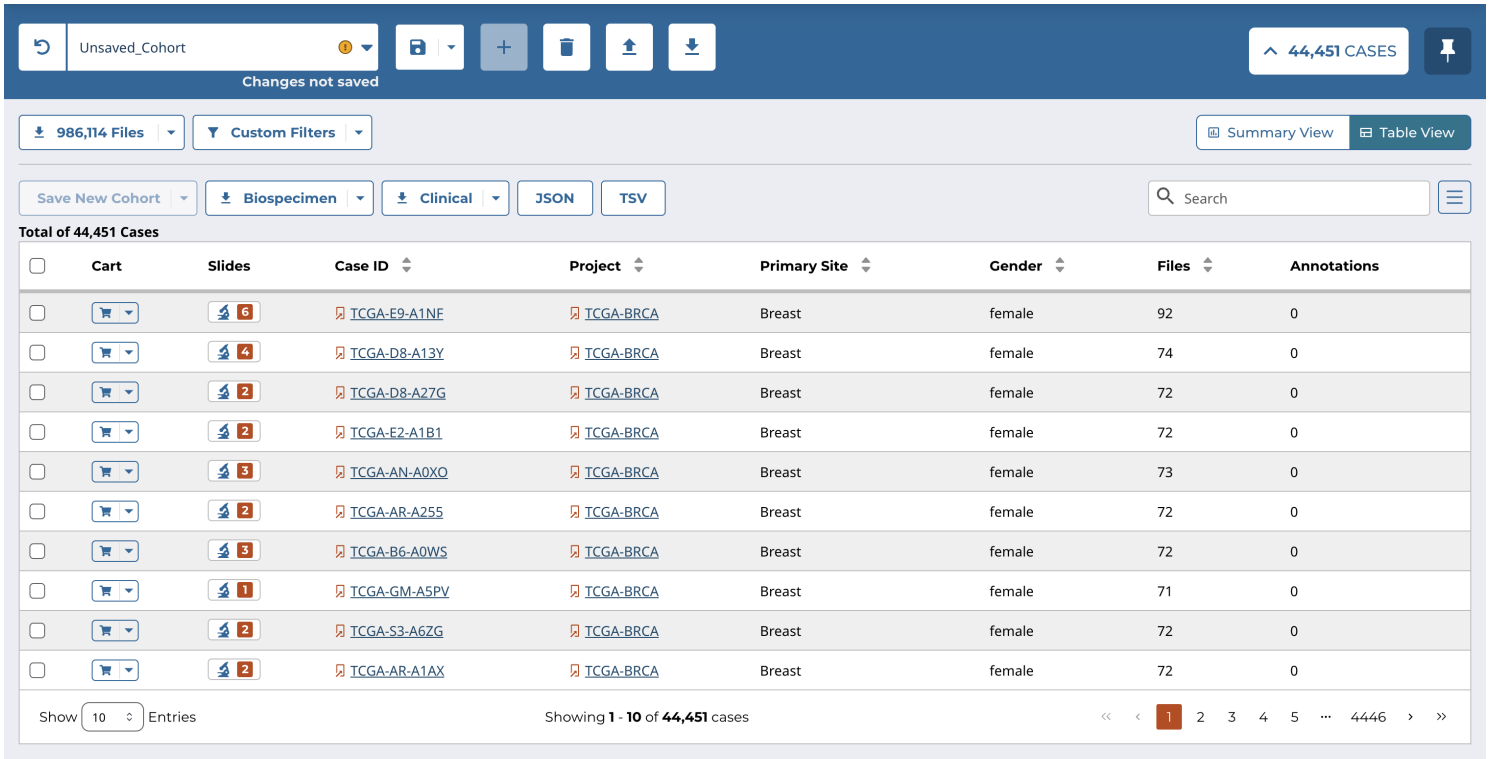


The middle button at the top right of each card displays the selection view. The selection view displays the same values as the default view as a number instead of a graph. In addition, each value can be selected to filter the active cohort for cases with a specific set of values.

Cohort Case Table

The case table displays a list of cases in the active cohort along with associated metadata. It also allows for each case to be selected with a checkbox for saving a new cohort or exporting metadata. All cases on the current page of the table can be selected

at the same time by using the checkbox in the header. Note that expanding the case table, with the cohort bar pinned, may obscure the view of the analysis center or analysis tools.



The top right of the case table features a search function that can be used to query specific cases. The following buttons are available at the top left of the case table:

- **Save New Cohort:** Allows for a new cohort to be created based on the selected cases from the table. The new cohort can comprise only the selected cases, the selected cases added to the active cohort, or the selected cases subtracted from the active cohort.
- **Biospecimen/Clinical:** Downloads the biospecimen/clinical metadata for the cohort in JSON or TSV format. When no cases are selected, the metadata pertains to the entire cohort. When cases are selected, the metadata pertains to only the selected cases.
- **JSON/TSV:** Downloads the information in the case table in JSON or TSV format

The case summary panel can be collapsed by selecting the 'Collapse' button that replaces the 'Expand' button.

Case Summary Page

Users can launch the Case Summary Page by clicking a Case ID in the Cohort Case Table. The Case Summary Page displays case details including the project and disease information, data files that are available for that case, and the experimental strategies employed. A button in the top-left corner of the page allows the user to add all files associated with the case to the file cart.

CA TCGA-UCEC / TCGA-DI-A1NO

Add all files to the cart

TOTAL OF 70 FILES 7 ANNOTATIONS

SUMMARY

Case UUID	d3e93d76-6565-454a-a085-4492024350e9	**Disease Type**	Adenomas and Adenocarcinomas
Case ID	TCGA-DI-A1NO	**Program**	TCGA
Project	TCGA-UCEC	**Primary Site**	Corpus uteri
Project Name	Uterine Corpus Endometrial Carcinoma	**Images**	3

FILE COUNTS BY DATA CATEGORY

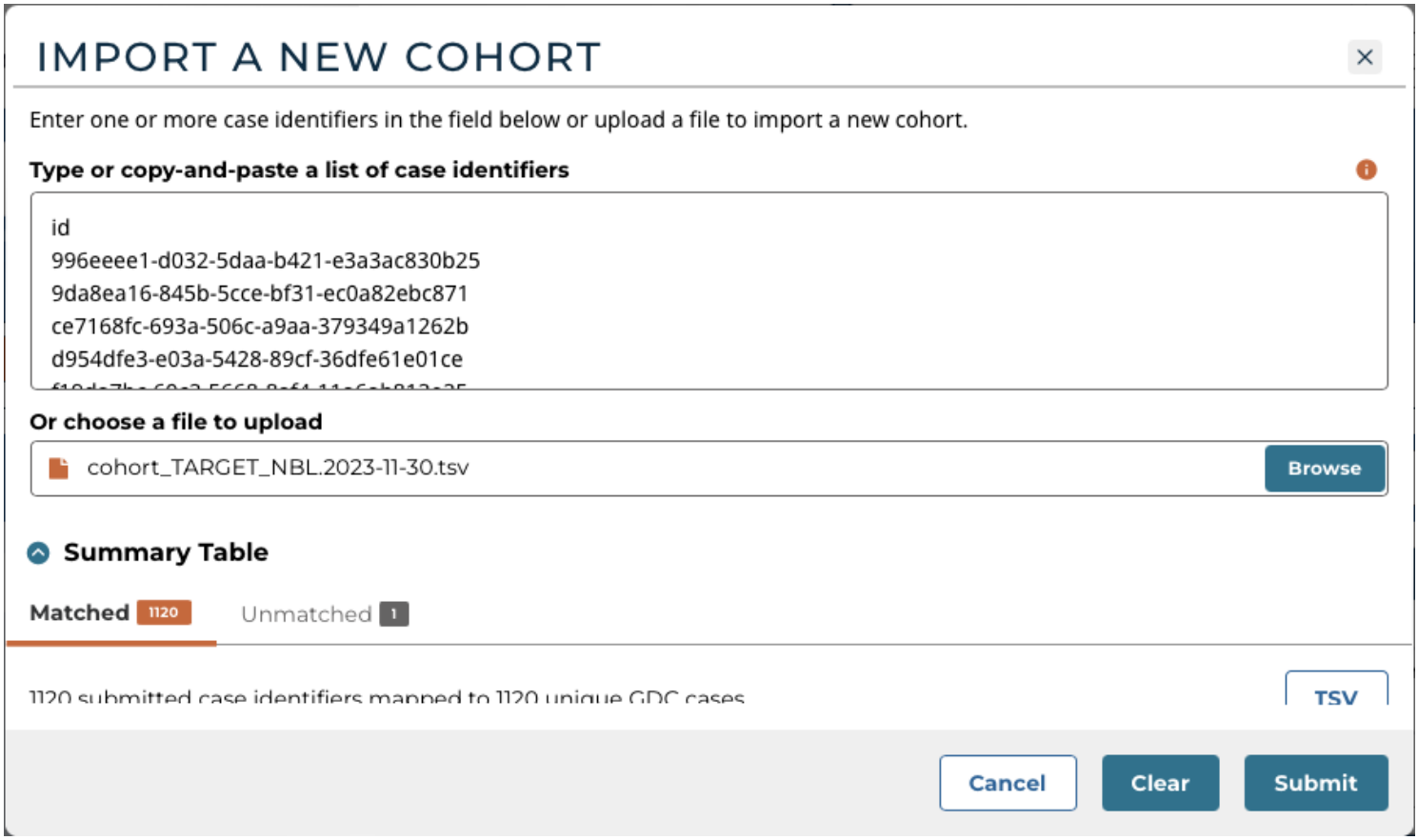
Data Category	**Files** (n=70)	
Biospecimen	17	24.29%
Clinical	11	15.71%
Copy Number Variation	11	15.71%
DNA Methylation	6	8.57%
Proteome Profiling	1	1.43%
Sequencing Reads	4	5.71%
Simple Nucleotide Variation	16	22.86%
Transcriptome Profiling	4	5.71%

FILE COUNTS BY EXPERIMENTAL STRATEGY

Experimental Strategy	**Files** (n=70)	
Diagnostic Slide	1	1.43%
Genotyping Array	13	18.57%
Methylation Array	6	8.57%
miRNA-Seq	3	4.29%
Reverse Phase Protein Array	1	1.43%
RNA-Seq	3	4.29%
Tissue Slide	2	2.86%
WXS	16	22.86%

Import New Cohort

The Import New Cohort button in the main toolbar allows for a set of cases to be imported. These can be entered directly into the text box as a plain text list of UUIDs or submitter_ids (barcodes) or imported as a CSV, TSV, or TXT file. Users can hover over the orange (i) to verify accepted case identifiers, delimiters, and file formats.



Clicking the `Submit` button will prompt users to name and save their new cohort, after which it will be made the active cohort.

Query Expressions

The query expressions section displays information about the filters applied to the current cohort and allows convenient operations to be performed on those filters.

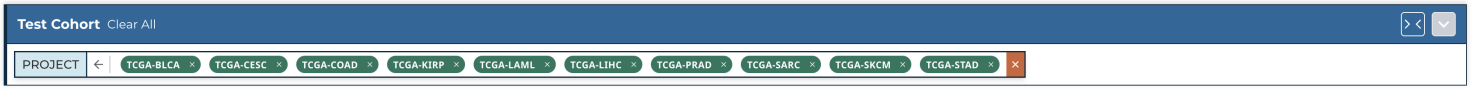
In the top-left corner of this section is the name of the current cohort. To its right is a "Clear All" option, which will remove all filtering applied on the current cohort.

On the top-right corner of this section are the following two buttons:

- **Collapse/Expand Selected Values**: by default, a full list of all the values that have been selected for each property is displayed. This button allows the user to switch from this default expanded view to a minimized view, which only displays the number of values selected for each property.
- **Collapse/Expand Filters Section**: by default, a maximum of three rows will be displayed at a time for the filters selected for the current cohort. This button allows the user to switch from displaying a maximum of three rows for the selected filters to displaying an unlimited number of rows. This button is only enabled if the display of the selected filters for the current cohort exceeds three rows.

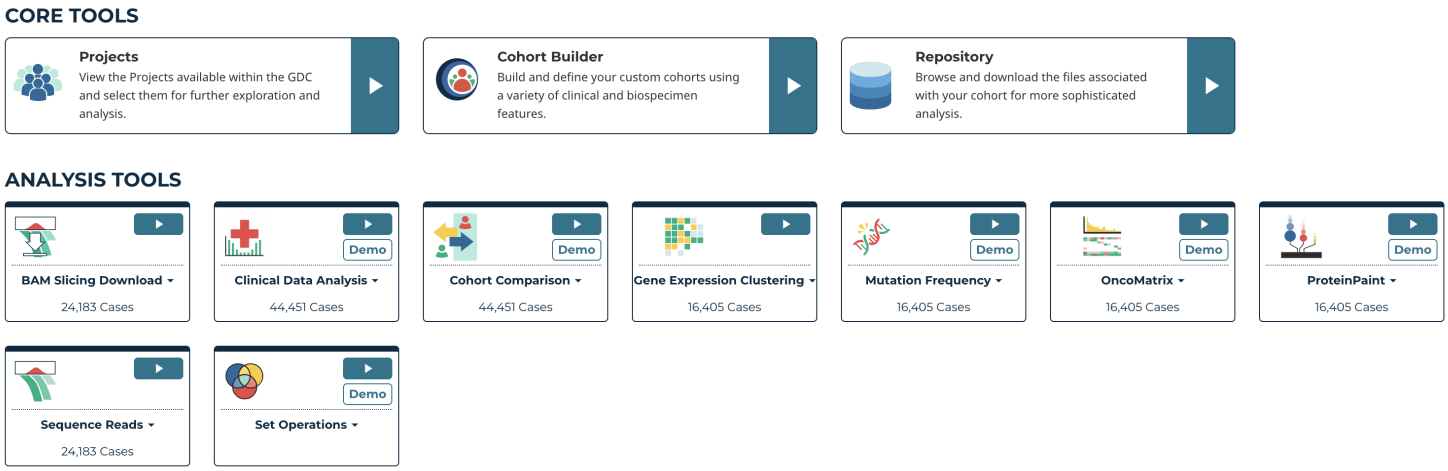
The main area of the query expressions section displays the filters applied to the active cohort. Individual values can be removed by clicking on them. Properties can be removed by clicking on the "X" to the extreme right of each property group of values.

If desired, selected values can be collapsed by clicking on the left arrow on the left of the values. When collapsed, values can be expanded again by clicking on the right arrow.



Analysis Center Tools

Available tools are displayed under the Query Expression section of the Analysis Center.

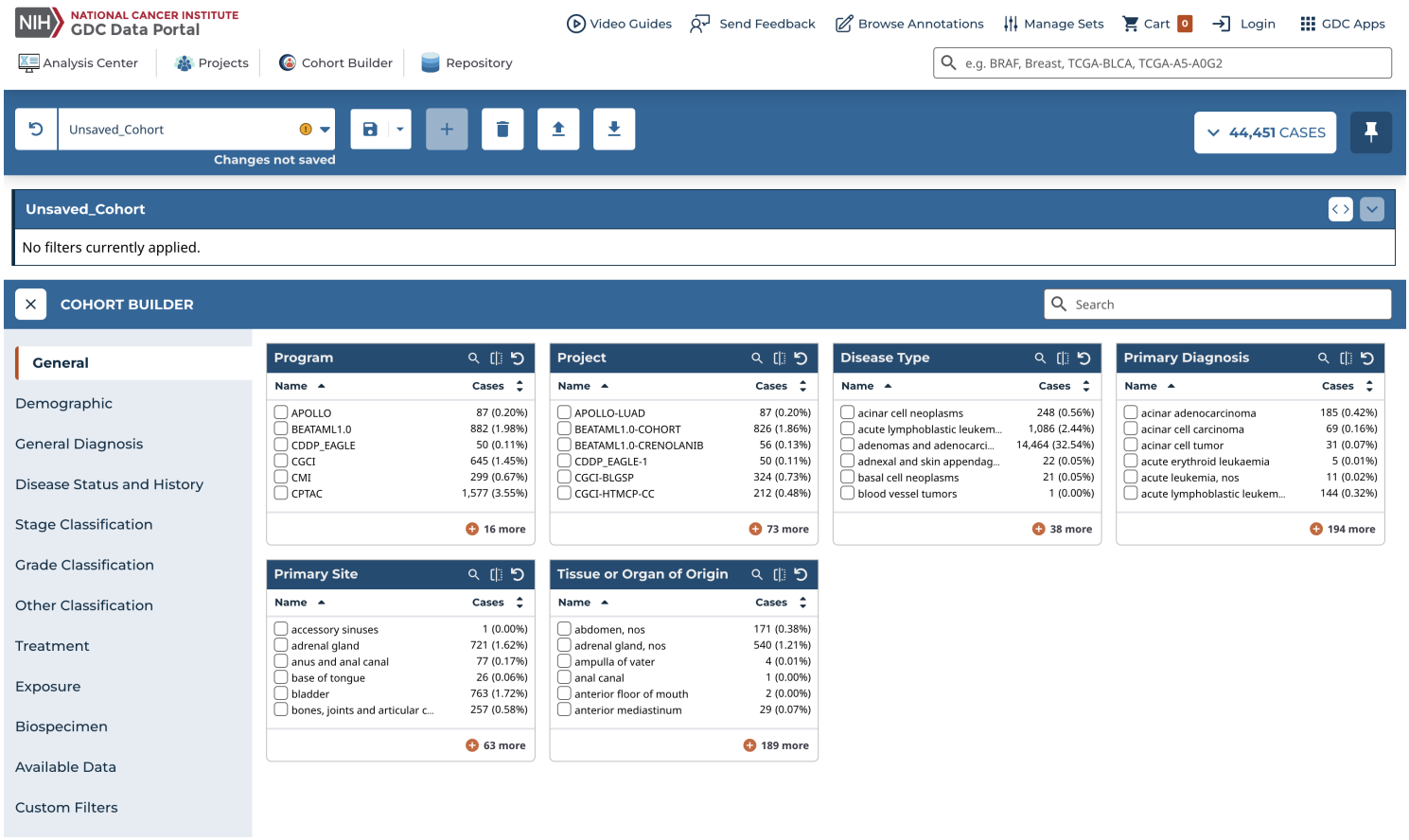


Each tool is showcased within a 'card', which can be launched using the teal 'Play' button.

Cohort Builder and Cohort Analysis

To build and analyze a cohort of interest using an analysis tool in the Analysis Center:

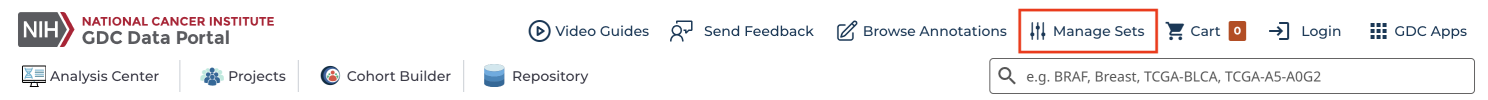
1. Choose the Cohort Builder icon on either the GDC Data Portal header, or click on the Cohort Builder card in the Analysis Center. The Cohort Builder will appear on the screen.
2. Create a custom cohort based on filters available in the Cohort Builder



3. Either choose the Analysis Center icon on the GDC Data Portal header, or click on the "X" on the left of the Cohort Builder header. All the tools in the Analysis Center will be displayed on the screen.
4. Choose an analysis tool from the list of tools in the Analysis Center to perform an analysis of a cohort

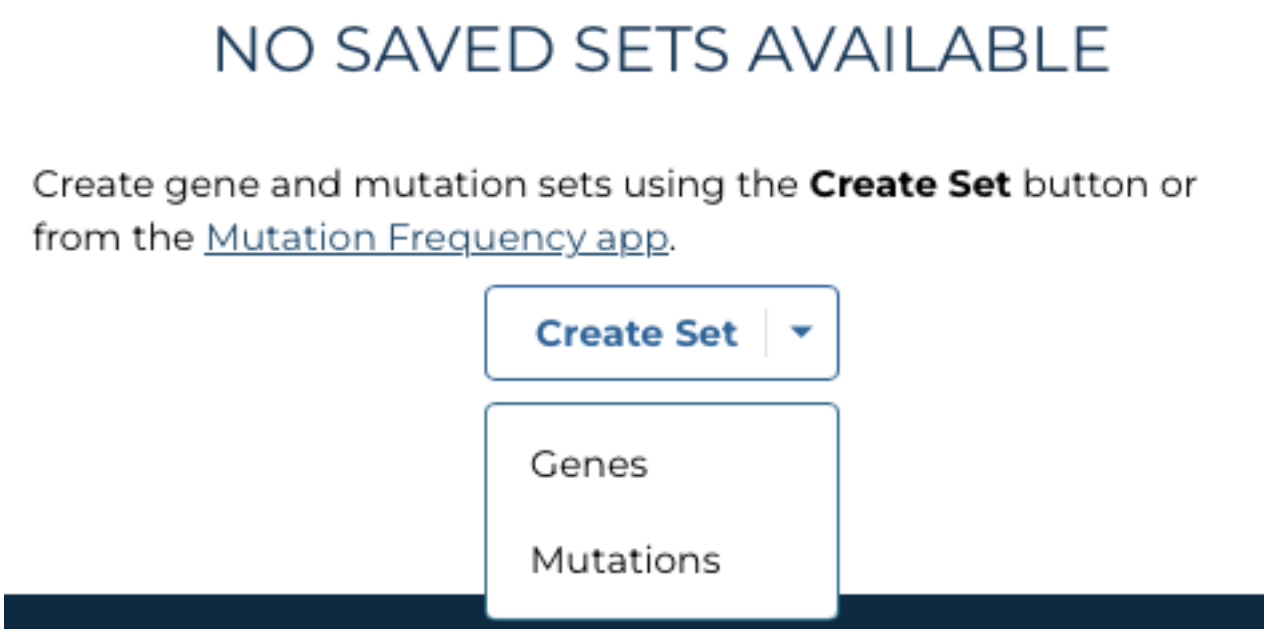
Manage Sets

The `Manage Sets` button at the top of the GDC Portal stores sets of genes or mutations of interest. On this page, users can review the sets that have been saved as well as upload new sets and delete existing sets.

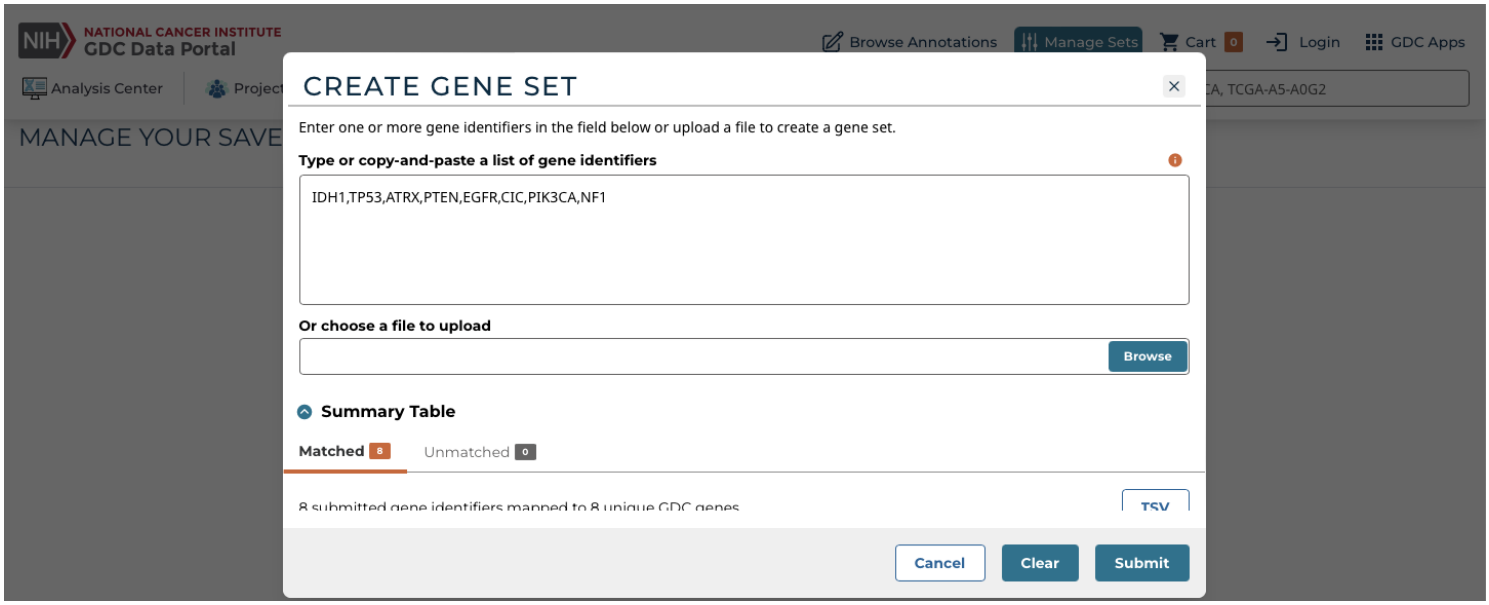


Upload Sets

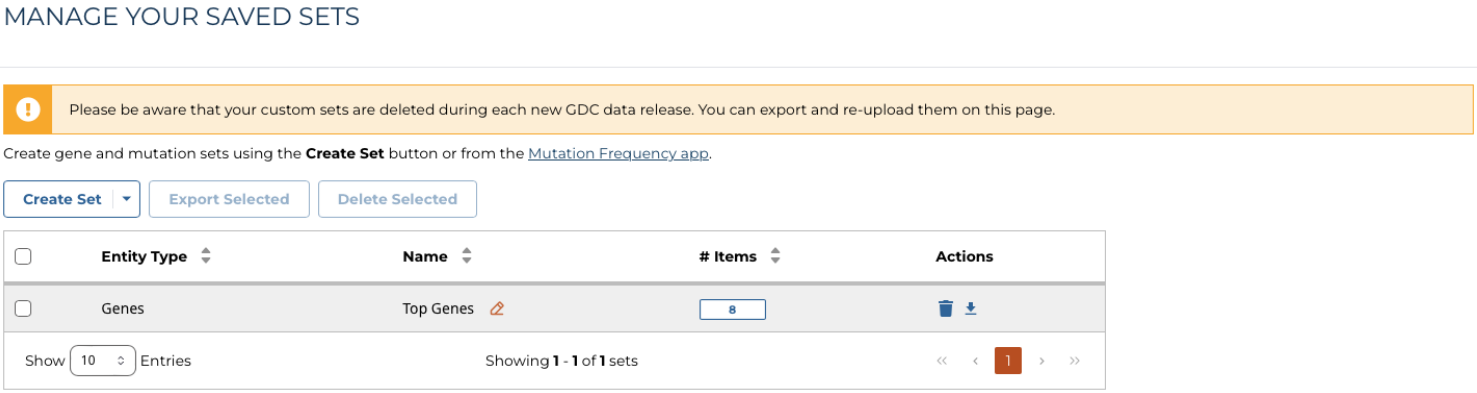
Clicking the `Create Set` button shows options for creating Gene or Mutation sets.



Upon clicking one of the menu items, users are shown a dialog where they can enter unique identifiers (i.e. gene symbols, mutation UUIDs, etc.) that describe the set.

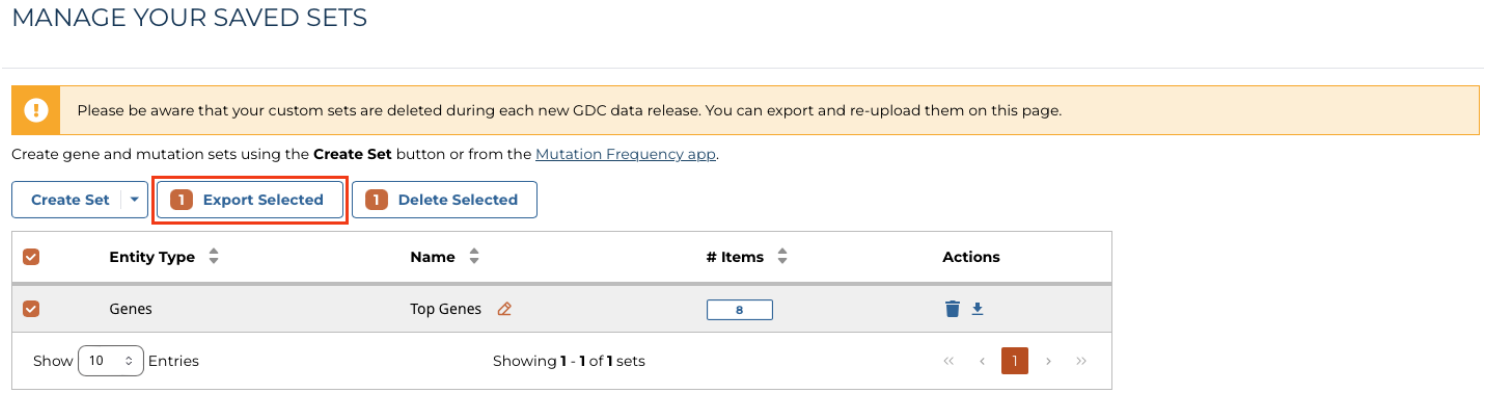


Clicking the `Submit` button will add the set of items to the list of sets on the Manage Sets page.



Export Sets

Users can export selected sets on this page by first clicking the checkboxes next to each set, then clicking the `Export selected` button at the top of the table.



A text file containing the ID of each gene or UUID of each mutation is downloaded after clicking this button.

Review Sets

There are a few buttons in the list of sets that allows a user to get further information about each one.

- ___ Items___: Clicking the button under the Items column launches a table with Gene ID and Symbol for gene sets or Mutation ID and Consequence for mutation sets
- **Delete/Download**: To the right of the Items column are buttons that will delete the set or download the list as a TSV file

Chapter 2

Quick Start

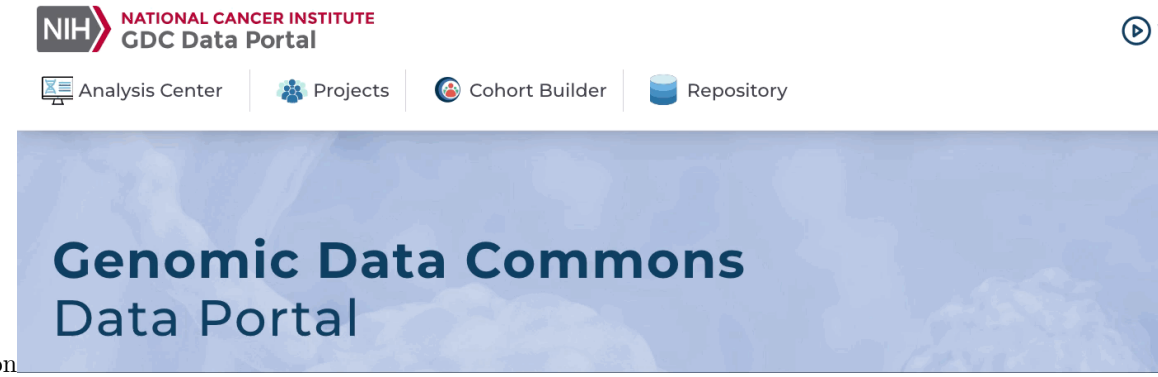
Quick Start Page

The purpose of this guide is to quickly introduce researchers to the GDC Data Portal. This is not a comprehensive overview of the Data Portal and may not contain details for your specific use-case. Please see the rest of the Data Portal documentation pages for information about specific tools.

Start at https://portal.gdc.cancer.gov/.

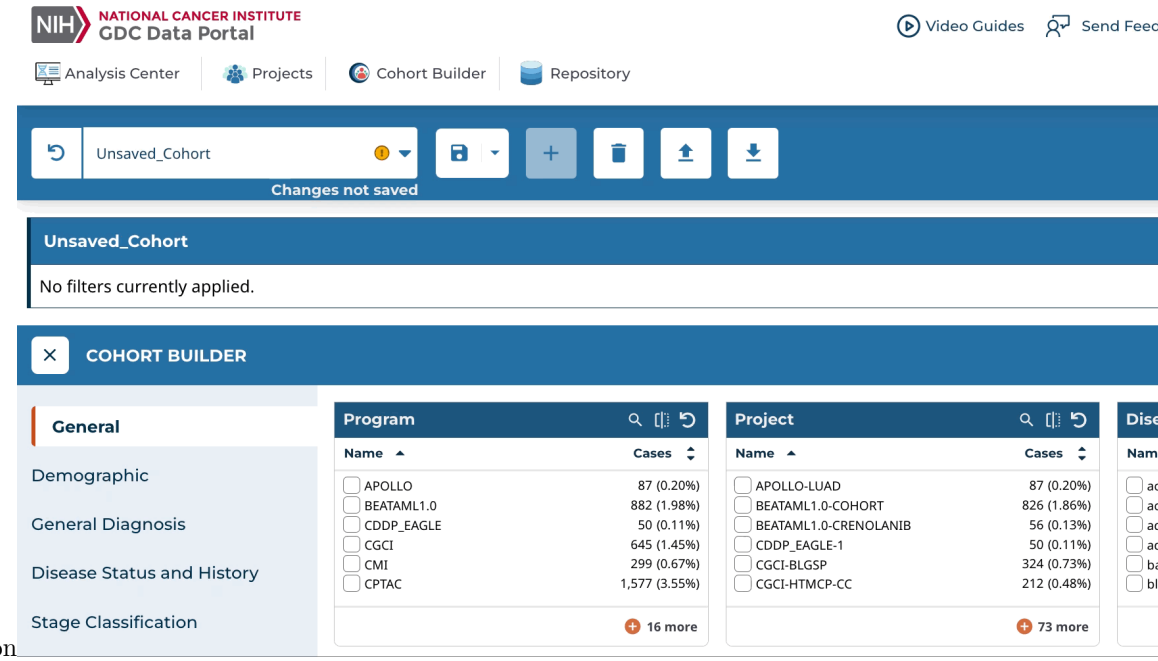
Building and Analyzing a Cohort

Step 1: Go to the Cohort Builder at the top left of the Data Portal.



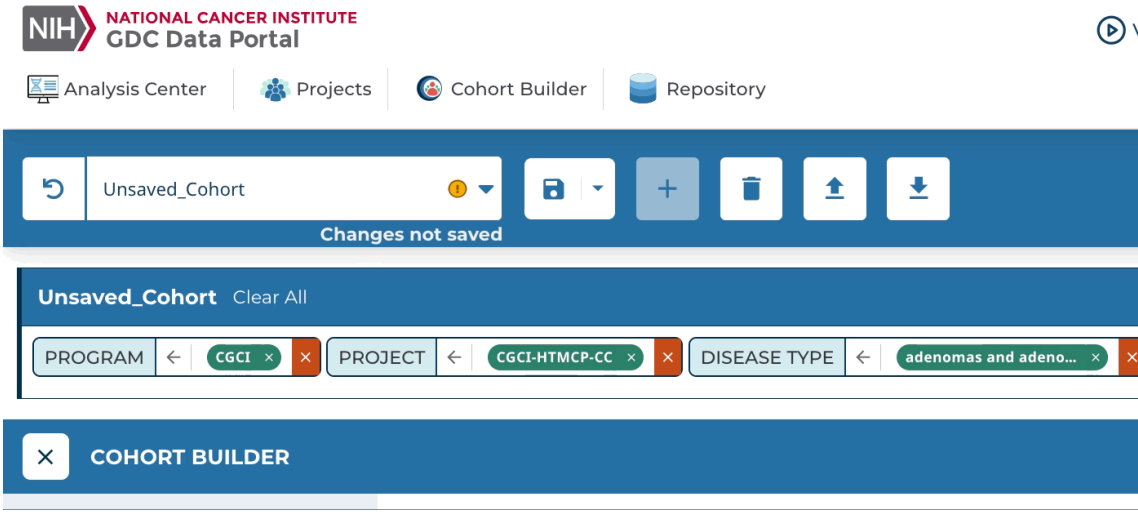
Click to Expand/Collapse Animation

Step 2: Use the filters in the Cohort Builder to filter down the full set of GDC cases to a subset you are interested in. Filter categories can be selected in the left panel.



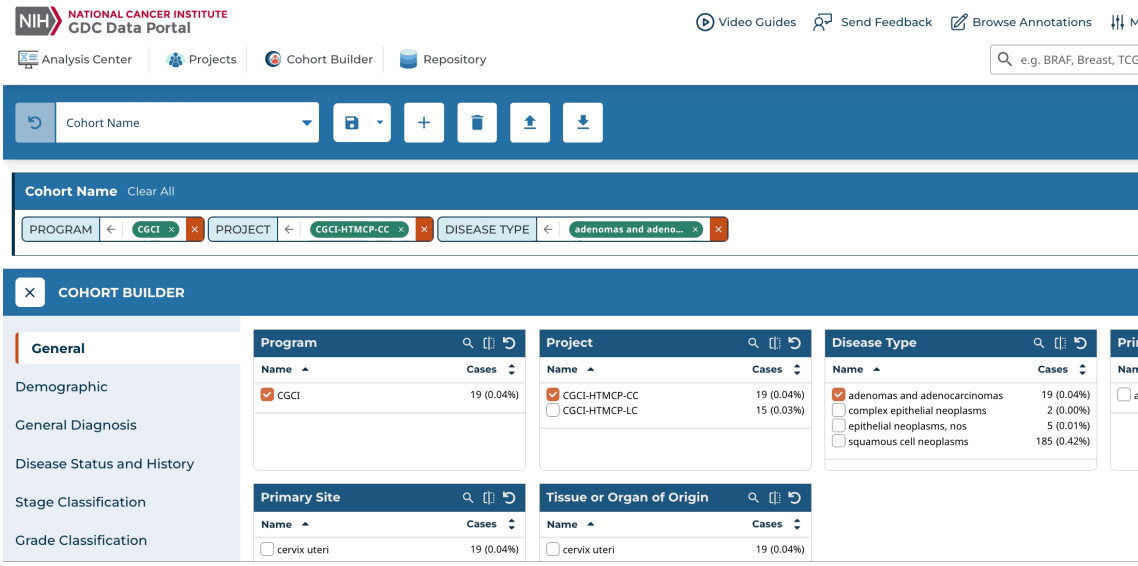
Click to Expand/Collapse Animation

Step 3: Save your cohort by clicking the "Save" icon in the cohort bar, choosing "Save", and naming your cohort when prompted.



Click to Expand/Collapse Animation

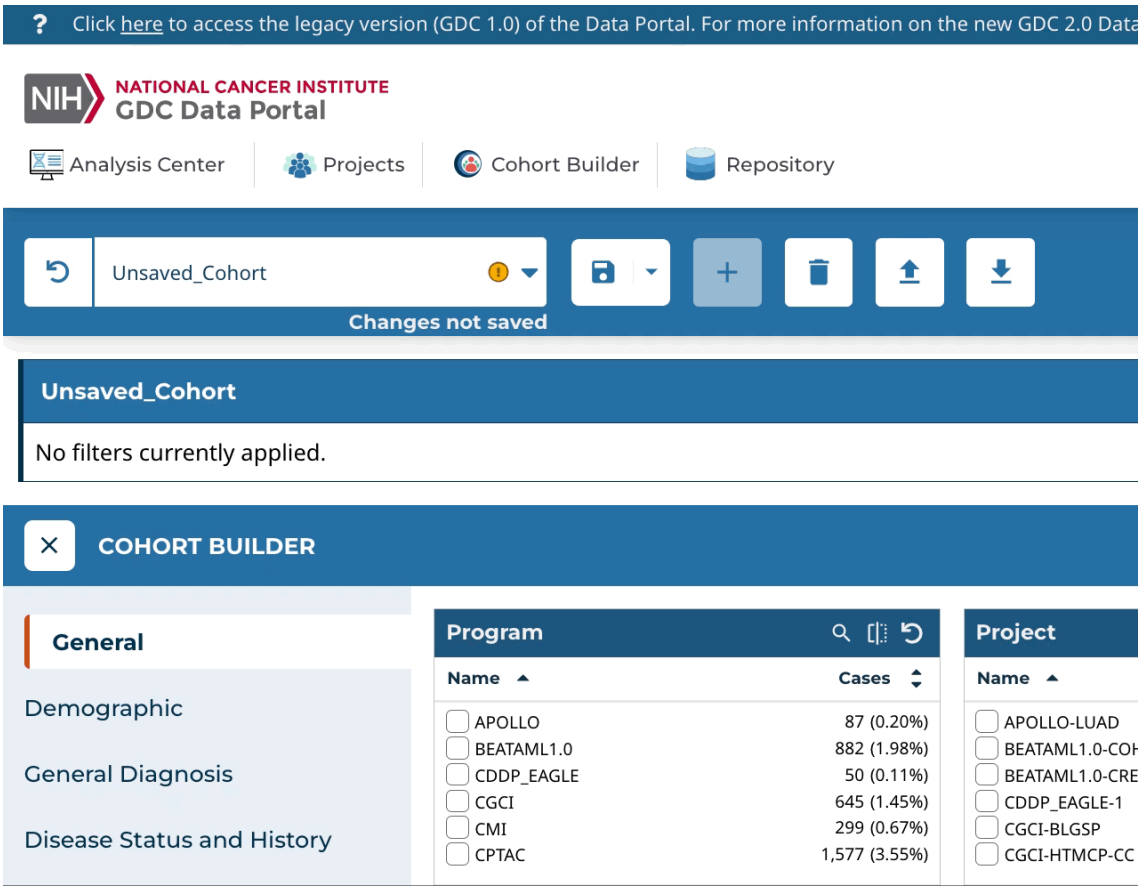
Step 4: The cohort you created is now your active cohort. Go to the Analysis Center at the top left of the Data Portal. Choose the tool you would like to use from the list. The analysis will apply to the data from your cohort.



Click to Expand/Collapse Animation

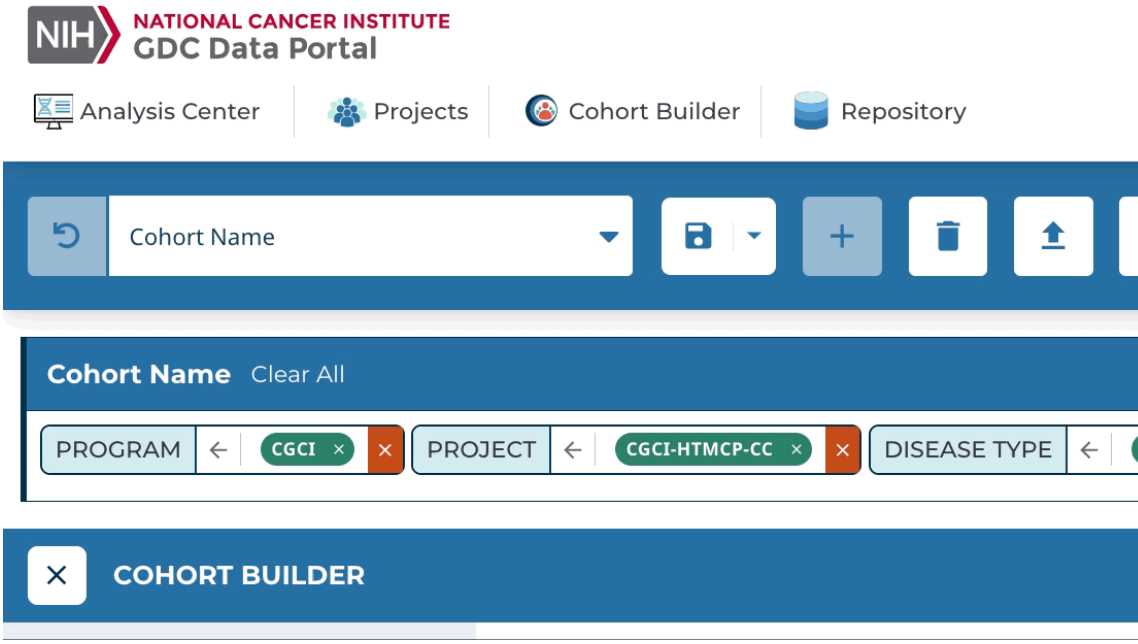
Downloading Files

Step 1: Select your cohort of interest, create one if you have not already.



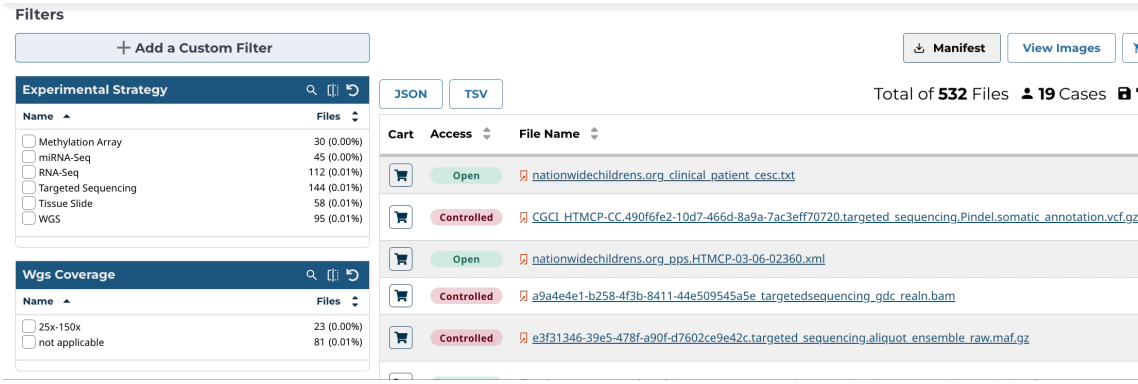
Click to Expand/Collapse Animation

Step 2: Go to the Repository tool from the Analysis Center or the top left of the portal.



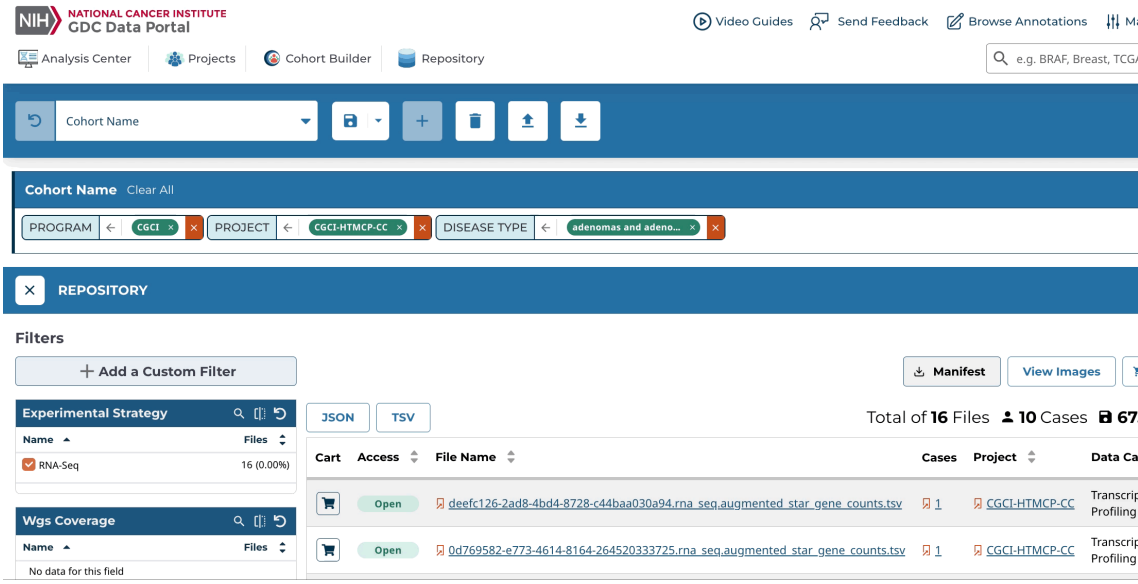
Click to Expand/Collapse Animation

Step 3: The Repository tool will display files that are associated with the cases in your active cohort. Narrow down your file selection by filtering them using the panel on the left.



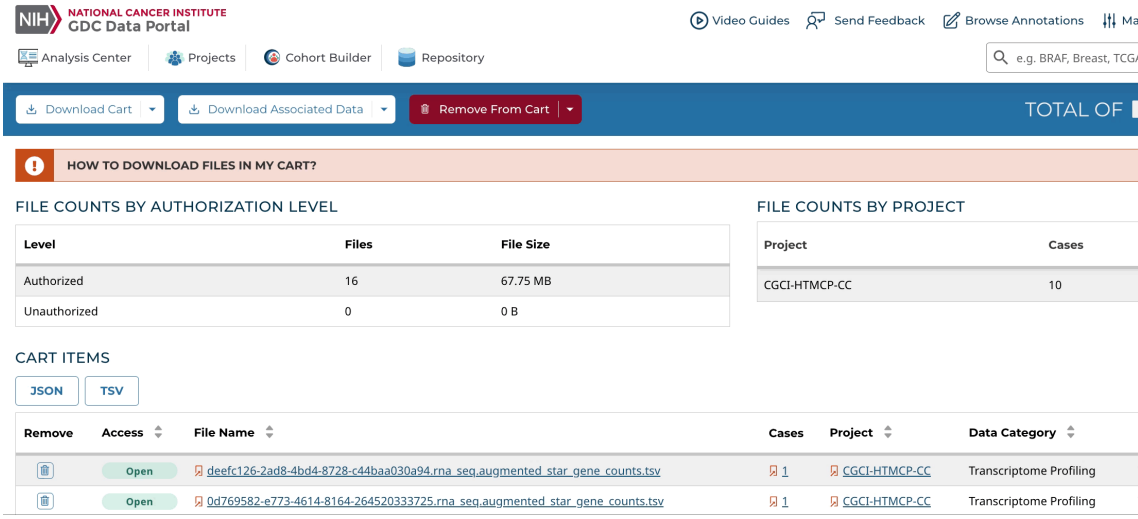
Click to Expand/Collapse Animation

Step 4: When you are done filtering, add the files to your cart. Then go to the cart.



Click to Expand/Collapse Animation

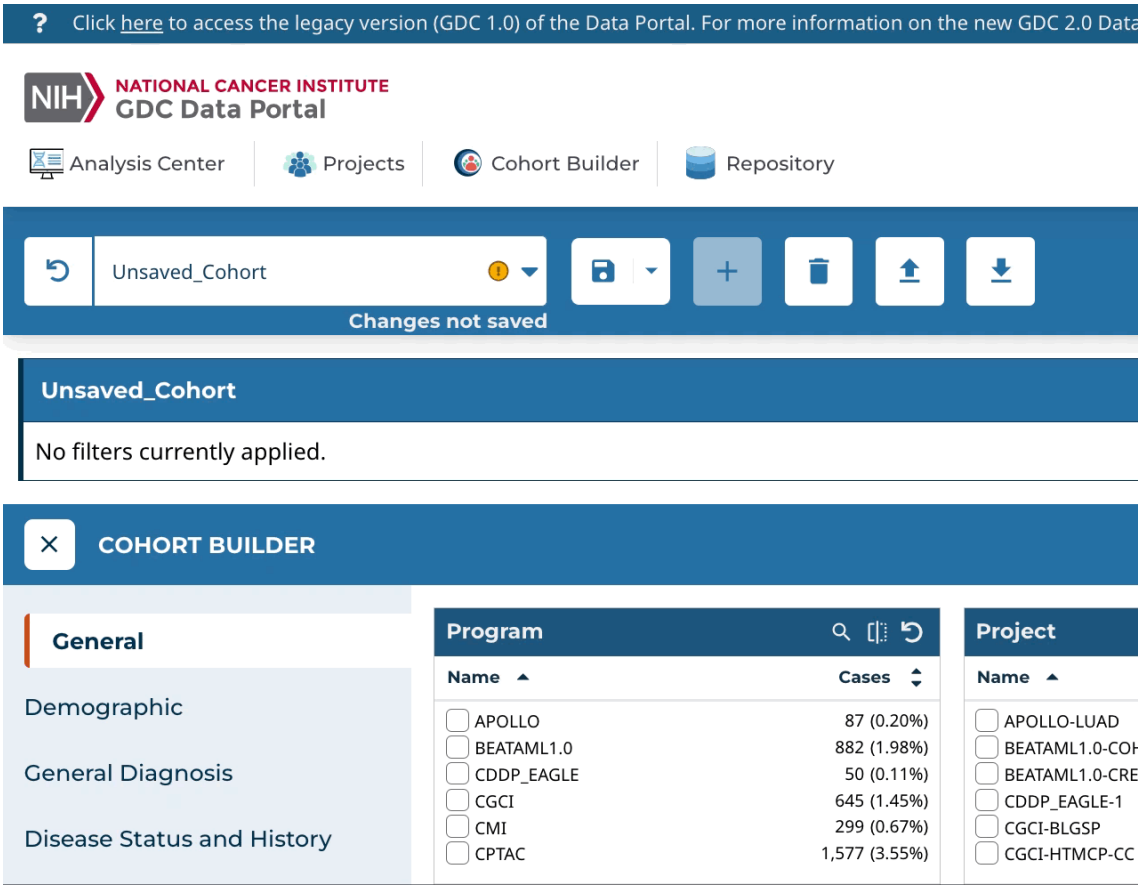
Step 5: The files in the cart can be downloaded directly from the browser or a manifest can be downloaded and passed to the GDC Data Transfer Tool.



Click to Expand/Collapse Animation

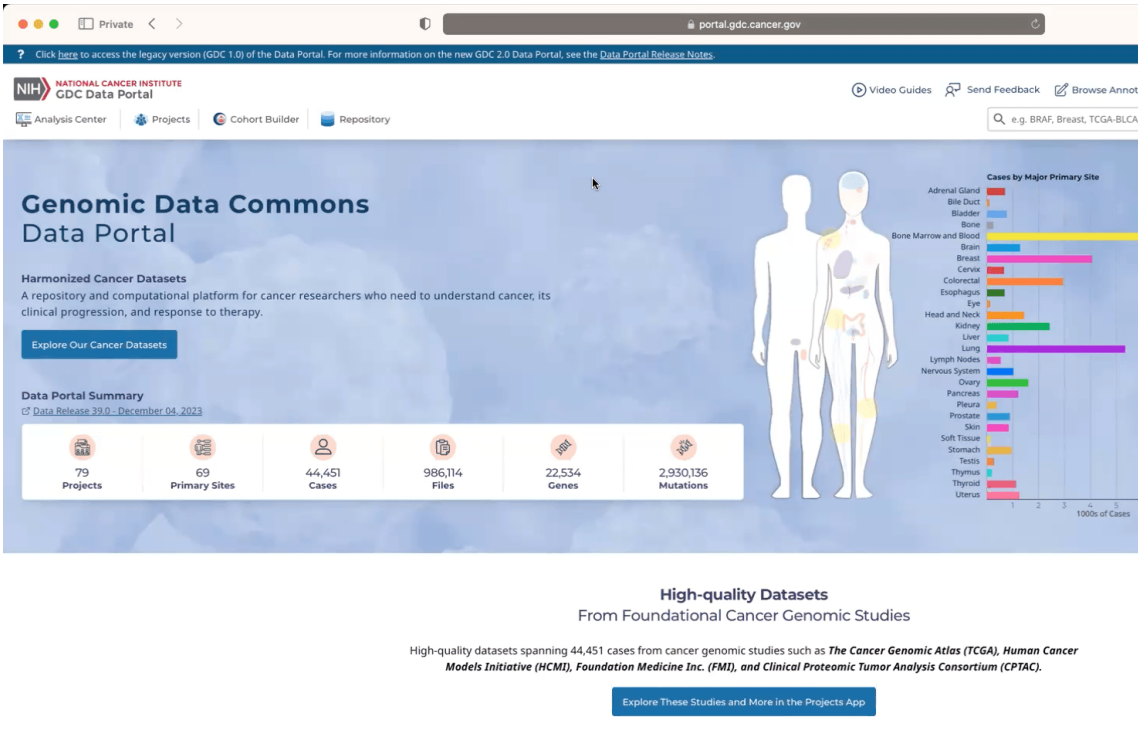
Viewing Mutations

Step 1: Select your cohort of interest or create one if you have not already.



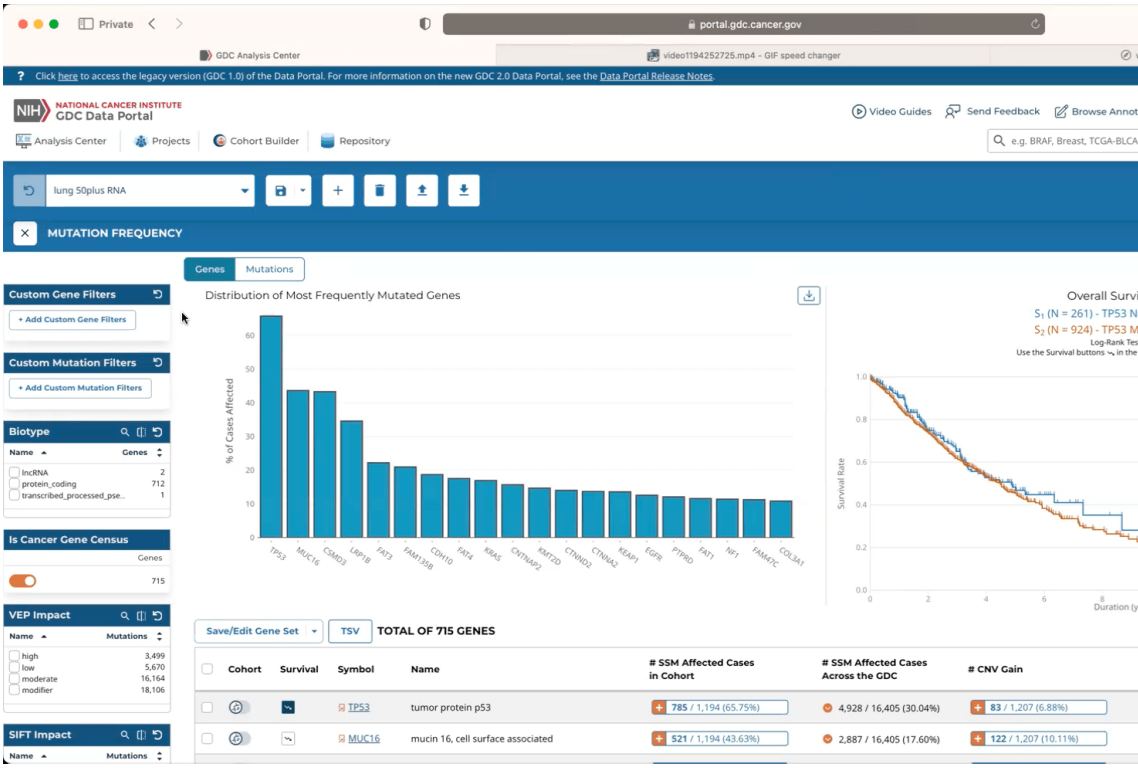
Click to Expand/Collapse Animation

Step 2: Launch the Mutation Frequency tool from the Analysis Center.



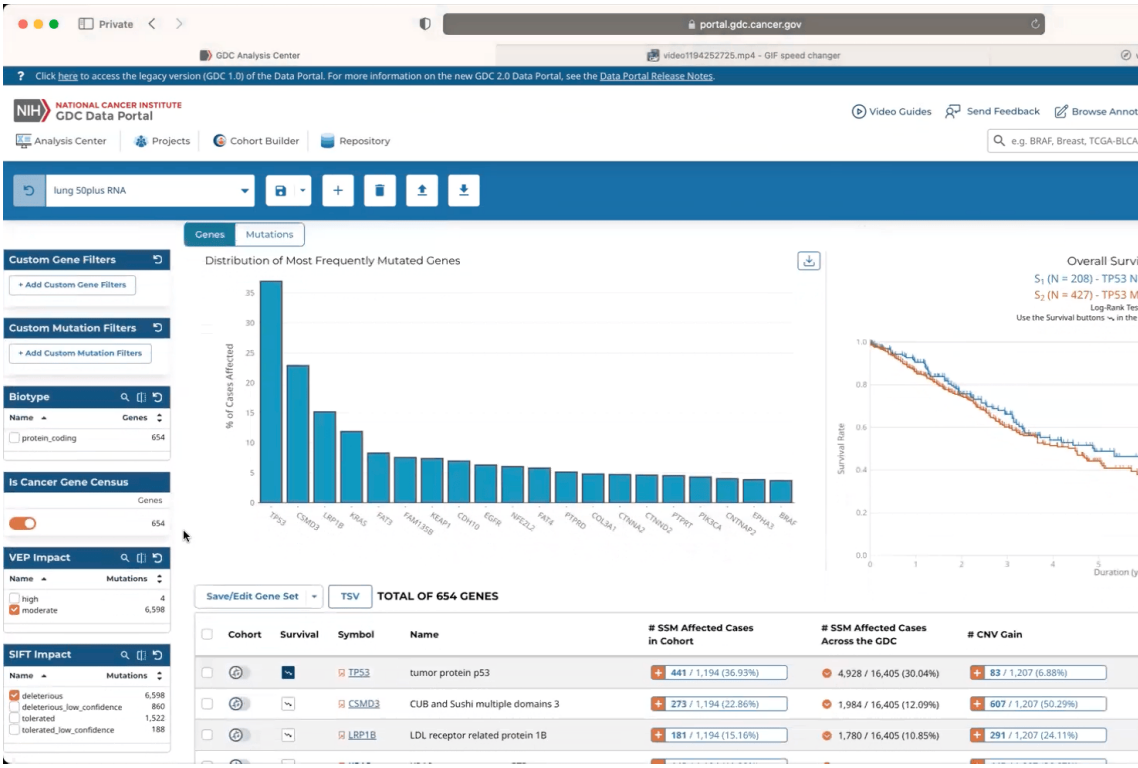
Click to Expand/Collapse Animation

Step 3: The Mutation Frequency tool visualizes the most frequently mutated genes and the most frequent somatic mutations for the active cohort. Narrow down your results by filtering them using the panel on the left.



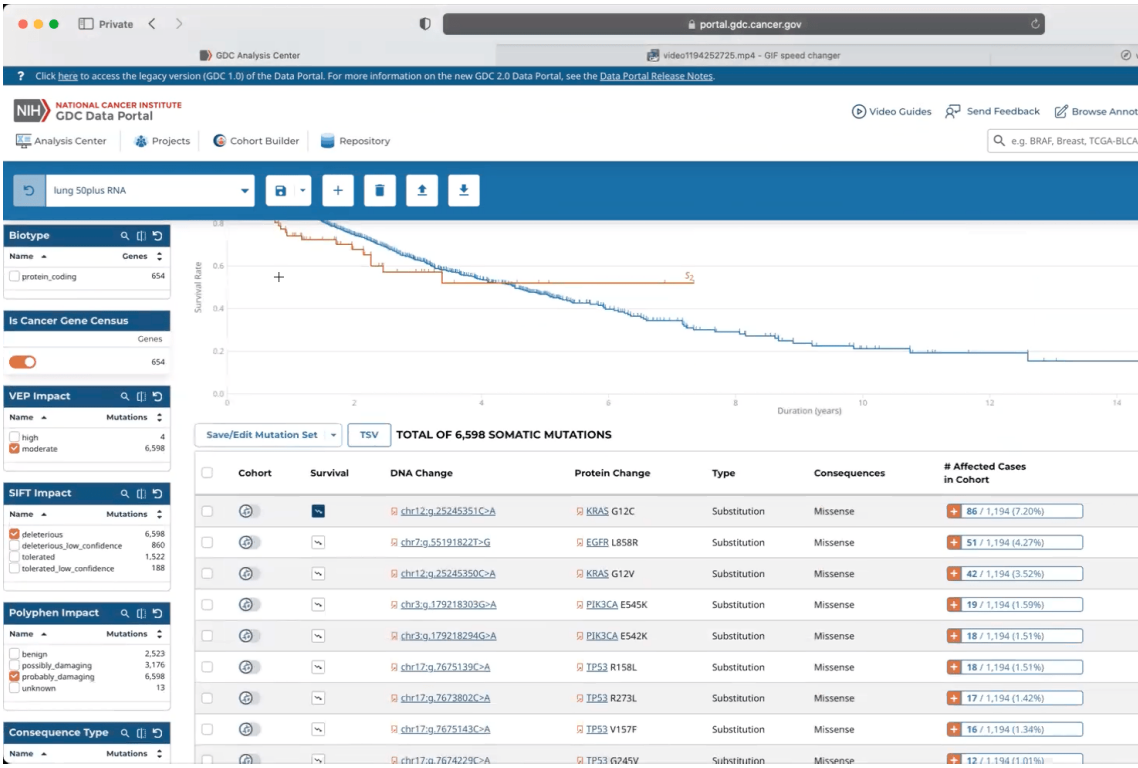
Click to Expand/Collapse Animation

Step 4: To view the most frequent somatic mutations, navigate to the Mutations tab.



Click to Expand/Collapse Animation

Step 5: The mutation table can be downloaded by clicking on the TSV button at the top of the table.



Click to Expand/Collapse Animation

Chapter 3

Cohort Builder

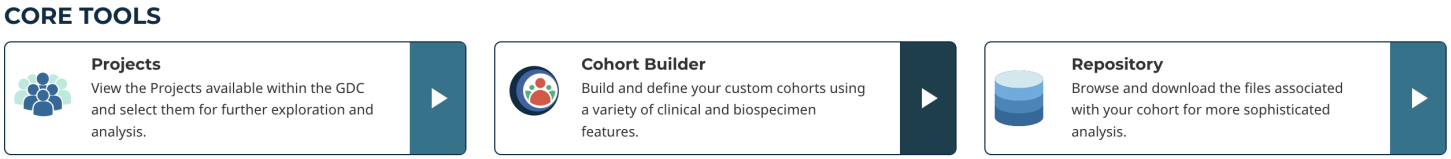
Cohort Builder

The Cohort Builder is a good starting point for users looking to gather information for a specific disease, project, or group of patients. Building a cohort allows users to download files, perform analyses, and query metadata for the same group of cases in multiple sections of the GDC Data Portal. This section will cover the process of building a cohort and downstream actions will be documented in their respective sections.

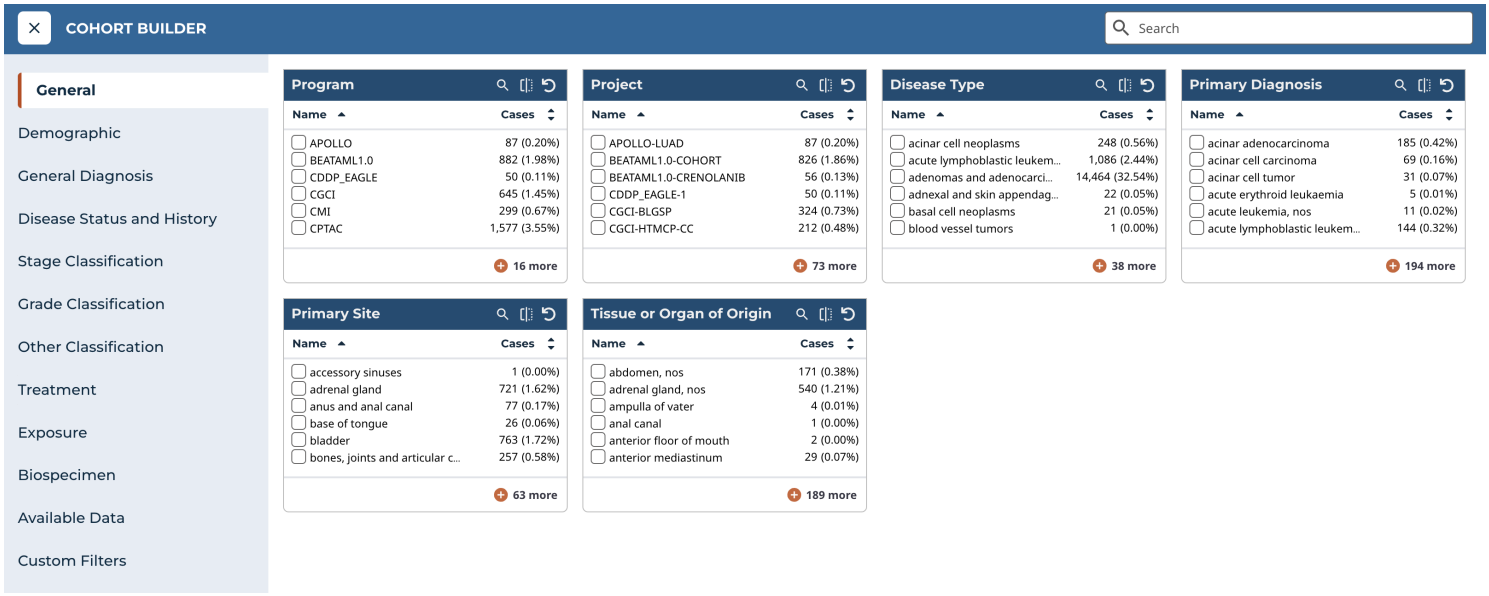
The Cohort Builder can be accessed in one of the following ways:

- Selecting the Cohort Builder link in the GDC Data Portal header

- Selecting the play button on the Cohort Builder card in the Analysis Center



Cohort Builder Panel



The Cohort Builder tool will be displayed as a panel in the Analysis Center and is used to filter the current cohort to a specific set of cases. The current cohort is always displayed in the main toolbar and can be changed from the main toolbar.

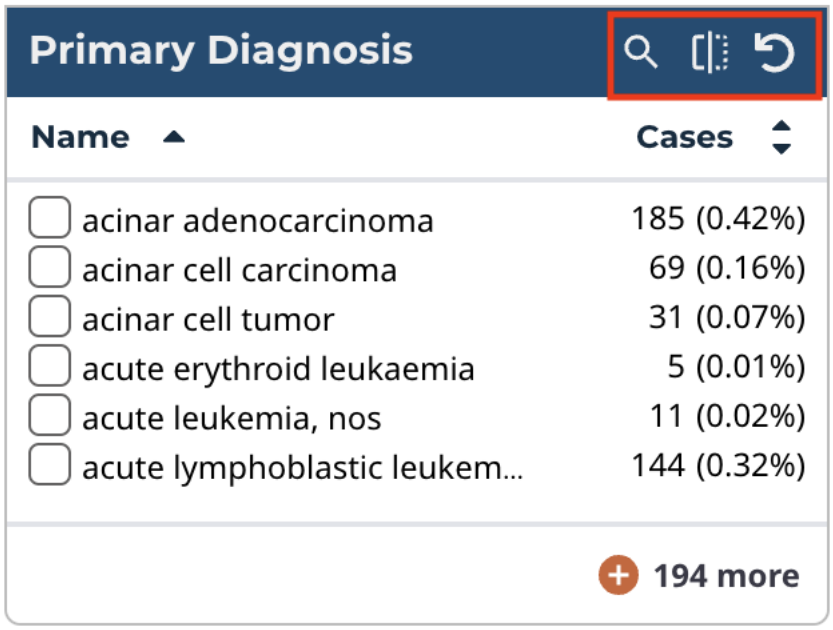
At the left side of the panel are a series of broad filter categories can be selected. Each filter category contains a set of specific filters within cohort builder cards that can be used to narrow your cohort to the desired set.

Cohort Builder Cards

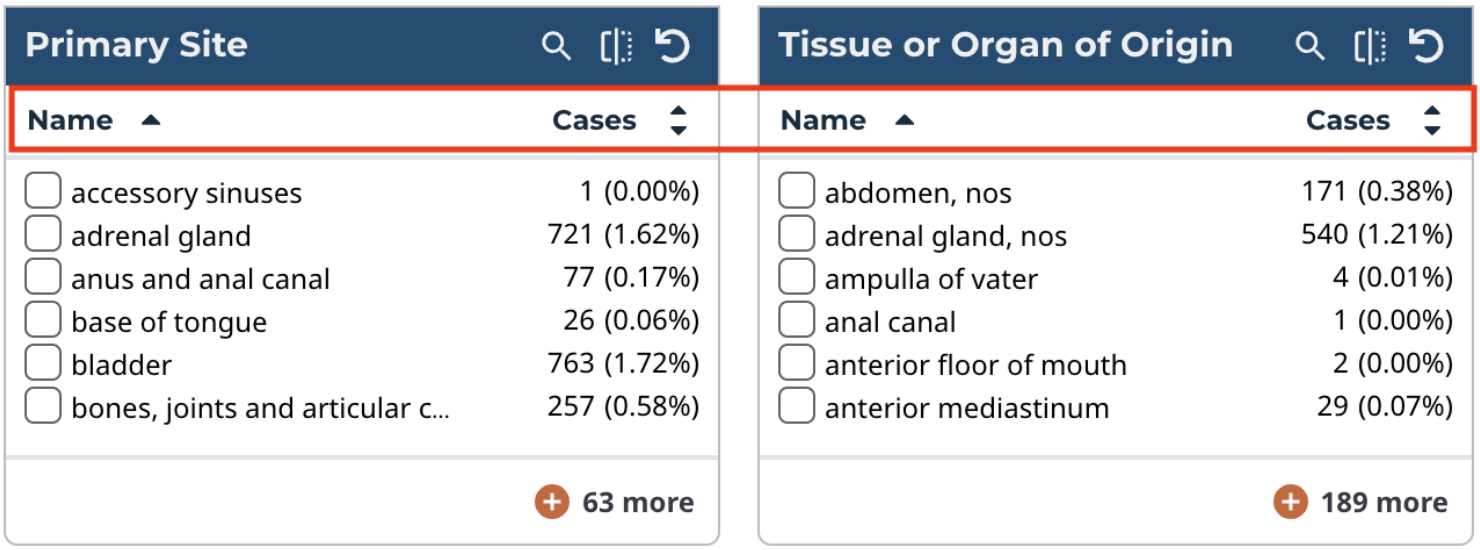
Each card within the Cohort Builder can be used to apply the corresponding filters on the current cohort. As filters are applied, they will be displayed on the Query Expressions section.

Additional features can be accessed at the top right of each card's header to facilitate filtering:

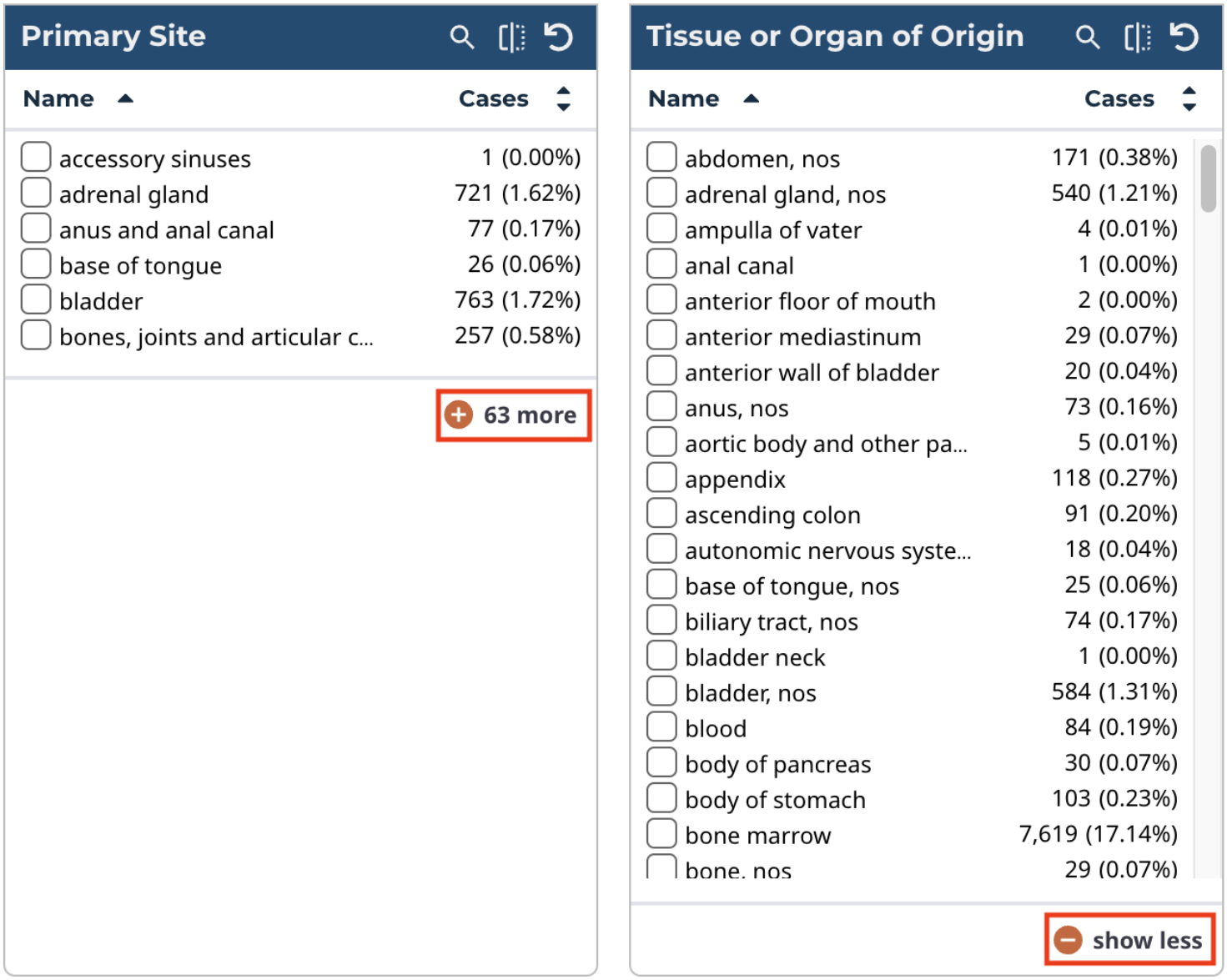
- **Search**: the search icon can be selected to reveal or hide a search field for entering text to search within the values of the current card. This feature is only available when the values are enums.
- **Flip Card**: cards can be flipped to reveal or hide a summary chart. This feature is only available when the values can be meaningfully displayed as bar graphs.
- **Reset Card**: this button will reset any filtering that has been applied within the card



In addition, filters in each card can be sorted, either alphabetically or by the number of cases based on current filters, by selecting one of the two headers directly underneath the card title. The default sort is alphabetical order.



The first six (or fewer) filters are shown for each card, but can be expanded to show 20 filters at once by clicking the "+" button which also indicates the number of additional filters not in view. The expanded view can be toggled off by clicking the resulting "show less" button.

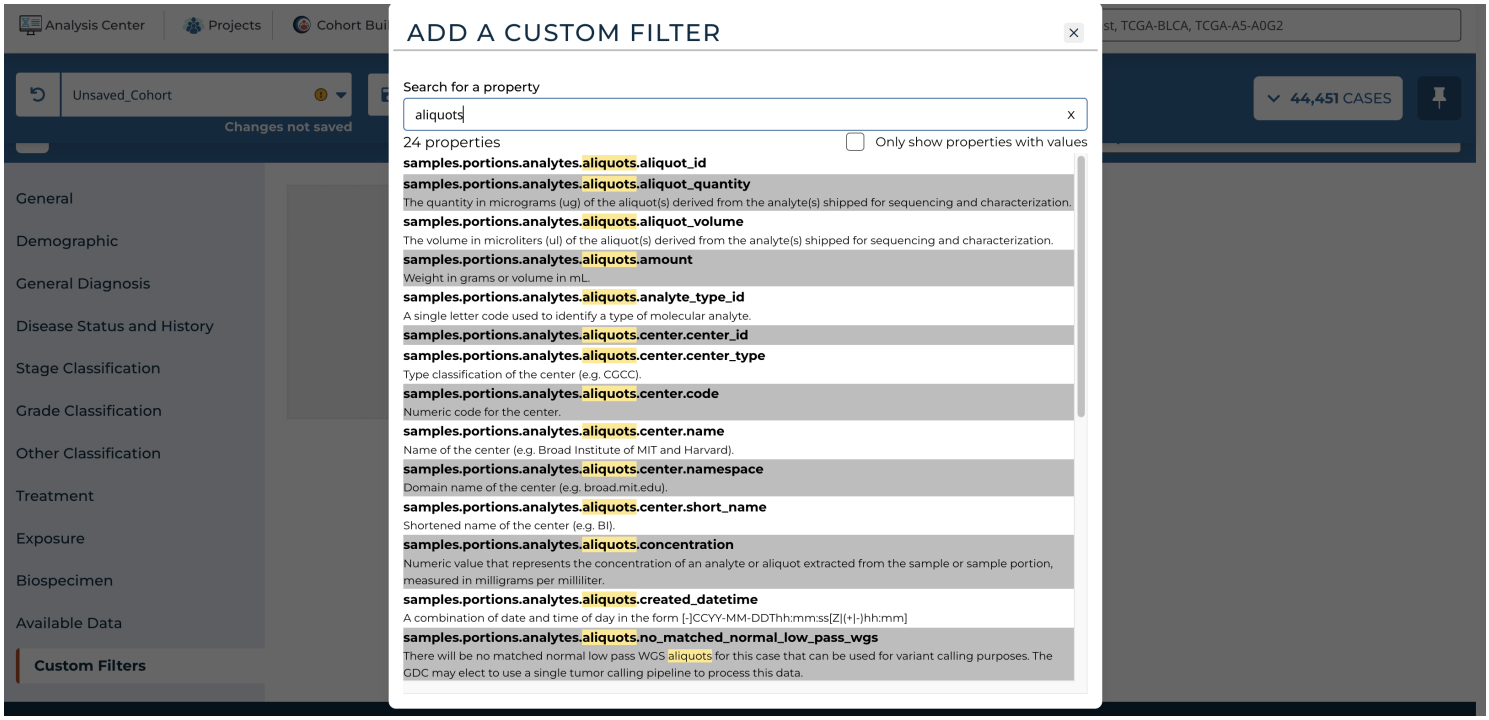


Available Data Filters

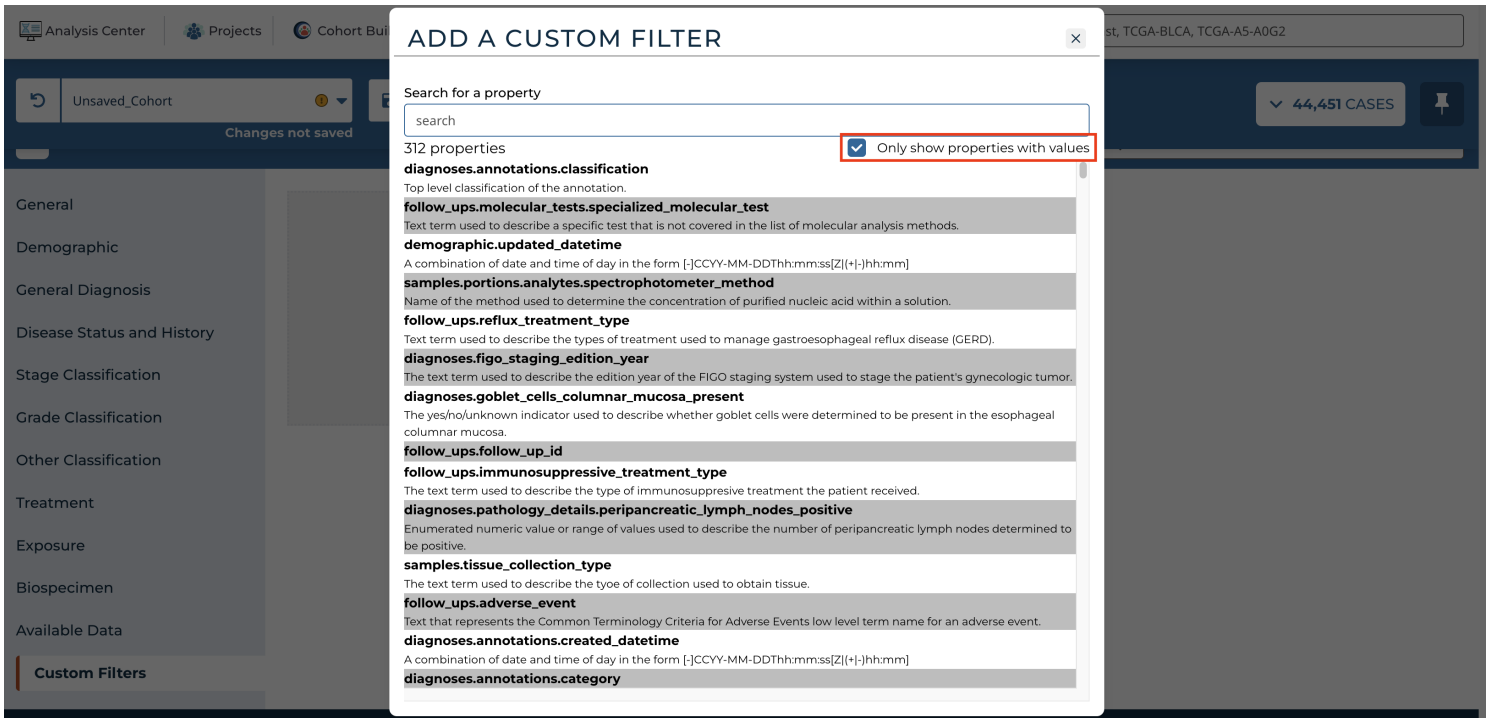
Toward the bottom of the list of filter categories, "Available Data" can be selected. These filters differ from the other default filters as they allow for cases that have certain types of associated data files. For example, filtering in the "Experimental Strategy" card for "RNA-Seq" will only display cases in the active cohort that have associated RNA-Seq files. These filters may be useful for studies that require only cases for which a certain type of analysis was performed.

Custom Filters

If a filter cannot be found within one of the categories, use the "Add a Custom Filter" button in the "Custom Filters" category to access any filters that are not displayed. Browse through the list of additional filters, or use the search box to search for filters by name. Once a filter is selected, it is then added to the "Custom Filters" category. A custom filter can be removed from this category by choosing the "X" at the top right of the filter card.



Filters that exist in the GDC but do not have any cases that have a value for the filter can also be removed from the "Custom Filters" list by selecting the "Only show properties with values" box.

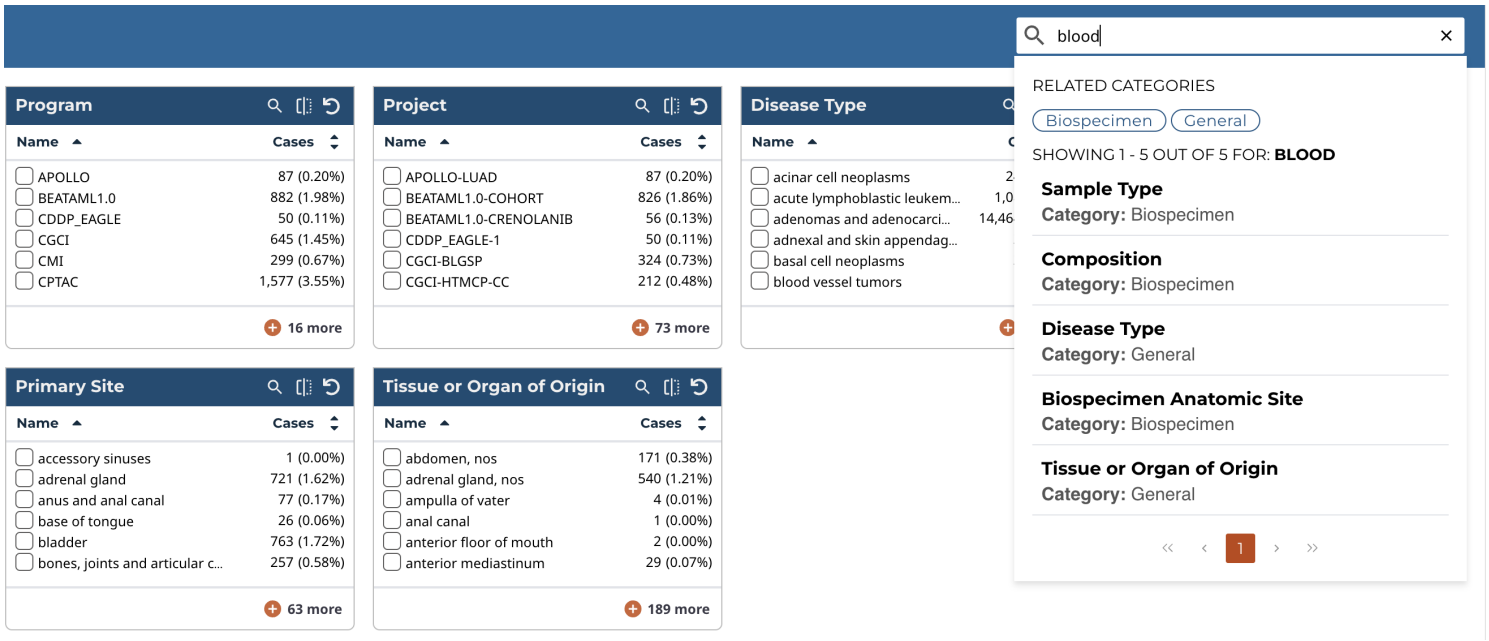


Cohort Builder Search

The Cohort Builder includes the ability to search across all the cards within it. This feature is located on the right of the Cohort Builder header.

As a search term is entered, the Cohort Builder Search feature will display a list of properties that contain matching results.

When a result is moused over, additional information is displayed to its left, including a description of the property and a list of values that match the search term.



When a result is selected, the card corresponding to the selected result will be displayed.

Closing the Cohort Builder

Once a custom cohort is built and filtering is complete, users can close the Cohort Builder and use the custom cohort with other tools.

To close the Cohort Builder panel and display all the tools within the Analysis Center, click on the "X" button on the left of the Cohort Builder header.

Alternatively, users can select the Analysis Center link or any of the other links on the GDC Data Portal header to close the Cohort Builder.

Changes made to the cohort with the Cohort Builder will persist through the other sections of the GDC Data Portal.

Users can then perform the following actions:

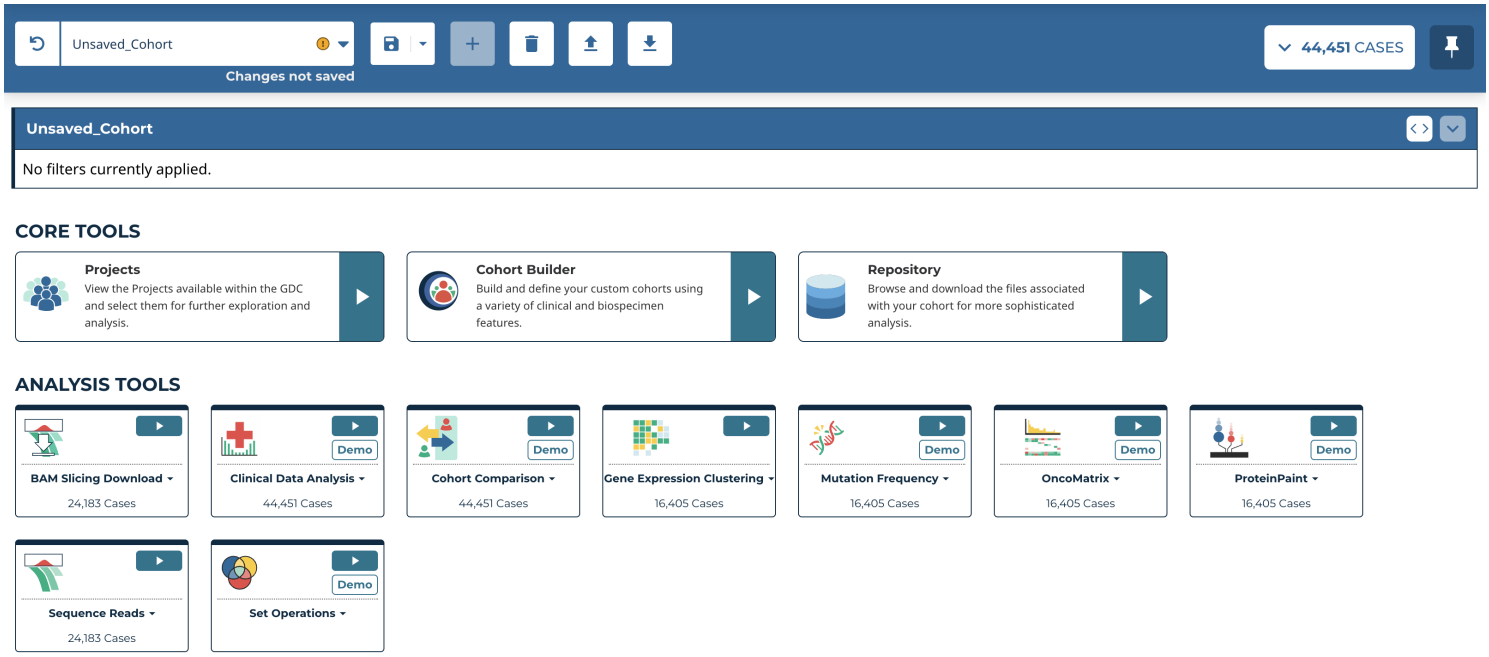
- Download files associated with the cohort from the Repository
- Analyze data from the cohort in the Analysis Center

Chapter 4

Analysis Center

GDC Analysis Center

The Analysis Center is the central hub for accessing the tools that support cohort analysis. The Analysis Center can be accessed by clicking on the 'Analysis Center' icon in the GDC Data Portal header, the "Explore Our Cancer Datasets" button on the home page, or one of the sites in the human anatomical outline or bar graph.



The Analysis Center consists of a main toolbar and a query expressions section, both of which are always displayed. The main toolbar displays the active cohort and can be used to create and manage custom cohorts. The query expression section displays the filters applied to the active cohort.

Available tools are displayed under the Query Expression section of the Analysis Center. Each analysis tool is showcased within a tool 'card', which has several items related to the analysis tool such as:

- A teal 'Play' button to launch the analysis tool on the given cohort
- A 'Demo' button that launches a demonstration of the analysis tool on an example cohort
- Clicking on the name of the analysis tool in the tool card toggles a drop down description of the analysis tool
- The number of cases from the cohort that the analysis will be performed on is at the bottom of the card

Core Tools

The 'Core Tools' section contains the GDC tools that constitute the main functionality of the Data Portal.

- Projects tool
- Cohort Builder
- Repository

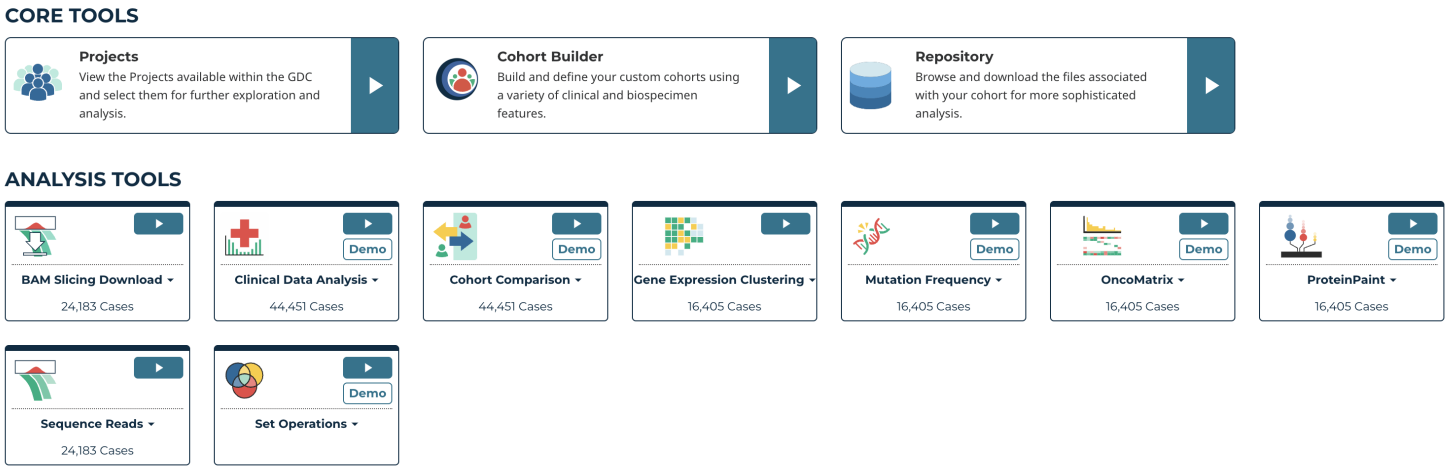
Analysis Tools

The 'Analysis Tools' section contains the tools available for specific analyses available for the active cohort.

- BAM Slicing Download
- Clinical Data Analysis
- Cohort Comparison
- Gene Expression Clustering
- MAF Aggregation
- Mutation Frequency
- OncoMatrix
- ProteinPaint
- Sequence Reads
- Set Operations

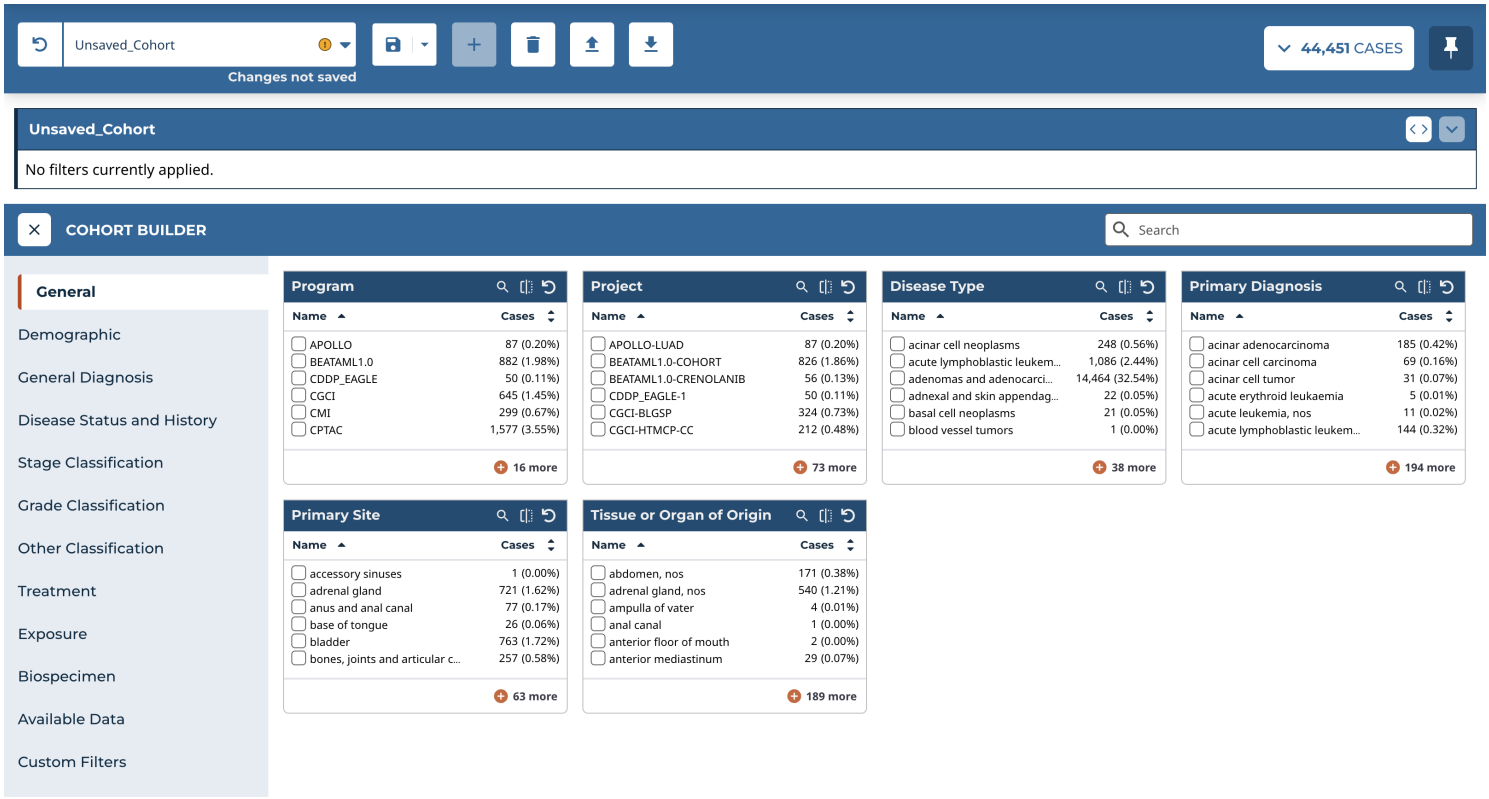
Each can be launched by clicking the Play buttons in each of the tool cards.

If there is not sufficient data in the active cohort to use a particular tool, the play button will be grayed out and will not be usable until a new cohort with sufficient data is selected.



Tool Panel

As each tool is selected, it is loaded in the 'Analysis Center' within a panel.



To close a tool and return to the default view that displays all the tool cards within the Analysis Center, click the “X” to the left of the tool’s header.

Chapter 5

Repository

Repository

Introduction

The Repository tool is where data files associated with each case in the current cohort can be browsed and downloaded. It also offers file filters for identifying files of interest.

The Repository tool can be reached in one of these two ways:

- Choosing the Repository link in the GDC Data Portal header
- Clicking the play button on the Repository card in the Analysis Center

Choosing a Cohort

The files displayed in the Repository will reflect the files that are associated with the active cohort. The current active cohort is displayed in the Main Toolbar.

For users who want to browse all files that are available at the GDC, create a new cohort via the main toolbar and use it with the Repository tool.

Filtering a Set of Files

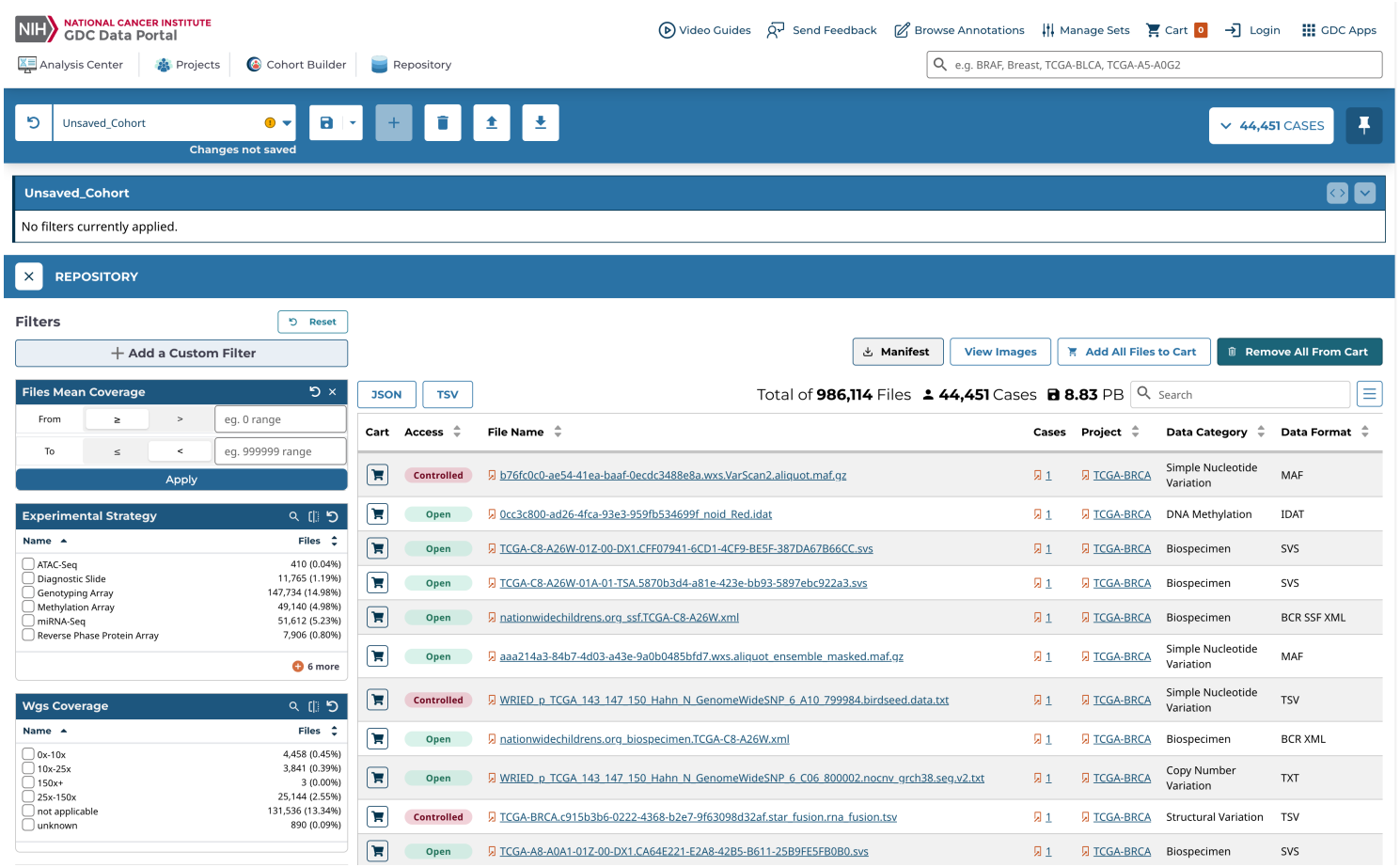
As most users are searching for specific types of files, a set of commonly-used default facet cards can be used in the left panel of the Repository tool to allow users to filter the files presented in the table on the right. The facet cards are as follows:

- **Experimental Strategy**: Experimental strategies used for molecular characterization of the cancer
- **WGS Coverage**: Range of coverage for WGS aligned reads
- **Data Category**: A high-level data file category, such as "Raw Sequencing Data" or "Transcriptome Profiling"
- **Data Type**: Data file type, such as "Aligned Reads" or "Gene Expression Quantification". Data Type is more granular than Data Category.
- **Data Format**: Format of the data file

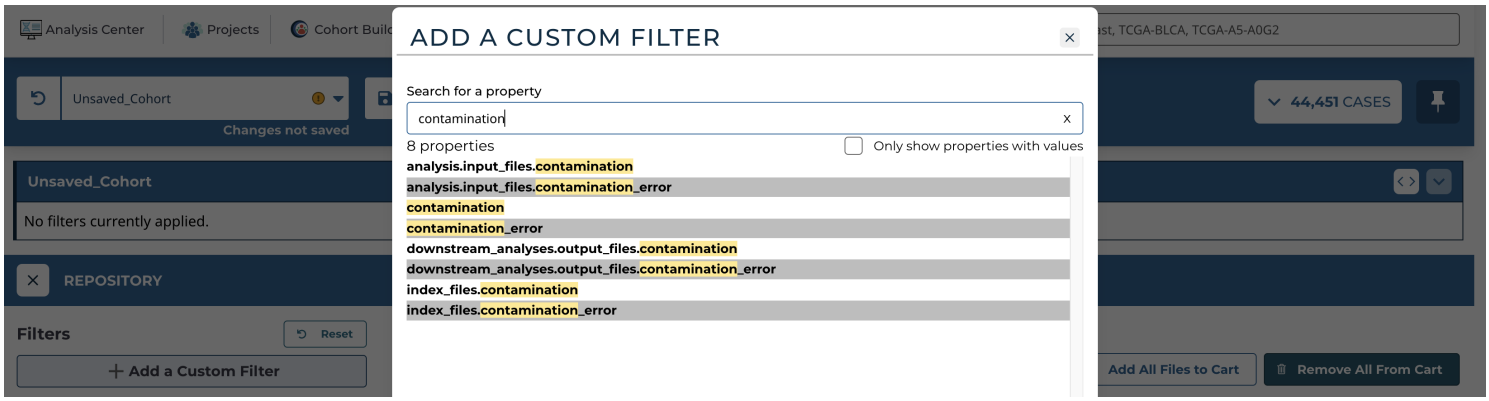
- **Workflow Type**: Bioinformatics workflow used to generate or harmonize the data file
- **Platform**: Technological platform on which experimental data was produced
- **Access**: Indicator of whether access to the data file is open or controlled

Values within each facet can be sorted alphabetically by choosing the "Name" header on the top left of each card. Alternatively, the "Files" header may be selected to sort the values by the number of files available.

Note that the categories displayed in the filters represent the values available for the active cohort.



If a different filter needs to be used, a custom filter can be applied by choosing the "Add a Custom Filter" button at the top of the default filters. Each custom filter can then be searched and chosen within the pop-up window. Once a custom filter is selected, a new filter card will appear at the top of the default filters. Custom filters can be removed from the Repository by choosing the X at the top right of each filter card.



Viewing Images

To view images associated with the active cohort, select the View Images button above the files table to launch the Slide Image Viewer.

Files Table

The table shows the list of all the files associated with the active cohort, subject to any filtering that has been applied in the Repository. By default, the table provides the following information for each file:

- **Access**: Displays whether the file is open or controlled access. Users must login to the GDC Portal and have the appropriate credentials to access these files.
- **File Name**: Name of the file. Clicking the link will bring the user to the File Summary Page.
- **Cases**: The number of cases associated with the file
- **Project**: The Project that the file belongs to. Clicking the link will bring the user to the Project Summary Page.
- **Data Category**: Type of data
- **Data Format**: The file format
- **File Size**: The size of the file
- **Annotations**: Whether there are any annotations

Additional information such as Data Type and Experimental Strategy can be displayed using the Customize Columms button above the table. The table can be sorted by clicking on the headers, and the search bar above the table can be used to locate specific files.

The JSON / TSV buttons will download the files' details (file name, file size, data category, access type, etc.) in JSON and TSV format, respectively.

Downloading a Set of Files

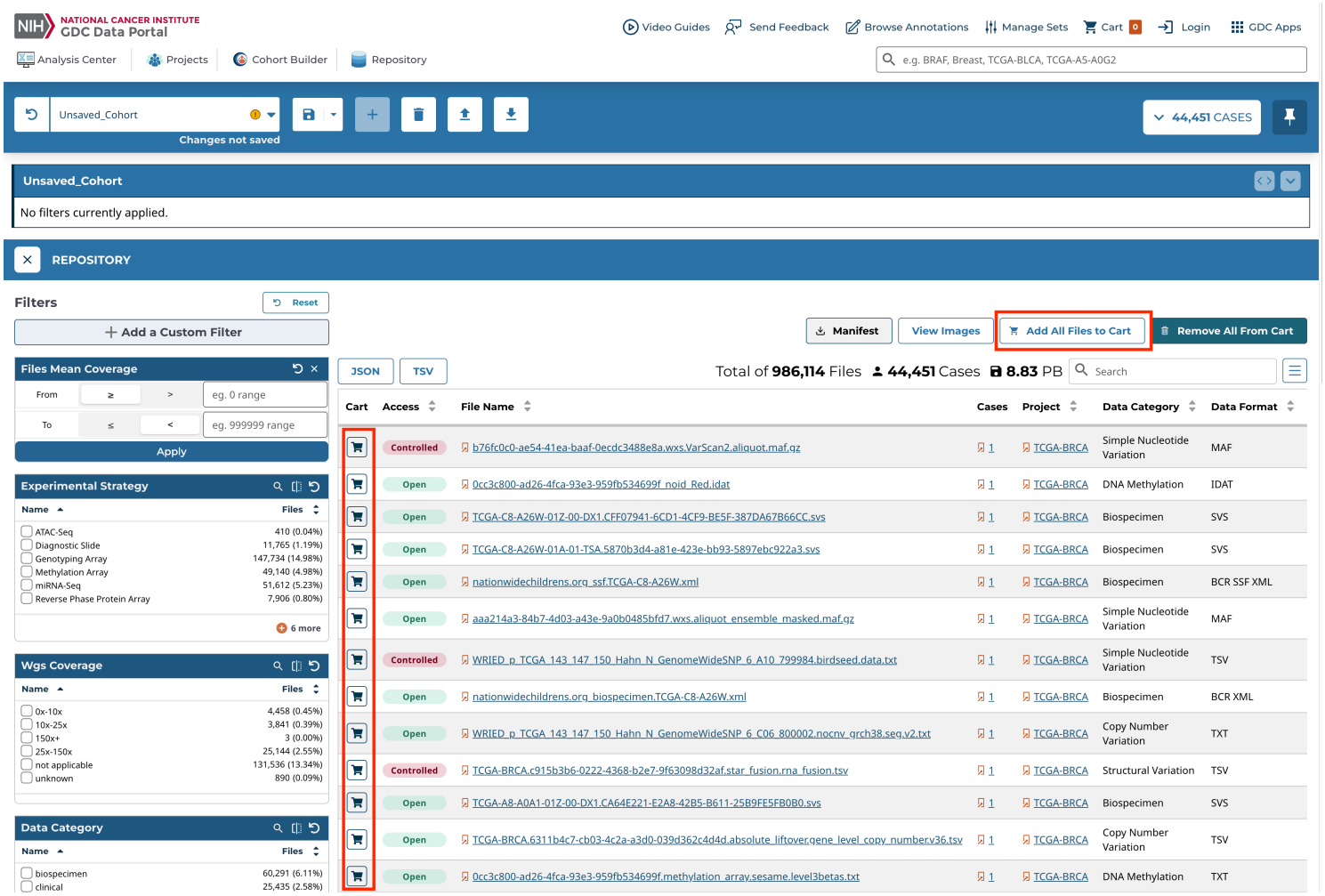
When filtering has been completed, files are ready to be downloaded. Depending on the number and size of files, the GDC has several options and recommendations for downloading them. While any amount of data can be downloaded using the GDC Data Transfer Tool or the API, files can be downloaded directly from the Data Portal if the size is 5 GB or less in total and the number of files does not exceed 10,000. For any downloads larger than 5 GB or 10,000 files, it's recommended that the download be performed using the GDC Data Transfer Tool.

Generating a Manifest File for the Data Transfer Tool

Select the Manifest button above the table to generate a manifest file required for batch download using the Data Transfer Tool. The manifest contains a list of the UUIDs corresponding to the files associated with the active cohort, subject to any filtering in the Repository.

Adding/Removing Files to the Cart for Download

Downloads can also be performed using the Cart by first adding a set of files to the Cart. This can be done using the following methods: * Clicking the cart icon on the left of each file. This will toggle between adding to and removing the file from the cart. * Selecting the Add All Files to Cart button. This will add all the files in the current cohort to the Cart, subject to any filtering that has been applied in the Repository.

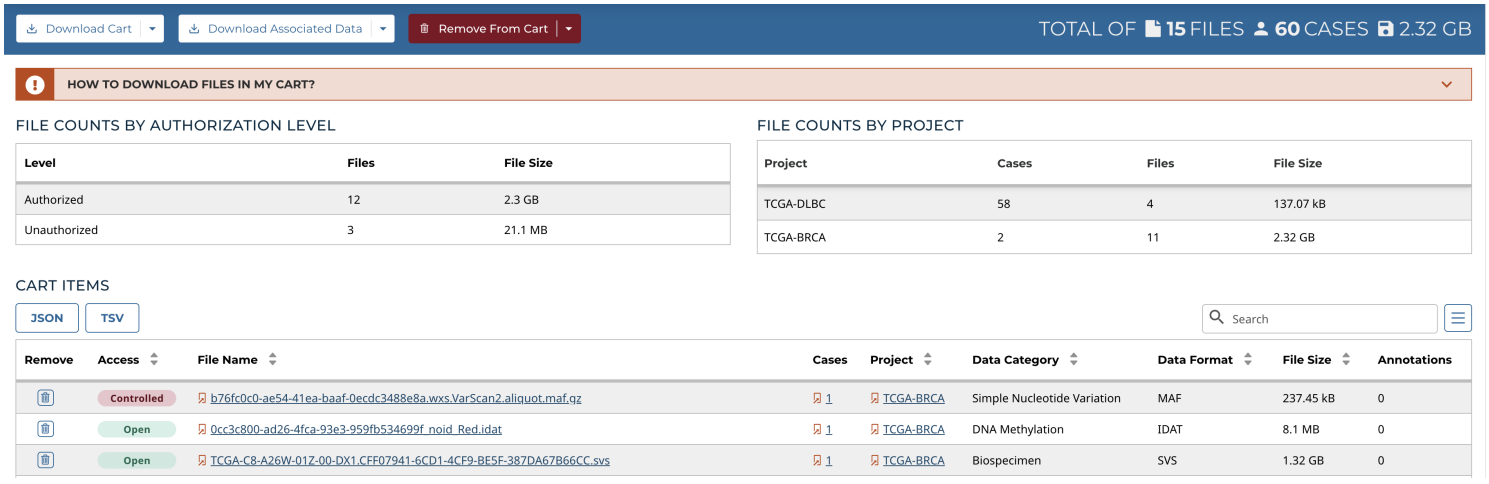


Cart

The Cart page can then be reached by clicking the Cart icon at the top right of the portal.

At the upper-right of the page is a summary of all files currently in the cart: * Number of files * Number of cases associated with the files * Total file size

The Cart page displays the file count by project and authorization level, as well as a table of all files that have been added to the Cart. Files can be removed from the Cart using the trash icons at the left of each file in the table or by selecting the “Remove from Cart” option at the top of the Cart page, which removes either all files or the unauthorized ones.



Cart Items Table

The Cart Items table shows the list of all the files that were added to the Cart and has the same functionality as the table in the Repository. By default, it displays the following information for each file:

- **Access**: Displays whether the file is open or controlled access. Users must login to the GDC Portal and have the appropriate credentials to access these files.
- **File Name**: Name of the file. Clicking the link will bring the user to the File Summary Page.
- **Cases**: The number of cases associated with the file
- **Project**: The Project that the file belongs to. Clicking the link will bring the user to the Project Summary Page.
- **Data Category**: Type of data
- **Data Format**: The file format
- **File Size**: The size of the file
- **Annotations**: Whether there are any annotations

Additional information can be displayed using the Customize Columms button above the table. Sort can be applied by clicking on the table headers, and the search bar provides additional options for locating specific files. Details of the files can be downloaded using the JSON and TSV buttons above the table.

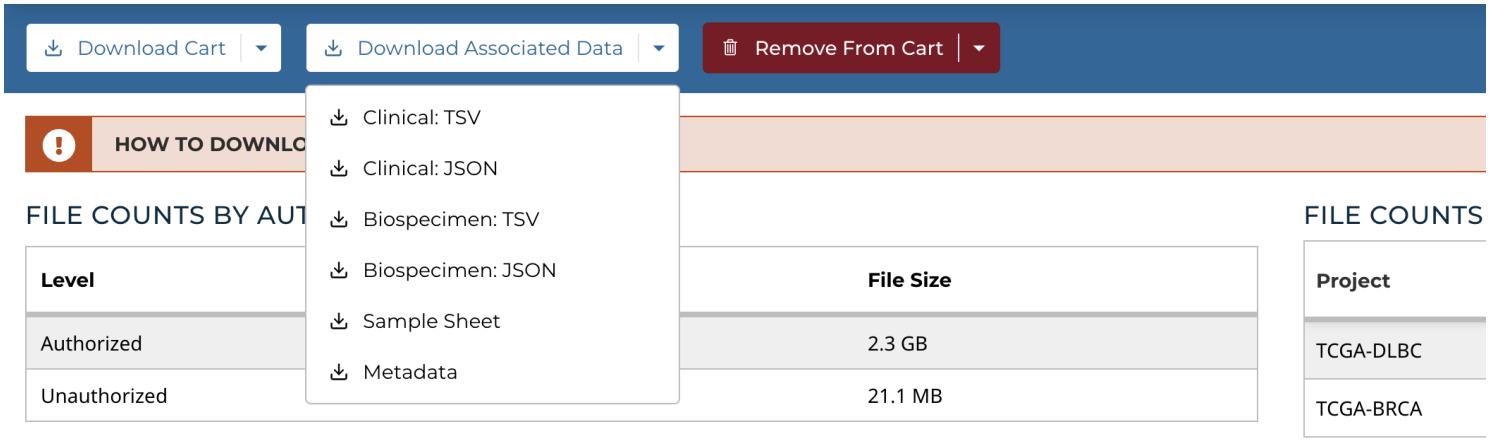
Downloading Files from the Cart

To download files in the Cart, select the Download Cart button and choose either:

- **Manifest**: Downloads a manifest for the files that can be passed to the GDC Data Transfer Tool. A manifest file contains a list of the UUIDs that correspond to the files in the cart.
- **Cart**: Download the files directly through the browser. Users have to be cautious of the amount of data in the cart since this option will not optimize bandwidth and will not provide resume capabilities. This option can only be used if the total size of the files in the Cart does not exceed 5 GB.

Additional Data Download

Additional data can be downloaded from the Cart page using the Download Associated Data button at the top of the page and choosing one of the available options.



- **Clinical:** TSV / Clinical: JSON - This includes all clinical information from the cases that are associated with the files (available as TSV or JSON)
- **Biospecimen:** TSV / Biospecimen: JSON - This includes all biospecimen information from the cases that are associated with the files (available as TSV or JSON).
- **Sample Sheet:** A TSV with commonly-used elements associated with each file, such as sample barcode and sample type.
- **Metadata:** This includes all of the metadata associated with each and every file in the cart. Note that this file is only available in JSON format and may take several minutes to download.

File Summary Page

Clicking on a file name, in the tables that appear on both the Repository and Cart pages, launches the File Summary Page. Each File Summary Page provides information about a data file, such as size, MD5 checksum, and data format; information on the type of data included; links to the associated cases and biospecimen; and information about how the data file was generated or processed.

The page also includes buttons to download the file, add it to the file cart, or (for BAM files) utilize the BAM slicing function.

FILE PROPERTIES

Name	365c5735-dce4-412a-803e-03e43ec829a5.rna_seq.chimeric.gdc_realn.bam
Access	controlled
UUID	0600cda9-b59c-4344-8a10-d68de17505af
Data Format	BAM
Size	254.87 MB
MD5 Checksum	bdfe31616d88e65e23ca9e7b78364b17
Project	TCGA-STAD

DATA INFORMATION

Data Category	Sequencing Reads
Data Type	Aligned Reads
Experimental Strategy	RNA-Seq
Platform	Illumina

In the lower section of the screen, the following tables provide more details about the file and its characteristics:

- **Associated Cases/Biospecimen**: List of cases or biospecimen the file is directly attached to
- **Analysis and Reference Genome**: Information on the workflow and reference genome used for file generation
- **Read Groups**: Information on the read groups associated with the file
- **Metadata Files**: Experiment metadata, run metadata, and analysis metadata associated with the file
- **Downstream Analysis Files**: List of downstream analysis files
- **File Versions**: List of all versions of the file

Chapter 6

Projects

Projects

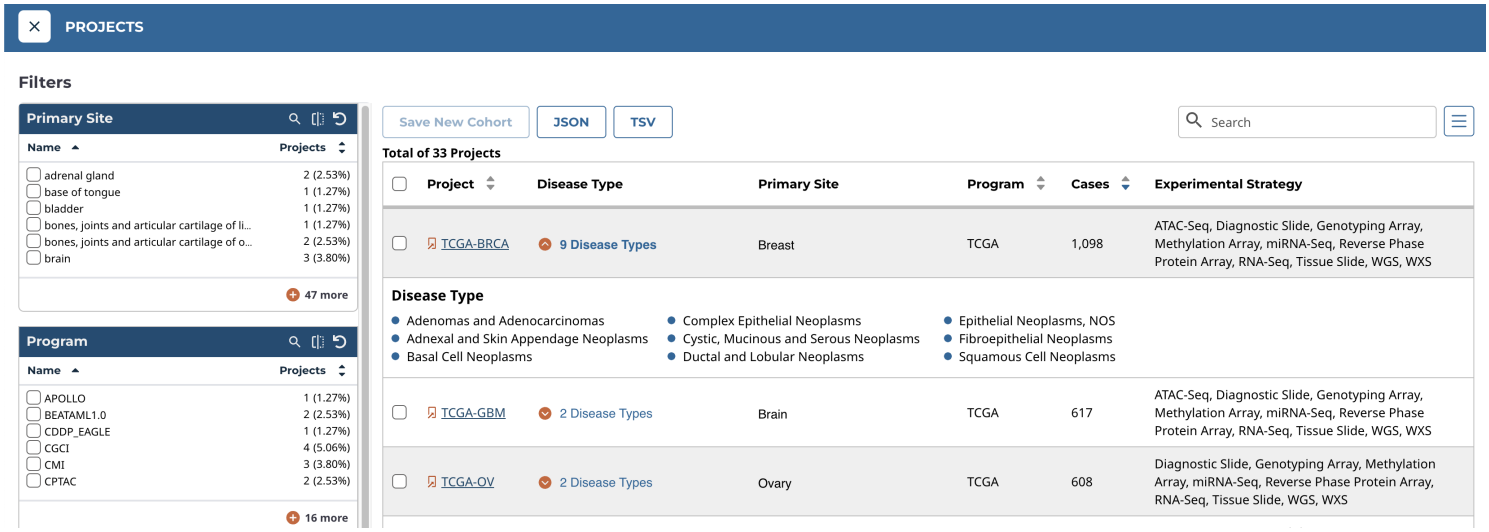
At a high level, data in the Genomic Data Commons is organized by project. Typically, a project is a specific effort to study a particular type(s) of cancer undertaken as part of a larger cancer research program. The GDC Data Portal allows users to access aggregate project-level information via the Projects tool and Project Summary Pages.

Projects Tool

The Projects tool provides an overview of all harmonized data available in the GDC, organized by project. It also provides filtering, navigation, and advanced visualization features that allow users to identify and browse projects of interest. Users can access the Projects tool from the GDC Data Portal header.

On the left, a panel of facets allows users to apply filters to find projects of interest. When filters are applied, the table on the right is updated to display only the matching projects. When no filters are applied, all projects are displayed.

The right side of the Projects tool displays a table that contains a list of projects and specific details about each project, such as the number of cases, types of diseases and primary sites, the program involved, and the experimental strategies available. When a project contains more than one value for the disease type and primary site properties, the full list of values can be expanded by choosing the drop down icon next to the name of the property.



Cohorts can be created by selecting individual projects and using the Save New Cohort button above the table. The checkbox in the header allows all projects on the current page of the table to be selected at the same time.

Facets Panel

Facets represent properties of the data that can be used for filtering. The facets panel on the left allows users to filter the projects presented in the Table.

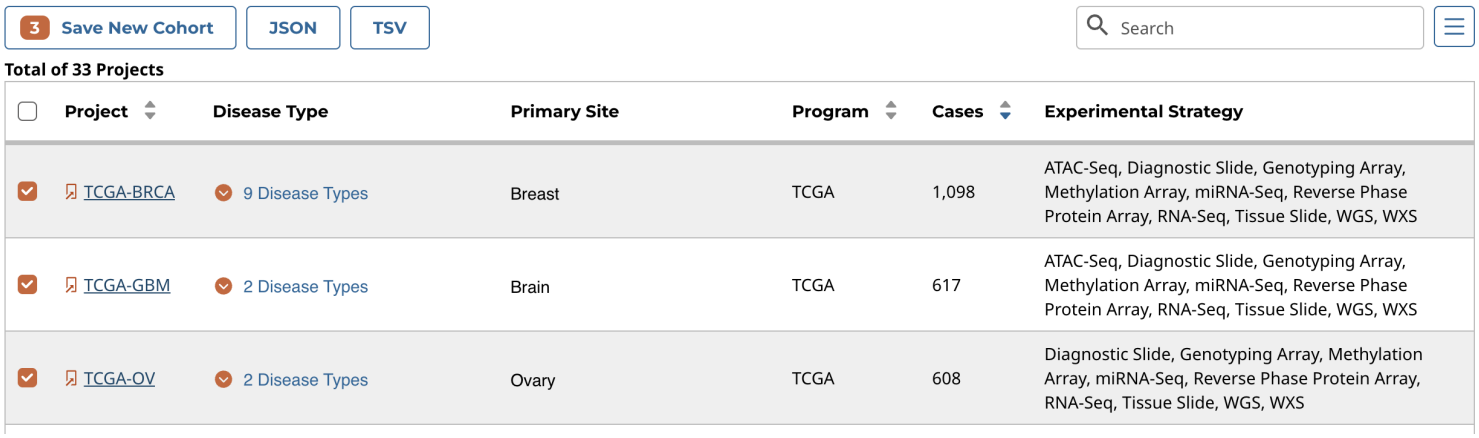
Users can filter by the following facets:

- **Primary Site**: Anatomical site of the cancer under investigation or review
- **Program**: Research program that the project is part of
- **Disease Type**: Type of cancer studied
- **Data Category**: Type of data available in the project
- **Experimental Strategy**: Experimental strategies used for molecular characterization of the cancer

Filters can be applied by selecting values of interest in the available facets, for example “WXS” and “RNA-Seq” in the “Experimental Strategy” facet, and “Brain” in the “Primary Site” facet. When facet filters are applied, the Table is updated to display matching projects.

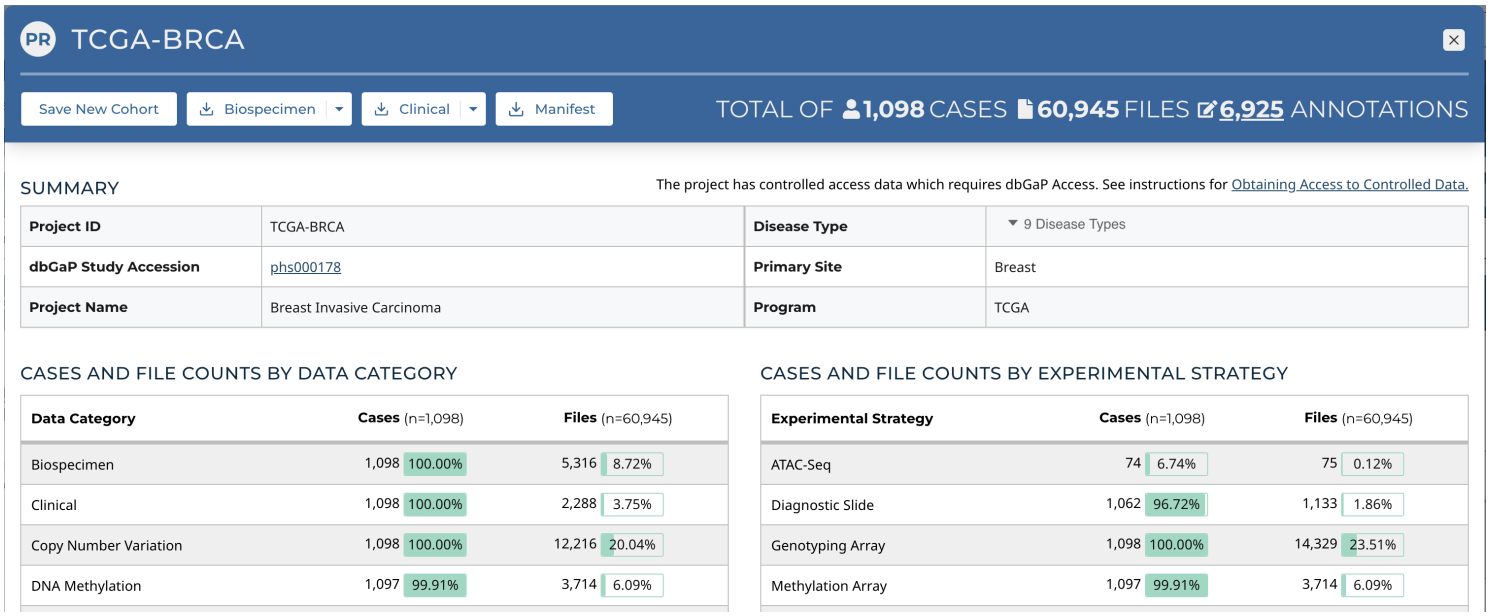
Creating Cohorts From Selected Projects

Custom cohorts consisting of specific projects can be created by selecting those projects in the table using the check boxes next to the project names and clicking the “Save New Cohort” button above the table.



Project Summary Page

Clicking the link for each project name on the table will bring users to that specific project's summary page. This page contains basic information about the contents of a project as well as the percentages of cases within the project that contain a specific experimental strategy or data category.



PR TCGA-BRCA

Save New Cohort | Biospecimen | Clinical | Manifest

TOTAL OF 1,098 CASES 60,945 FILES 6,925 ANNOTATIONS

SUMMARY

The project has controlled access data which requires dbGaP Access. See instructions for Obtaining Access to Controlled Data.

Project ID	TCGA-BRCA	**Disease Type**	9 Disease Types
dbGaP Study Accession	phs000178	**Primary Site**	Breast
Project Name	Breast Invasive Carcinoma	**Program**	TCGA

CASES AND FILE COUNTS BY DATA CATEGORY

Data Category	Cases (n=1,098)	Files (n=60,945)
Biospecimen	1,098 100.00%	5,316 8.72%
Clinical	1,098 100.00%	2,288 3.75%
Copy Number Variation	1,098 100.00%	12,216 20.04%
DNA Methylation	1,097 99.91%	3,714 6.09%

CASES AND FILE COUNTS BY EXPERIMENTAL STRATEGY

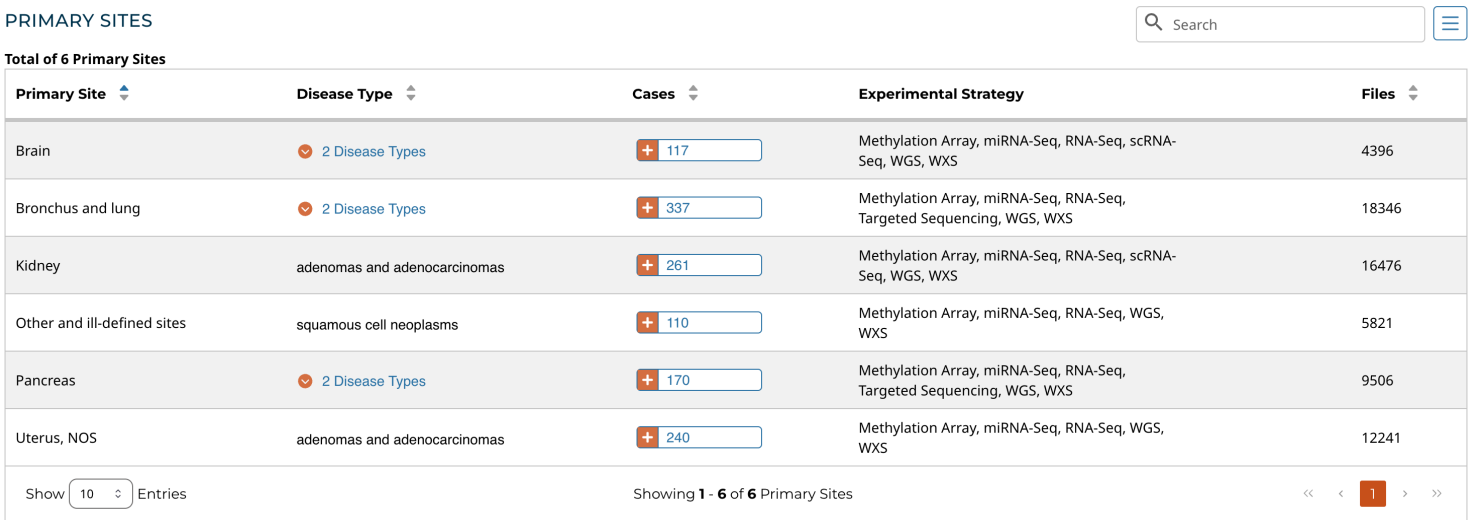
Experimental Strategy	Cases (n=1,098)	Files (n=60,945)
ATAC-Seq	74 6.74%	75 0.12%
Diagnostic Slide	1,062 96.72%	1,133 1.86%
Genotyping Array	1,098 100.00%	14,329 23.51%
Methylation Array	1,097 99.91%	3,714 6.09%

Four buttons on the left of the header allow the user to perform a variety of actions related to the project:

- **Save New Cohort**: Creates a new cohort consisting of all the cases in the project
- **Biospecimen**: Downloads biospecimen metadata associated with all cases in the project in either TSV or JSON format
- **Clinical**: Downloads clinical metadata about all cases in the project in either TSV or JSON format
- **Manifest**: Downloads a manifest for all data files available in the project. The manifest can be used with the GDC Data Transfer Tool to download the files.

Primary Sites Table

Summary pages for projects with multiple primary sites also include a Primary Sites table. Each row of the table contains information relevant to a specific primary site within the project, and additional cohorts can be created using buttons located within the table.



PRIMARY SITES

Search

Total of 6 Primary Sites

Primary Site	Disease Type	Cases	Experimental Strategy	Files
Brain	2 Disease Types	117	Methylation Array, miRNA-Seq, RNA-Seq, scRNA-Seq, WGS, WXS	4396
Bronchus and lung	2 Disease Types	337	Methylation Array, miRNA-Seq, RNA-Seq, Targeted Sequencing, WGS, WXS	18346
Kidney	adenomas and adenocarcinomas	261	Methylation Array, miRNA-Seq, RNA-Seq, scRNA-Seq, WGS, WXS	16476
Other and ill-defined sites	squamous cell neoplasms	110	Methylation Array, miRNA-Seq, RNA-Seq, WGS, WXS	5821
Pancreas	2 Disease Types	170	Methylation Array, miRNA-Seq, RNA-Seq, Targeted Sequencing, WGS, WXS	9506
Uterus, NOS	adenomas and adenocarcinomas	240	Methylation Array, miRNA-Seq, RNA-Seq, WGS, WXS	12241

Show 10 Entries

Showing 1 - 6 of 6 Primary Sites

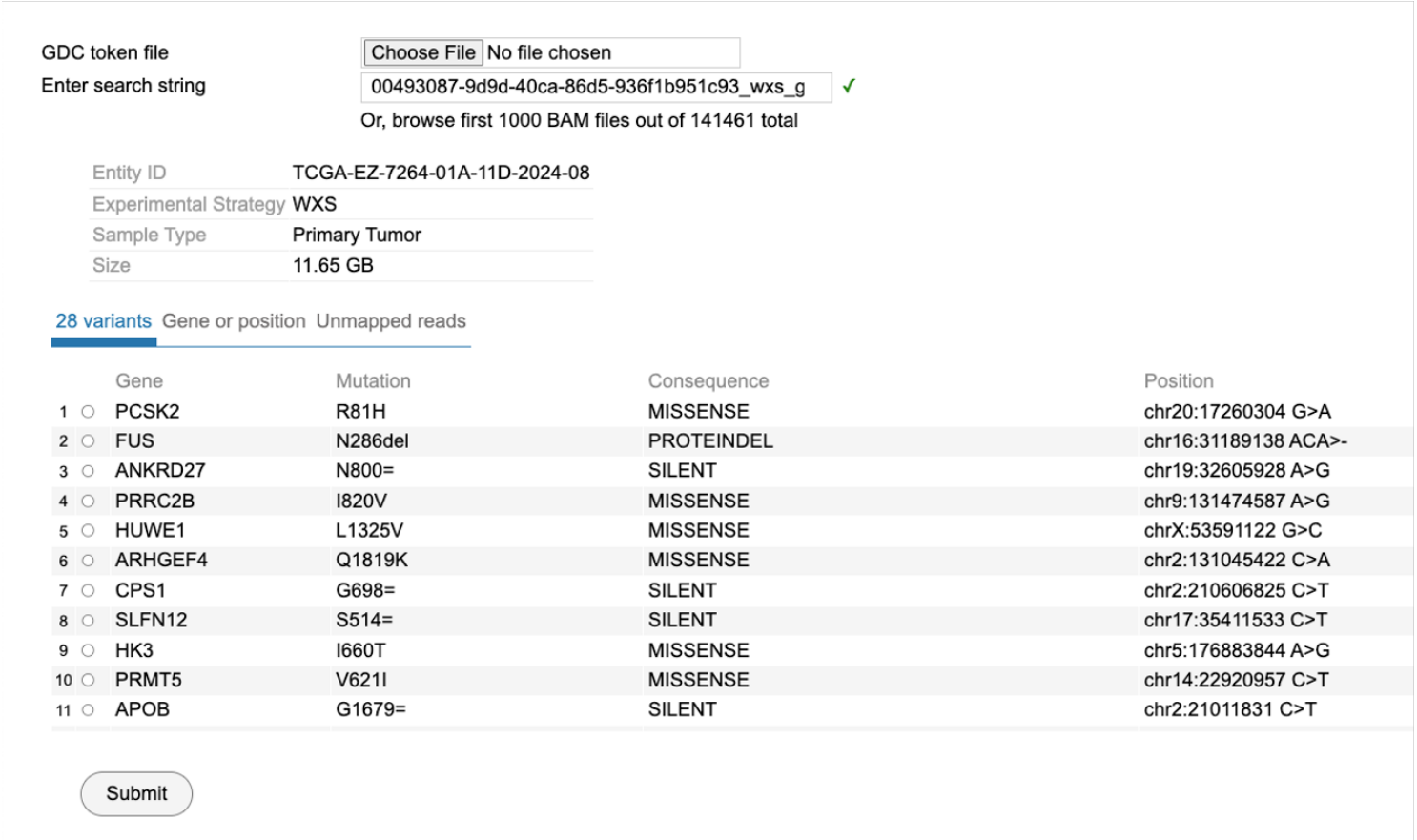
Chapter 7

BAM Slicing

GDC BAM Slicing Download User Guide

Introduction to BAM slicing

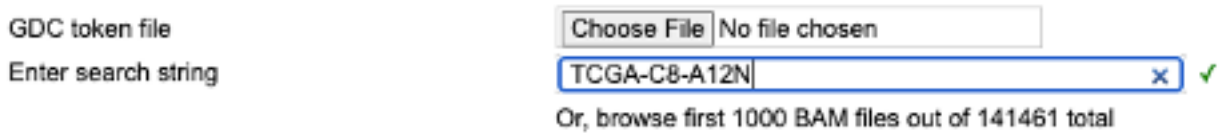
The GDC BAM slicing download feature is a tool for slicing individual BAM files based on the variant, gene, position, or SNPs for individual case/entities in the NCI GDC. In addition, it also allows users to download unmapped reads from a BAM file. To begin, follow the following steps as outlined.



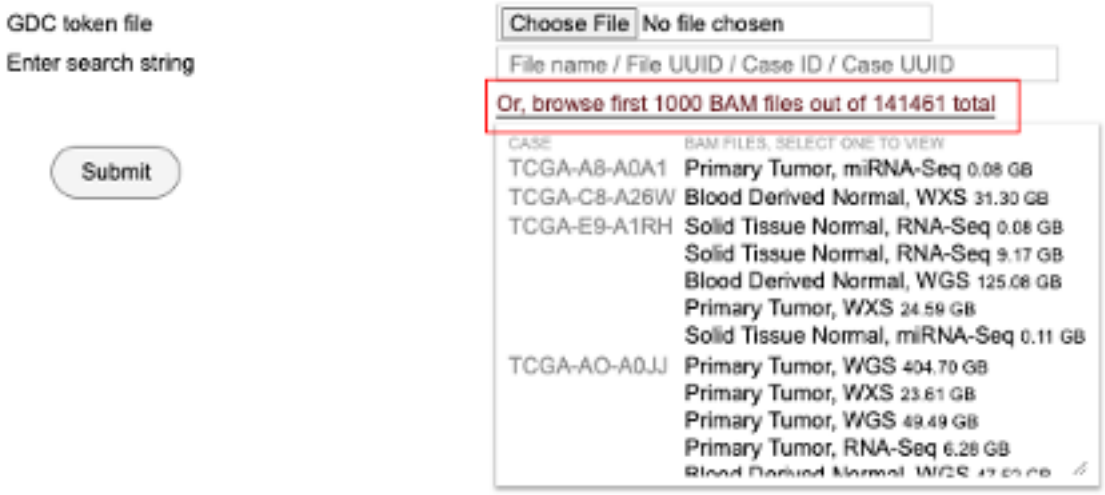
Download

Searching for a case

To search for a case, enter a string id which could be a file name, file id, case UUID or case ID. For example, enter and search for the case 'TCGA-C8-A12N' as shown. Please note that the complete id must be used. Partial ids are not allowed.

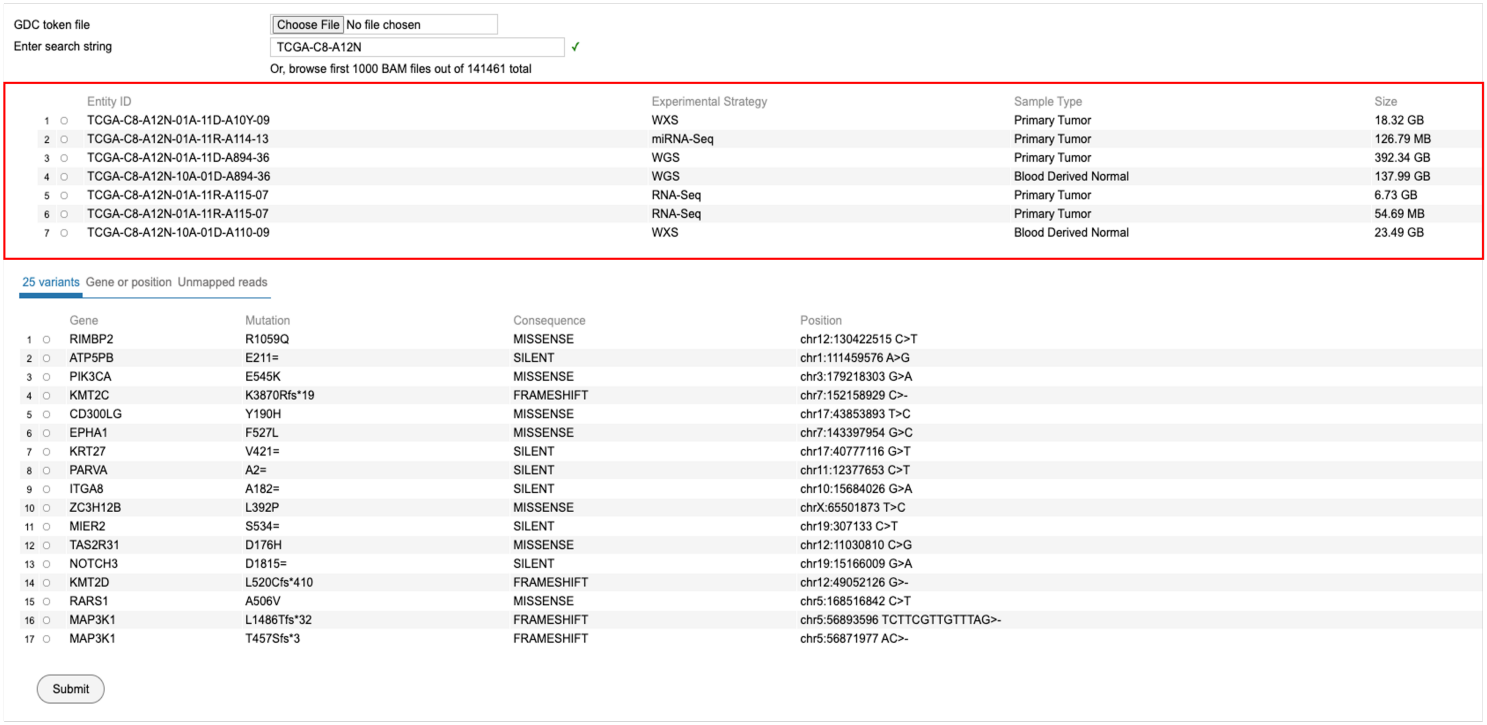


A user may also choose to browse the first thousand BAM files. Click on the label "Or, browse first 1000 BAM files out of 141461 total" to load the following view. Scroll and select the case of interest.



Selecting Variant/Gene, Position or Unmapped reads

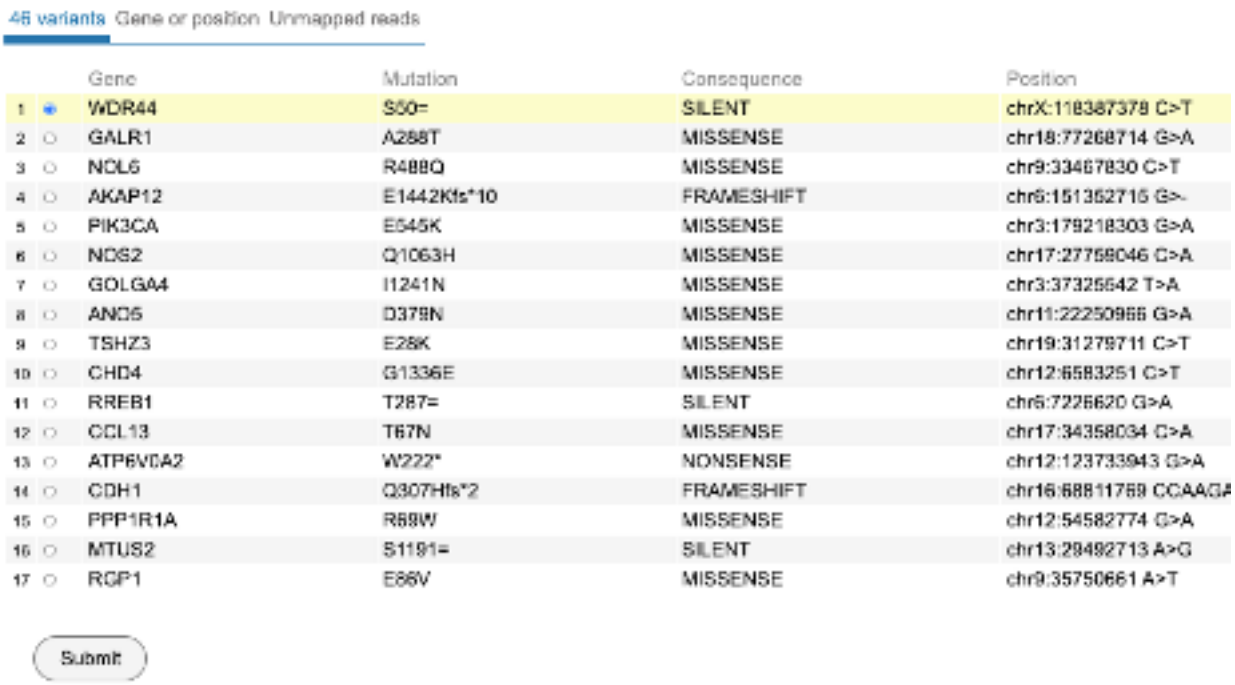
Upon searching by a string id, the following view is displayed. User must select an Entity ID associated with the case ID as shown. This view shows the 'Experimental strategy', 'Sample Type', and 'Size' of the file associated with that entity id.



Choosing a BAM file directly from the thousand files will display the following view. This view has selected an entity id chosen by the user with 'miRNA-Seq' as the experimental strategy, sample type being the 'Primary Tumor' with a file size of 83.03 MB.

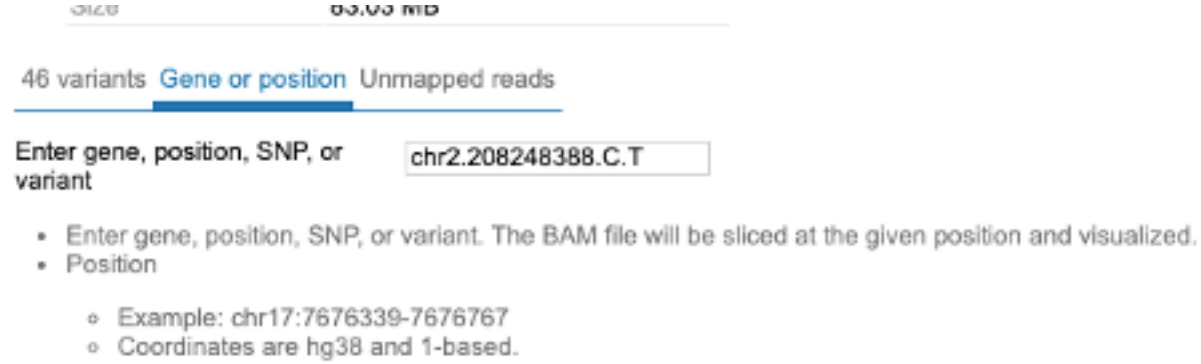
Selecting a Variant

From the view as shown above, user can choose from 48 variants. Select WDR44 as shown and click 'Submit' to download the BAM slice for this case and gene variant.



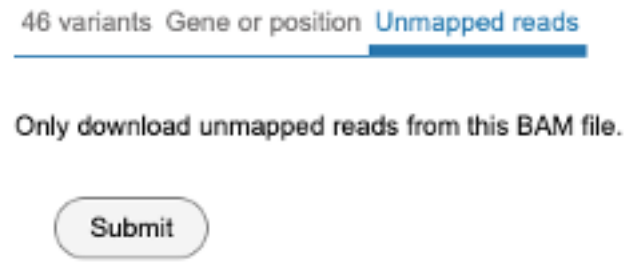
Selecting a Gene

Click on the next tab to access the view that allows selecting BAM files for a particular gene, snp or a specific position/range in the genome. After making your selection, click the 'Submit' button at the bottom of the view to download the slices.



Selecting unmapped reads

Click the tab for accessing the 'Unmapped reads' as shown below. Click 'Submit' to download the unmapped reads.



Saved Downloads

All downloads are saved in the ‘Downloads’ folder.

Chapter 8

Clinical Data Analysis

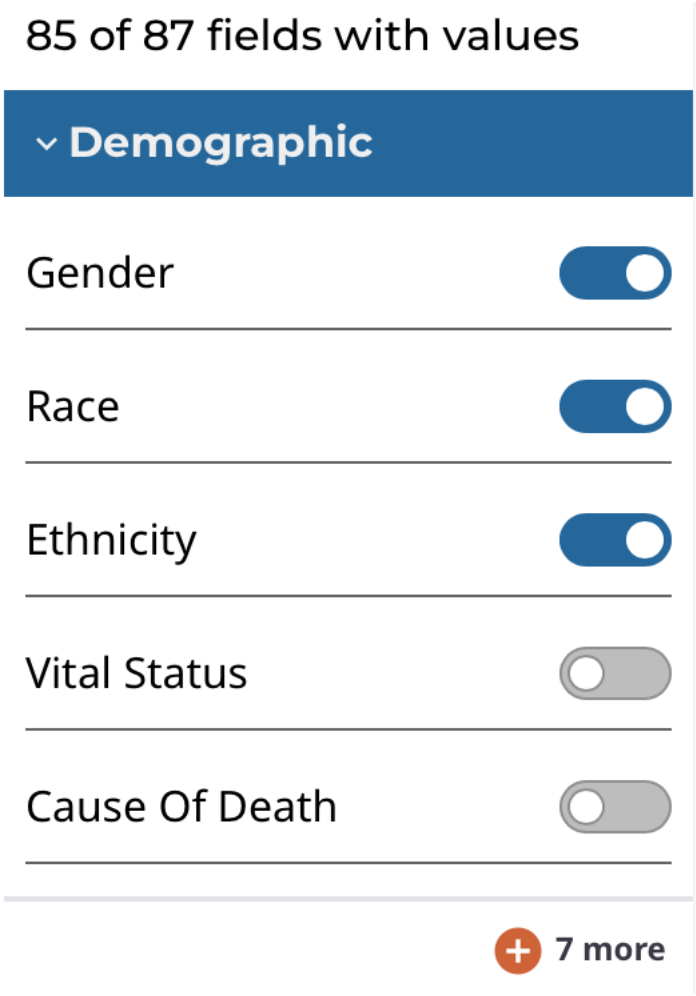
Clinical Data Analysis

The Clinical Data Analysis tool allows for a set of customizable charts to be generated for a set of clinical attributes. Users can select which clinical fields they want to display and visualize the data using various supported plot types. The clinical analysis features include:

- Ability to select which clinical fields to display
- Examine the clinical data of each field using these visualizations:
  - Histogram
  - Survival Plot
  - Box and QQ Plots
- Create custom bins for each field and re-visualize the data with those bins
- Select specific cases from a clinical field and use them to create a new cohort, or modify/remove from an existing cohort
- Download the visualizations of each plot type for each variable in SVG or PNG
- Download the data table of each field in JSON or TSV format
- Print all clinical variable cards in the analysis with their active plot to a single PDF

Enabling Clinical Variable Cards

- In the Analysis Center, select the *Clinical Data Analysis* tool card.
- In the Clinical Data Analysis tool, use the control panel on the left side of the analysis to display which clinical variables you want. To enable or disable specific variables for display, click the on/off toggle controls:



The clinical fields are grouped into these categories:

- **Demographic:** Data for the characterization of the patient by means of segmenting the population (e.g. characterization by age, sex, race, etc.).
- **Diagnosis:** Data from the investigation, analysis, and recognition of the presence and nature of the disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation.
- **Treatment:** Records of the administration and intention of therapeutic agents provided to a patient to alter the course of a pathologic process.
- **Exposure:** Clinically-relevant patient information not immediately resulting from genetic predispositions.

Exploring Clinical Card Visualizations

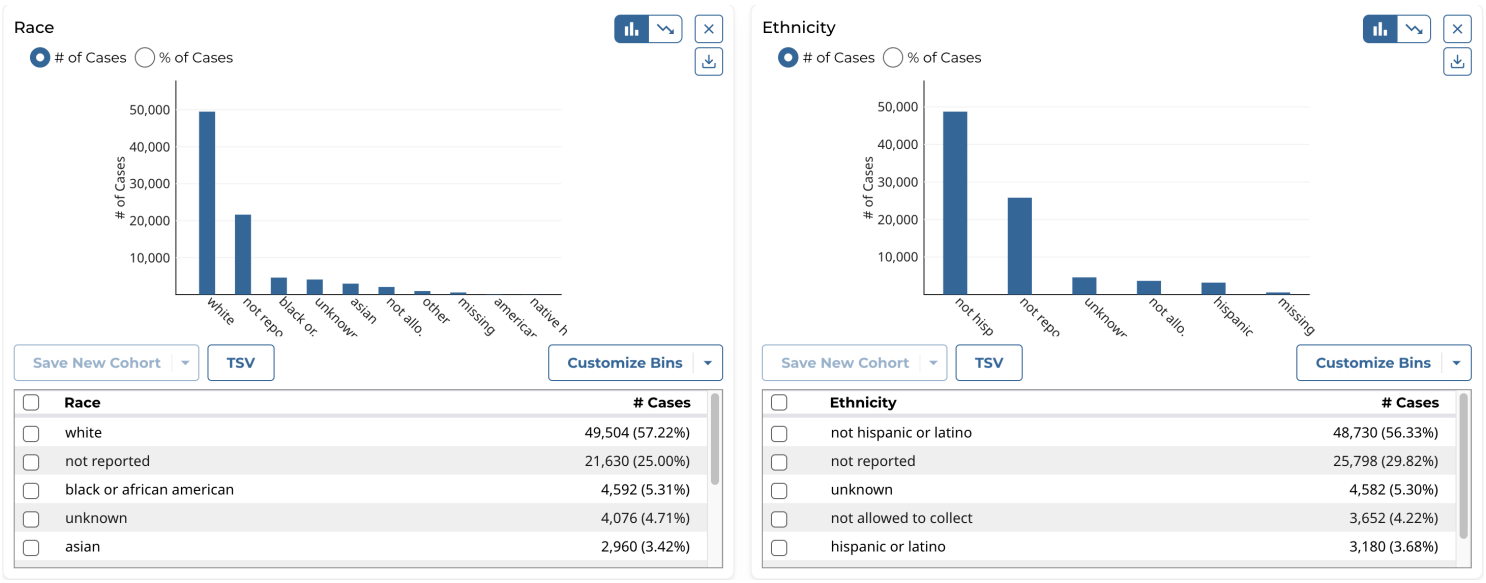
Users can explore different visualizations for each clinical field they have enabled for display. All cards support histograms and survival plots. Additionally, continuous variables can be graphically represented as box and QQ plots. To switch between plot types, click the different plot type icons in the top-right of each card.

Histogram

The histogram plot type supports these features:

- View the distribution of cases (and % of cases) in the cohort for the clinical field's data categories as a histogram
- View the distribution of cases in tabular format
- Select the cases for specific data categories to create new cohorts, append to existing cohorts, or remove from existing cohorts
- Download the histogram visualization in SVG or PNG format

- Download the raw data used to generate the histogram in JSON format



Note that the histogram plot applies to, and can be displayed for, both categorical and continuous variables.

Survival Plot

The survival analysis is used to analyze the occurrence of event data over time. In the GDC, survival analysis is performed on the mortality of the cases. Thus, the values are retrieved from GDC Data Dictionary properties and a survival analysis requires the following fields:

- Data on the time to a particular event (days to death or last follow up).
  - Fields: **demographic.days_to_death** or **demographic.days_to_last_follow_up**
- Information on whether the event has occurred (alive/deceased).
  - Fields: **demographic.vital_status**
- Data split into different categories or groups (i.e. gender, etc.).
  - Fields: **demographic.gender**

The survival analysis in the GDC uses a Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

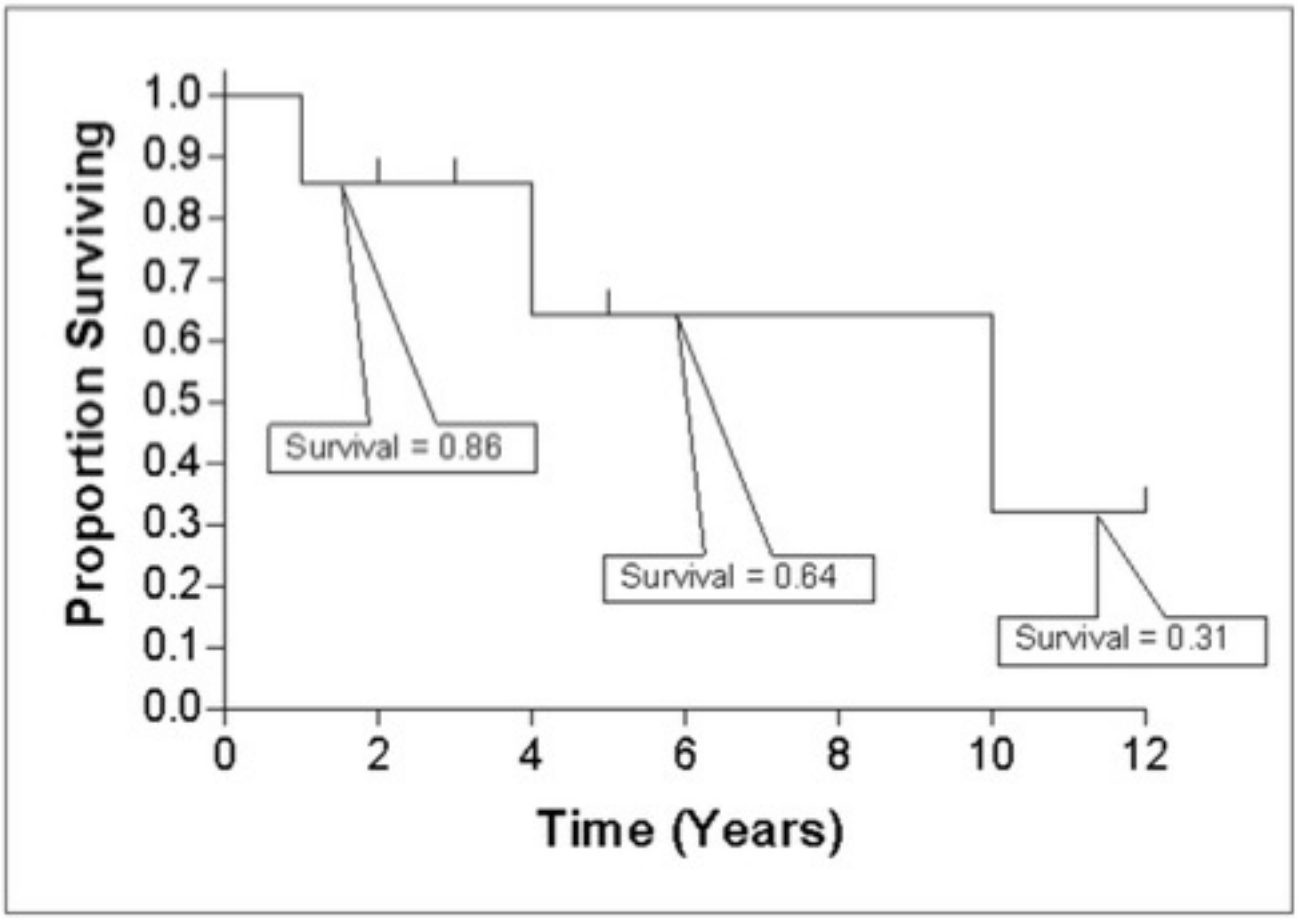
Where:

- S(t) is the estimated survival probability for any particular one of the t time periods.
- ni is the number of subjects at risk at the beginning of time period ti.
- and di is the number of subjects who die during time period ti.

The table below is an example data set to calculate survival for a set of seven cases:

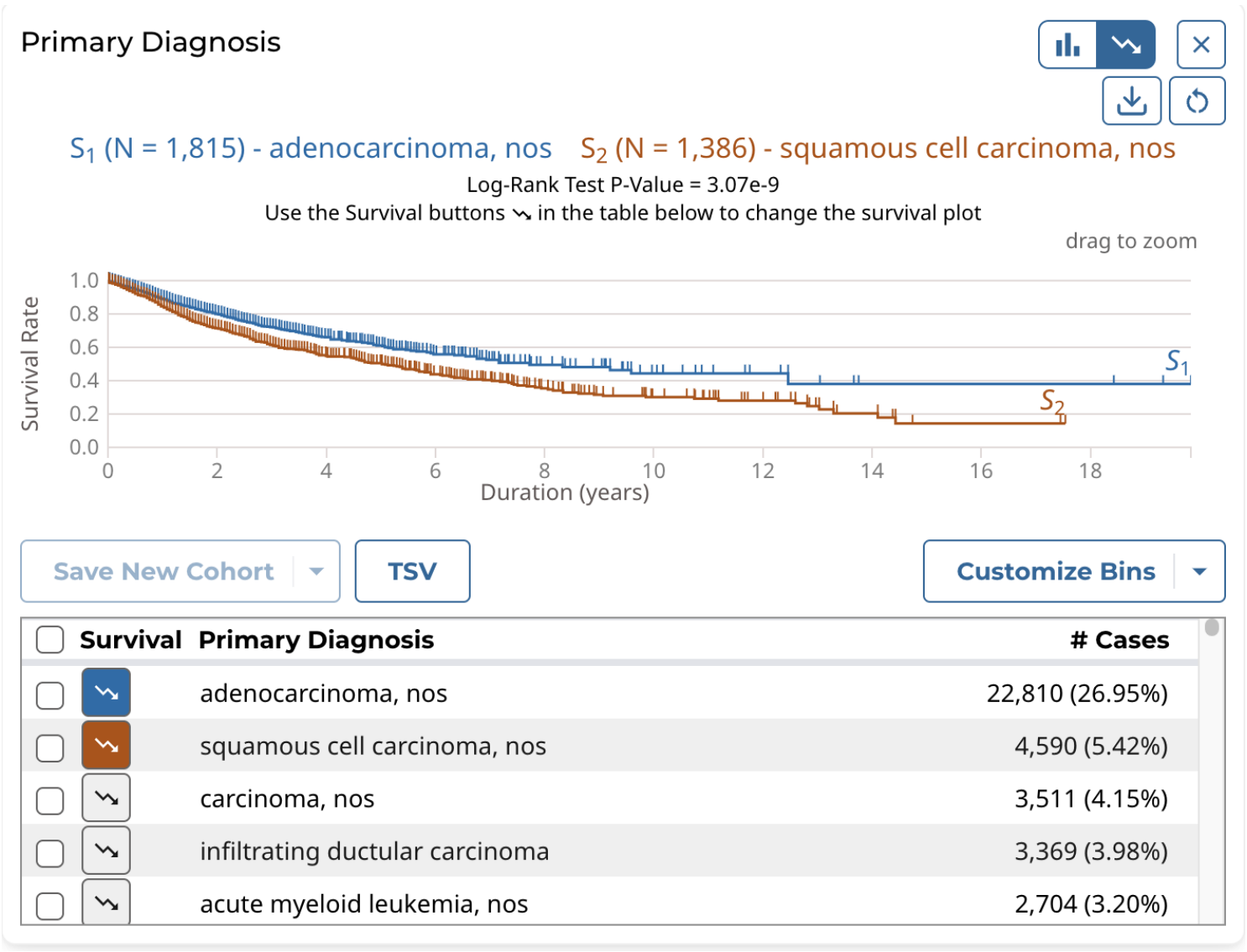
overall_survial_time (Years)	interval		# of donors at risk at start of interval (r)	# of censored donors during interval (c)	# of donors at risk at end of interval (n=r-c)	# of donors died at end of interval (d)	estimated interval survival probability ((n-d)/n)	estimated cumulative survival probability at end of interval (S)
	start	end						
0	0							1
1	0	1	7	0	7	1	(7-1)/7 = 0.86	1 * 0.86 = **0.86**
4	1	4	6	2	4	1	(4-1)/4 = 0.75	0.86 * 0.75 = **0.64**
10	4	10	3	1	2	1	(2-1)/2 = 0.5	0.86 * 0.75 * 0.5 = **0.31**
>12	10	12	1	0	1	0	(1-0)/1 = 1.0	0.86 * 0.75 * 0.5 * 1.0 = **0.31**

The calculated cumulated survival probability can be plotted against the interval to obtain a survival plot like the one shown below.



The survival plot type supports these features:

- View the distribution of cases (and % of cases) in the cohort for the clinical field's data categories as a table.
- Select and plot the survival analysis for the cases of specific data categories in the table:
  - By default the top 2 categories (highest of cases) are displayed.
  - Users can manually select and plot up to 5 categories at a time.
- Download the survival plot visualization in SVG or PNG format
- Download the raw data used to generate the survival plot in JSON or TSV format



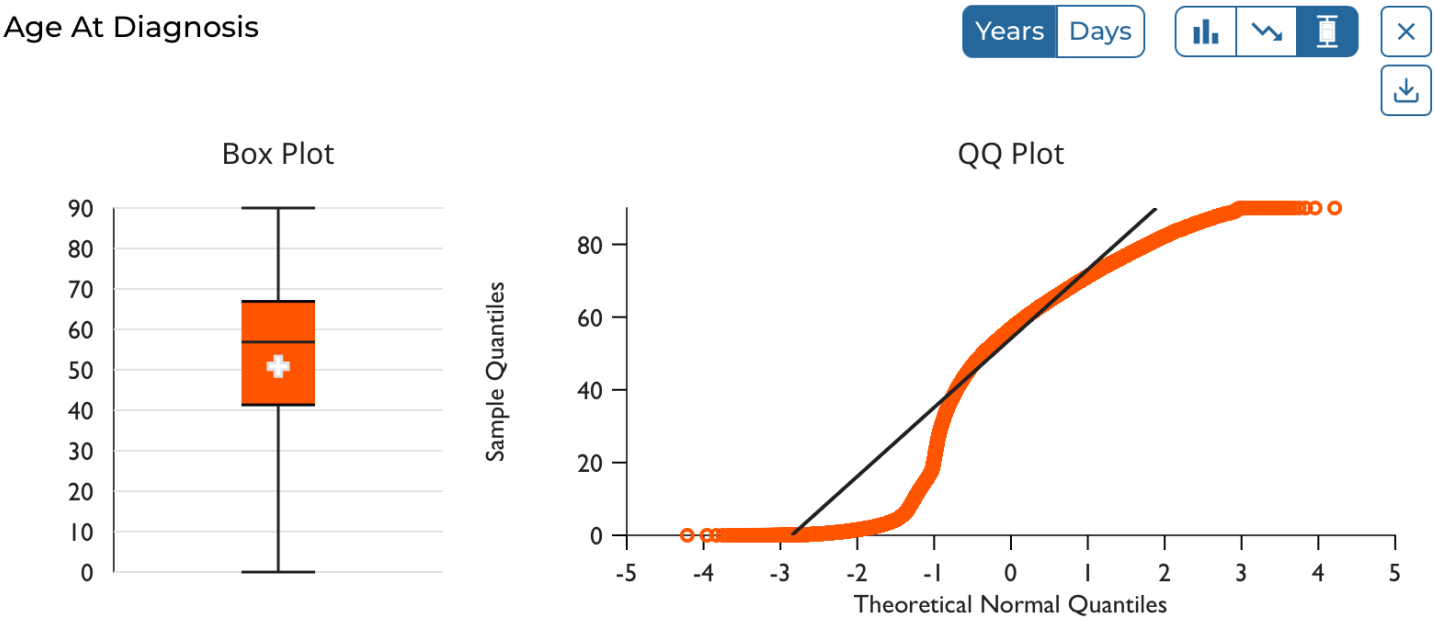
Note that the survival plot applies to, and can be displayed for, both categorical and continuous variables.

Box and QQ Plots

The box and QQ plot types support these features:

- View the quartiles (Q1, Q2/median, and Q3) as well as the mean, minimum, and maximum values in the cohort for the clinical field as a box plot
- View the descriptive statistics in the cohort for the clinical field in tabular format
- Plot the quantiles of the clinical field's distribution with quantiles of a theoretical normal distribution as a QQ plot
- Download the box and QQ plot visualizations in SVG or PNG format
- Download the raw data used to generate the QQ plot in JSON or TSV format

Age At Diagnosis



TSV

Statistics	Years
Minimum	0
Maximum	90
Mean	50.87
Median	56.87
Standard Deviation	22.62
IQR	25.58

Note that the box and QQ plots apply to, and can be displayed for, continuous variables only.

Certain continuous variables that are measured with units of time, such as Days to Birth, include a toggle to switch between displaying the data in years or days. A standard formula is employed for converting between years and days:

- 1 year = 365.25 days

Creating Custom Bins

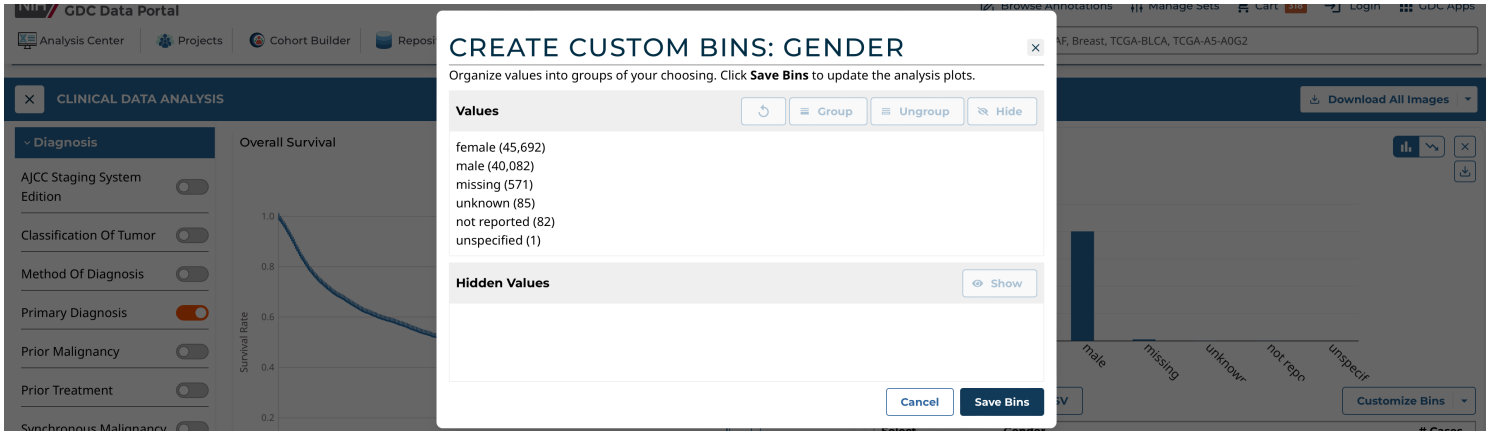
For each clinical variable, whether categorical or continuous, users can create custom bins to group the data in ways they find scientifically interesting or significant. Once saved, the bins are applied to these visualizations and they are then re-rendered:

- Histogram and associated data table
- Survival plot and associated data table

Custom bins can be reset to their defaults at any time for each card by selecting the ***Reset to Default*** option after clicking ***Customize Bins***.

Categorical Binning

To create custom bins for a categorical variable, click ***Customize Bins***, then ***Edit Bins***. A configuration window appears where the user can create their bins:



The user can:

- Group existing individual values into a single group
- Give a custom name to each group
- Ungroup previously grouped values
- Completely hide values from being shown in the visualization
- Re-show previously hidden values

Continuous Binning

To create custom bins for a continuous variable, click ***Customize Bins***, then ***Edit Bins***. A configuration window appears where the user can create their bins:

The user can choose one of these continuous binning methods:

- (1) Create equidistant bins based on a set interval:
  - User must choose the interval (e.g. equidistant bins of 1,825 days for the Age of Diagnosis field)
  - User can optionally define the starting and ending value between which the equidistant bins will be created
- (2) Create completely custom ranges:
  - User manually enters 1 or more bins with custom ranges
  - User must enter a name for each range and the start and end values
  - The ranges can be of different interval lengths

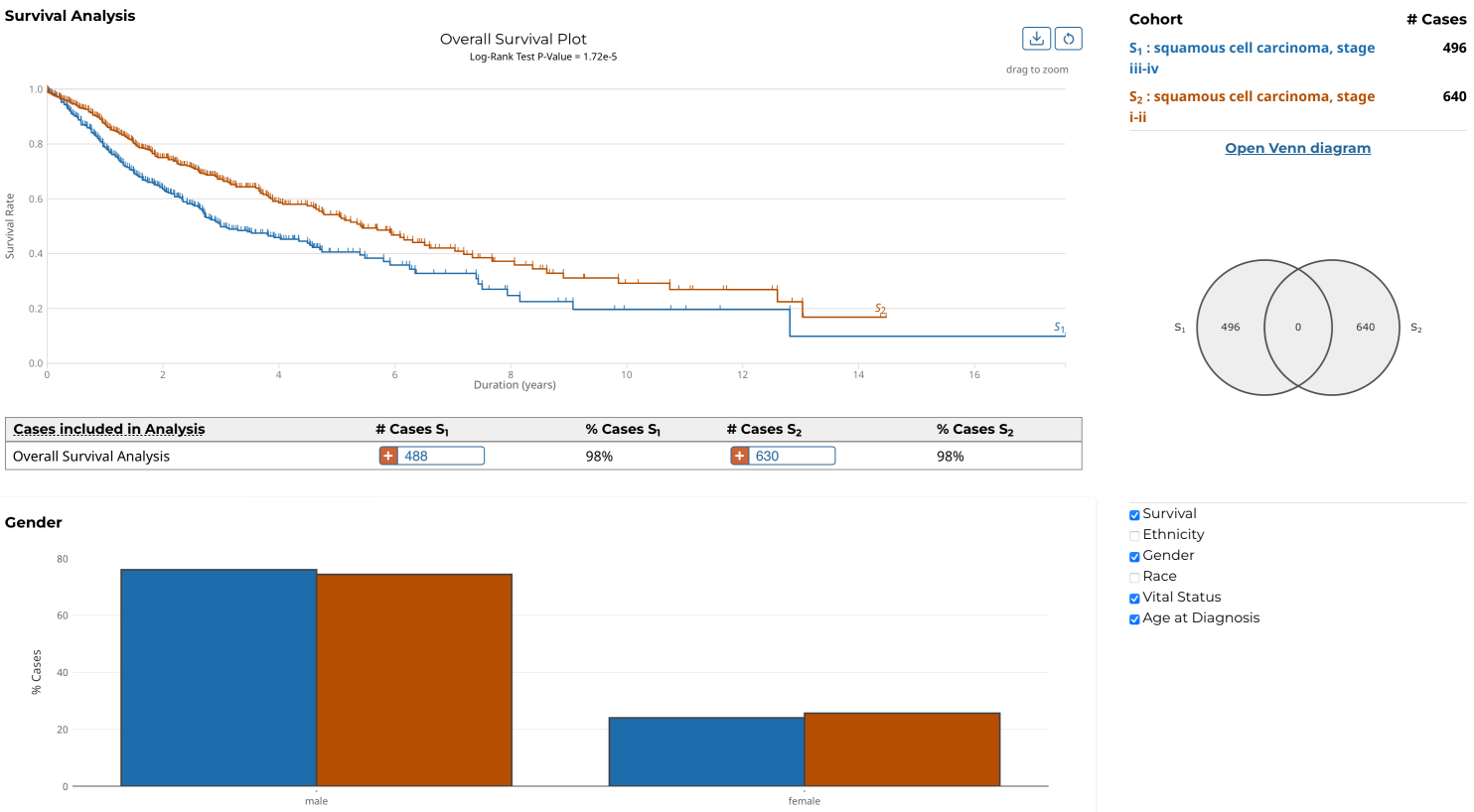
Chapter 9

Cohort Comparison

Cohort Comparison

The Cohort Comparison tool displays graphs and tables that demonstrate the similarities and differences between the active cohort and a different cohort. The following features are displayed for each of the two cohorts:

- A key detailing the number of cases in each cohort and the color that represents each (blue/orange)
- A Venn diagram, which shows the number of cases shared between the cohorts
- A selectable survival plot that compares both sets with information about the percentage of represented cases
- A breakdown of each cohort by selectable clinical facets with a bar graph and table. The facets included are `Vital_Status`, `Gender`, `Race`, `Ethnicity`, and `Age_at_Diagnosis`. A p-value (if it can be calculated from the data) that demonstrates whether the statuses are proportionally represented is displayed for the `Vital_Status`, `Gender`, and `Ethnicity` facets.
- Additional cohorts can be created containing subsets of these two cohorts



Note that clicking the “Open Venn diagram” link will launch the Set Operations tool with the same cohorts used in the Cohort Comparison tool.

Chapter 10

Gene Expression Clustering

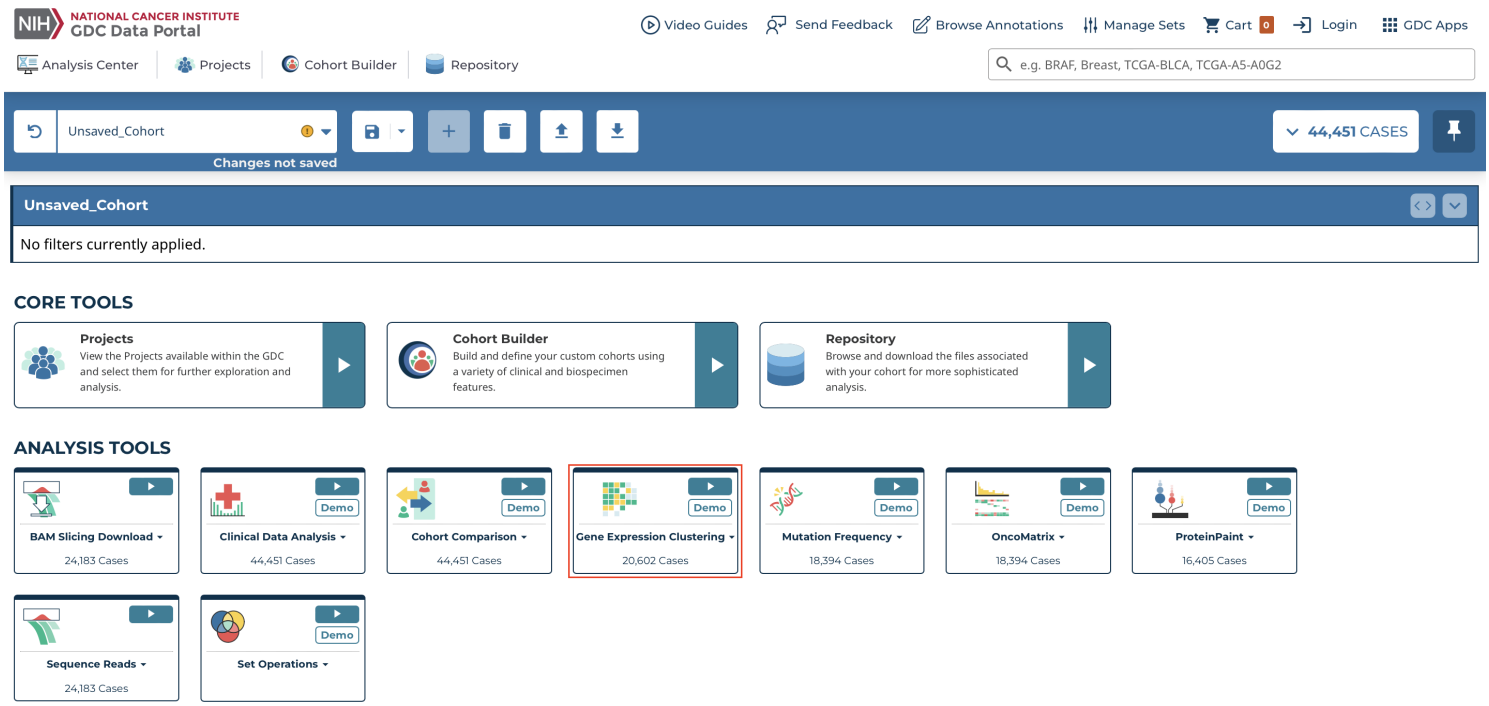
Gene Expression Clustering Tool

Introduction to Gene Expression Clustering

The Gene Expression Clustering tool is a web-based tool for performing sample clustering by selecting a desired set of genes from the NCI Genomic Data Commons (GDC), and visualizing a heatmap of a z-score transformed matrix.

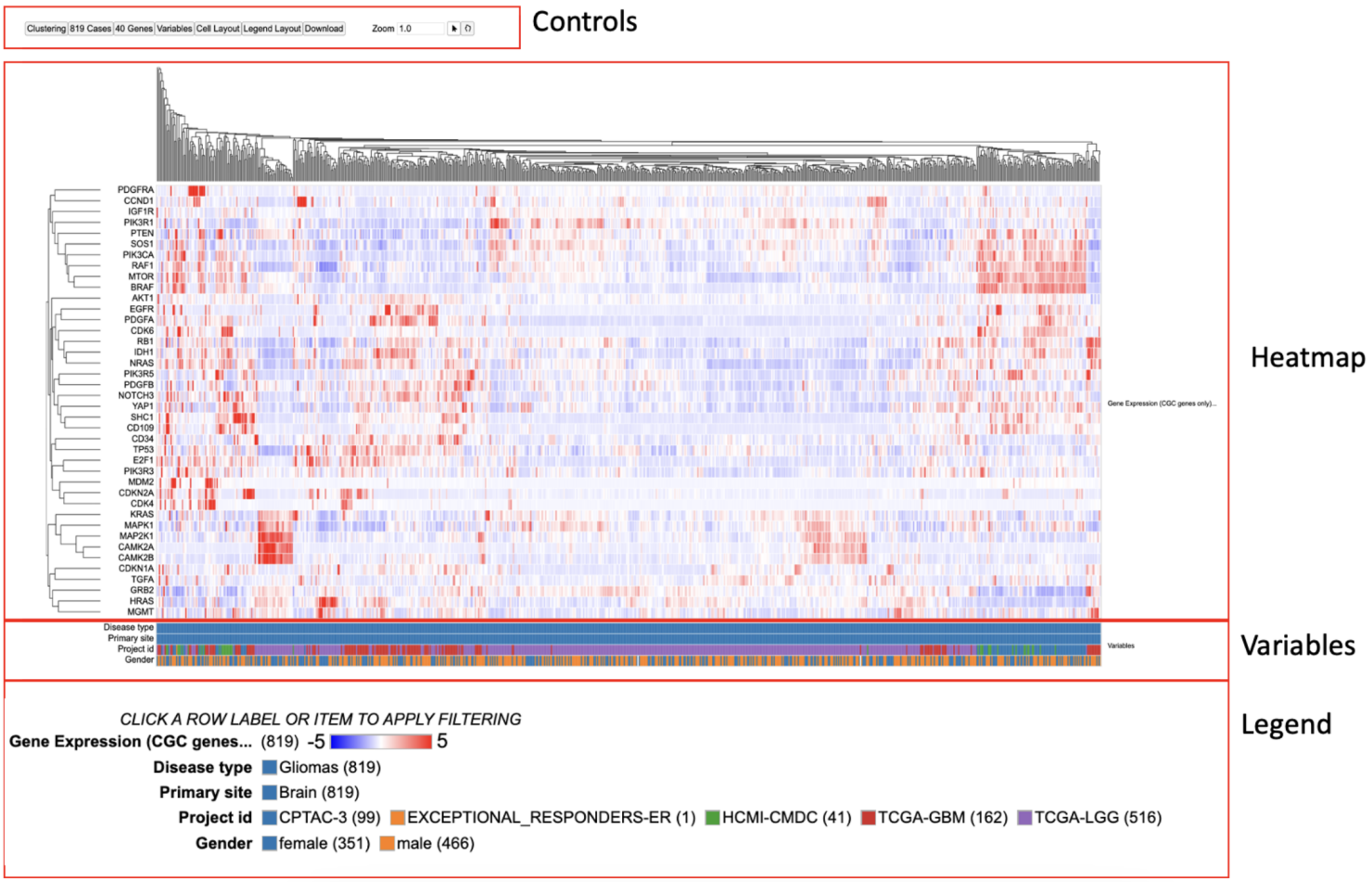
Quick Reference Guide

At the Analysis Center, click the 'Gene Expression Clustering' card to launch the heatmap.



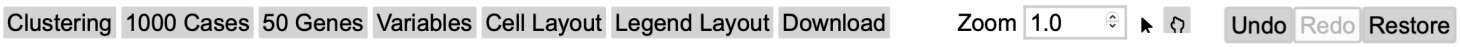
Users can view publicly available genes as well as login with credentials to access controlled data.

There are four main panels in the Gene Expression Clustering tool: controls, heatmap, variables, and legend.



Controls

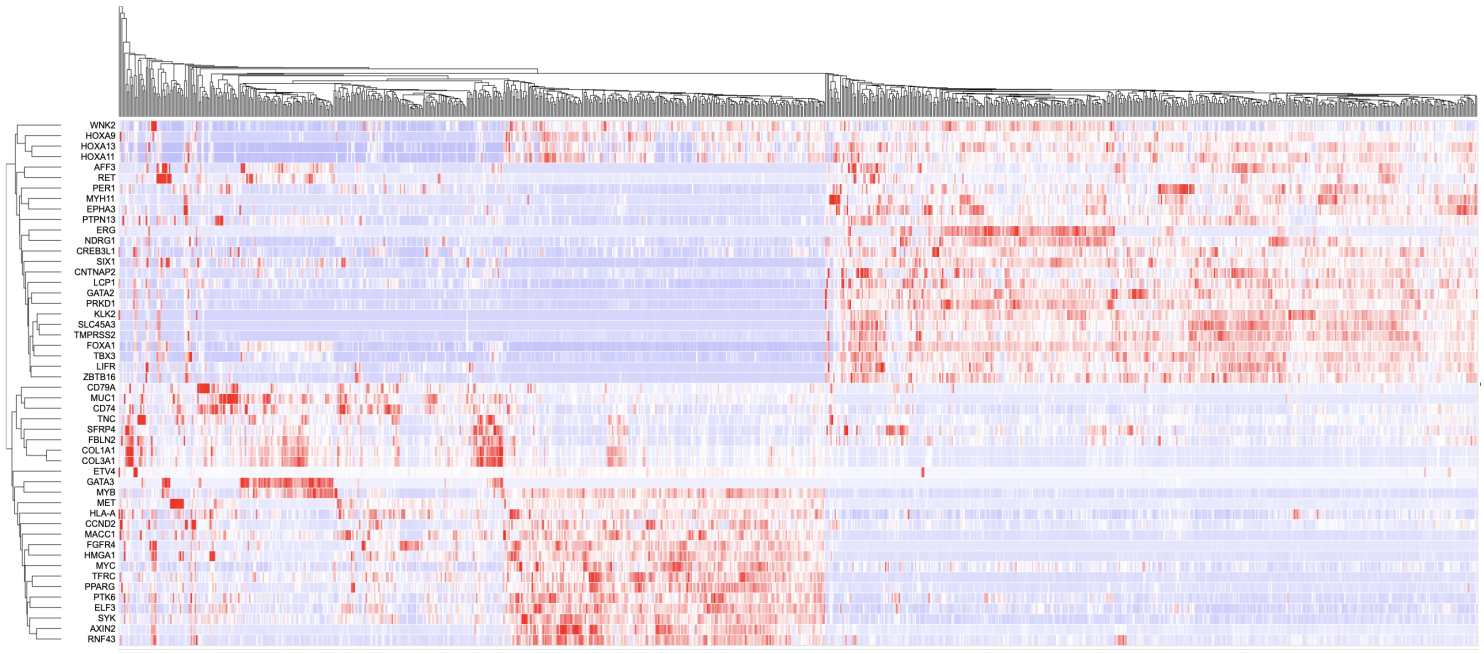
The control panel can modify the displayed data or the appearance of the matrix. Their functionalities are outlined below.



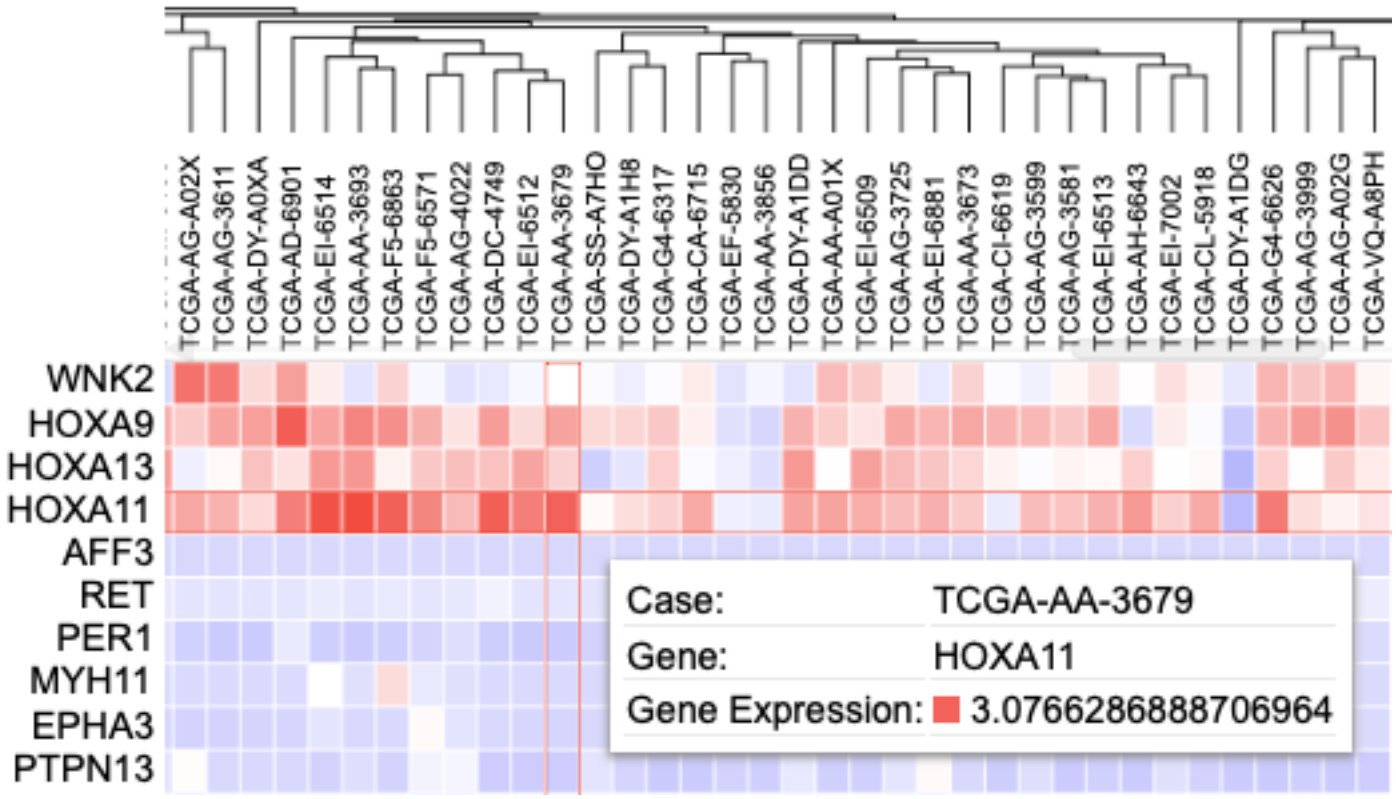
- **Clustering:** Modify the default clustering of the heatmap (Average or Complete), alter the column and row dendrogram dimensions, and change the z-score cap
- **Cases:** Adjust the visible characters of the case labels
- **Genes:** Modify how cases are represented for each gene (Absolute, Percent, or None), row group and label lengths, rendering style, and the existing gene set
  - **Edit Group:** Displays a panel of currently selected genes, which can be modified by clicking on a gene to remove it from the gene set, searching for a particular gene to add, loading top variably expressed genes, or loading a pre-defined gene set provided by the MSigDB database
  - **Create Group:** Create a new gene set by searching for a particular gene, loading top mutated genes, or loading a pre-defined gene set provided by the MSigDB database
- **Variables:** Search and select variables to add to the matrix below the heatmap
- **Cell Layout:** Modify the format of the cells by changing colors, cell dimensions, and label formatting
- **Legend Layout:** Alter the legend by changing the font size, dimensions, and other formatting preferences
- **Download:** Download the plot in svg format
- **Zoom:** Adjust the zoom level by using the up and down arrows on the input box, entering a number, or using the sliding scale to view the case labels

Heatmap

The Gene Expression Clustering heatmap displays the active cohort's cases along the top horizontally, genes along the left column, and the z-score transformed gene expression value.



Hovering over a cell in the heatmap displays the case submitter_id, gene name, and gene expression value.

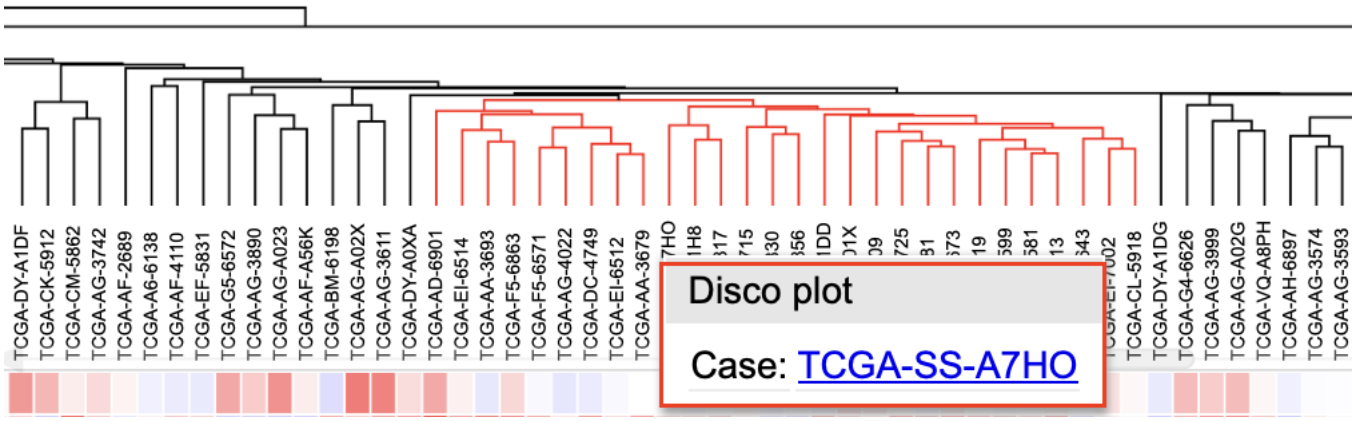


Clicking on a cell also gives users the option to launch the Disco plot, a circos plot displaying copy number data and consequences for that case.

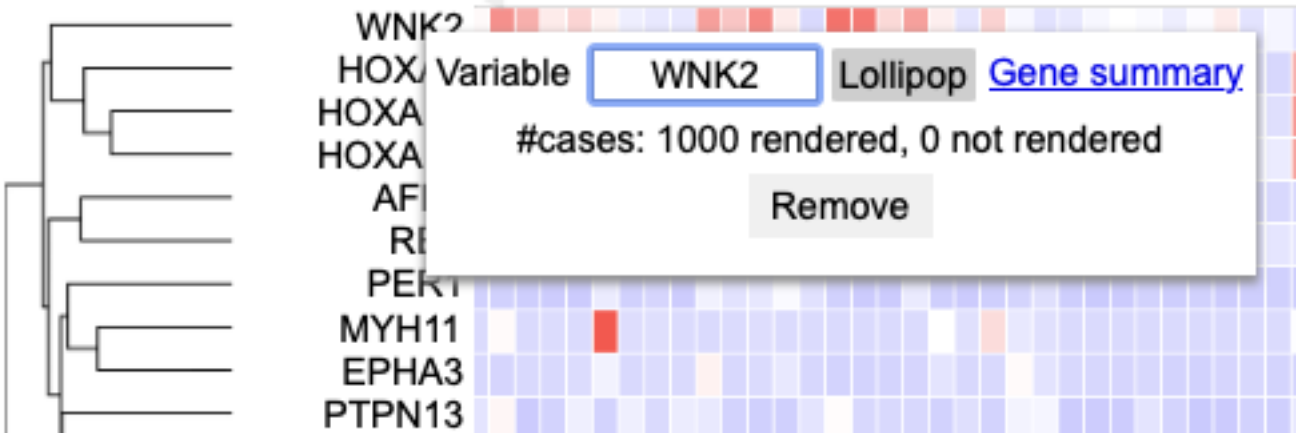
Selecting cases on the cluster

Cases on the cluster can be selected by clicking on the dendrogram. Once part of the dendrogram is selected, users can choose to zoom in to the cases, list all highlighted cases, or create a cohort of the selected cases.

Click on a case in the dendrogram to showcase the Disco plot or the GDC Case Summary Page.

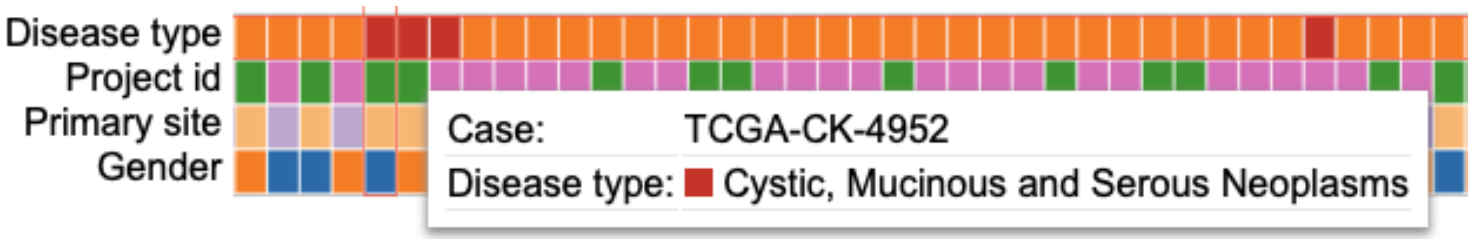


In the column of genes on the left, click on a gene to rename it, launch the ProteinPaint Lollipop plot, display the GDC Gene Summary Page, or remove the gene. The lollipop plot displays all cases across the GDC affected by SSMs in the selected gene.

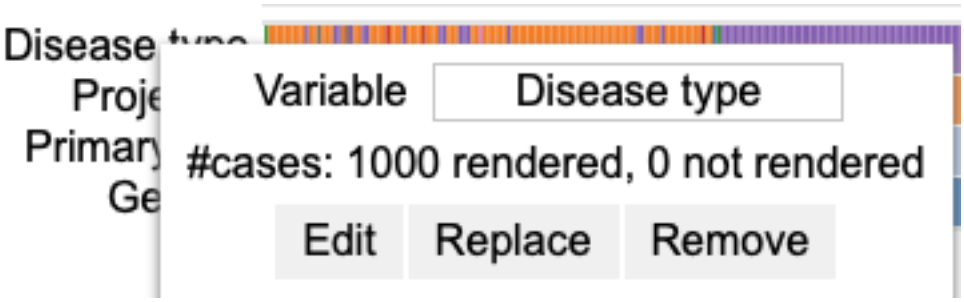


Variables

Any variables added to the matrix appear below the heatmap. Users can hover over a cell to display the case submitter_id and their value for the given variable.

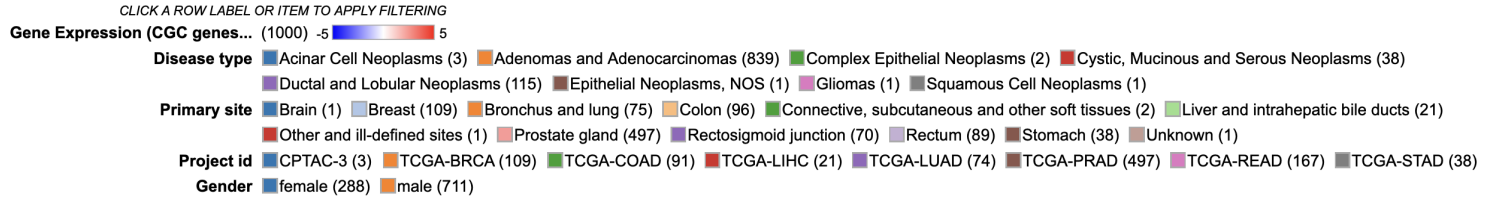


Click on a variable to rename it, edit it by excluding categories, replace it with a different variable, or remove it entirely.

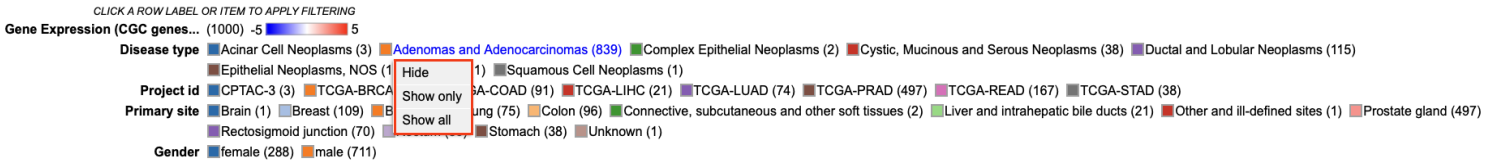


Legend

In addition to the color coding system for the gene expression values, the legend displays the number of cases from the active cohort in each category for all variables that are selected to appear in the matrix.

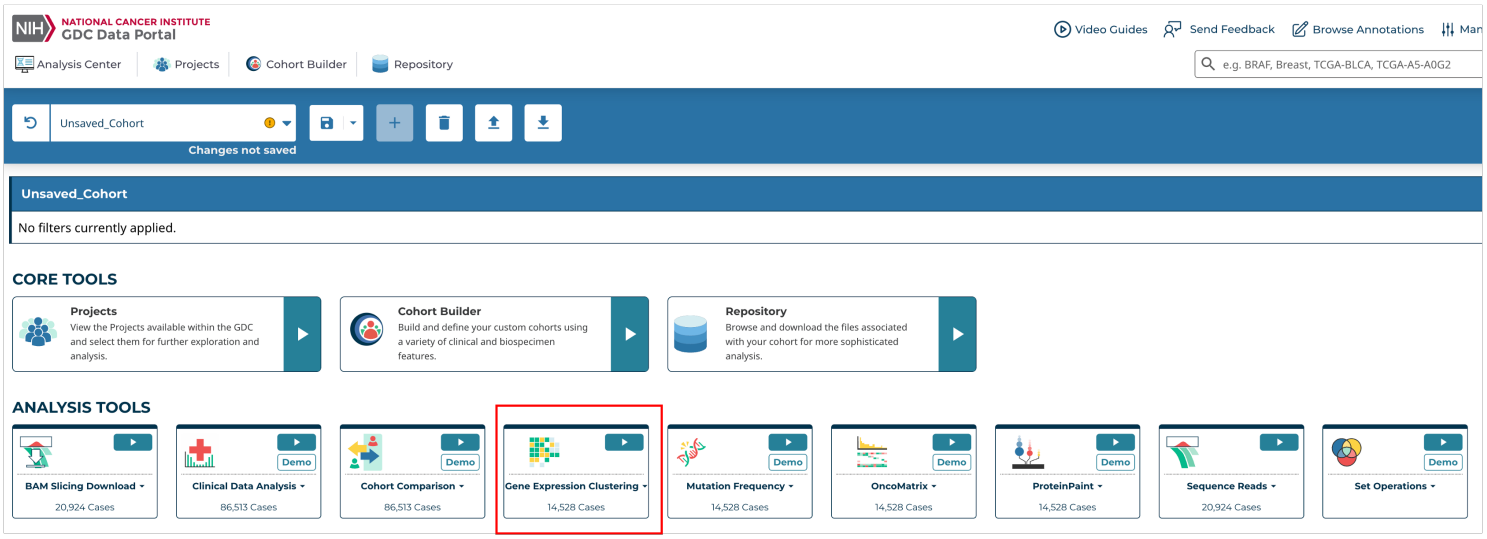


Users can click on a variable in the legend to hide a specific category, only show a specific category, or show all categories for the selected variable.



Accessing the Tool

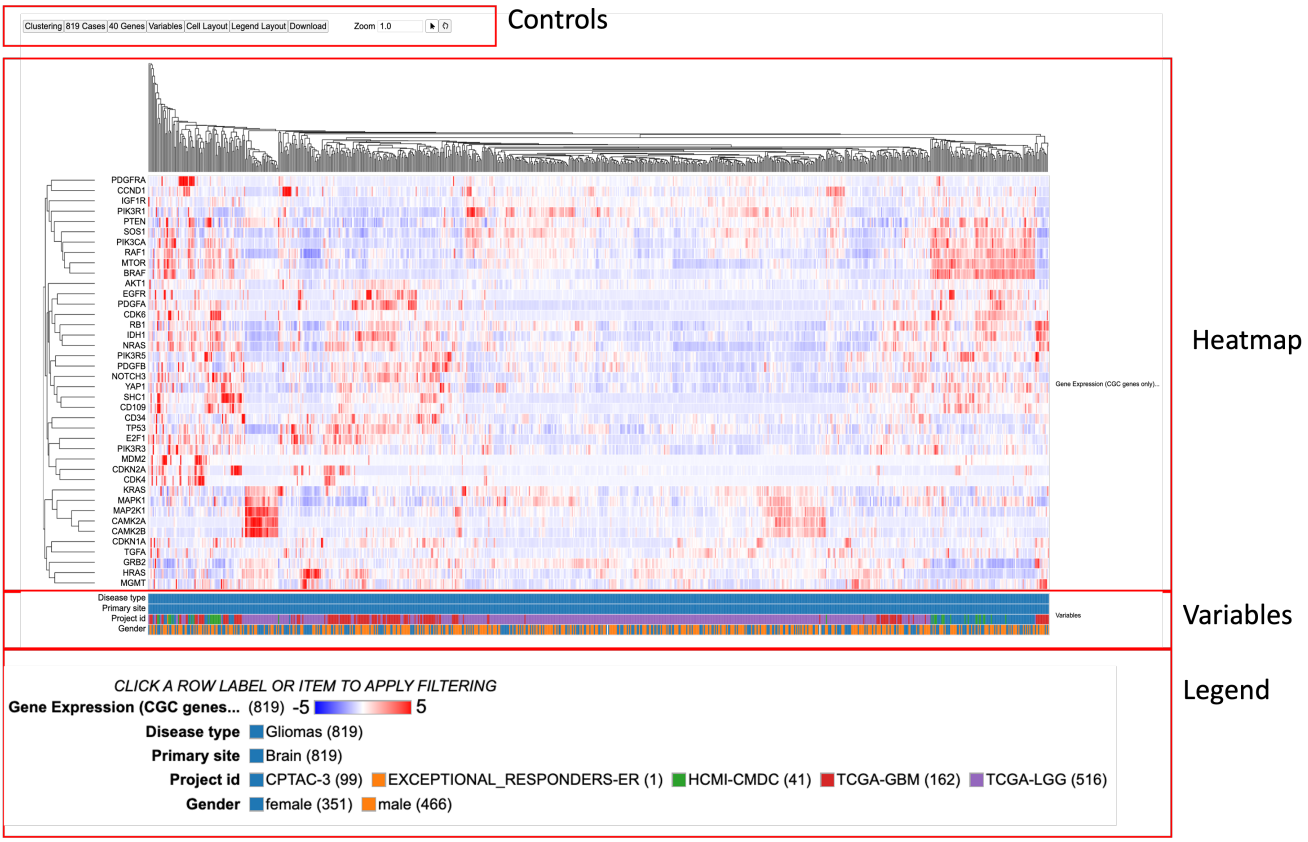
At the analysis center, click the 'Gene Expression Clustering' card to launch the heatmap.



View publicly available genes as well as login with credentials to access controlled data.

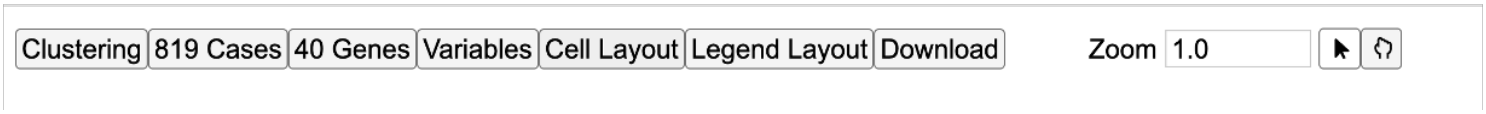
Features

The following features are viewable once the default heatmap is loaded. The default heatmap shows all the glioma cases. There are four main panels as outlined in the figure i.e., the 'Controls', 'Heatmap', 'Variables' and the 'Legend'. Each of the features and functionalities are described in detail in the following sections.



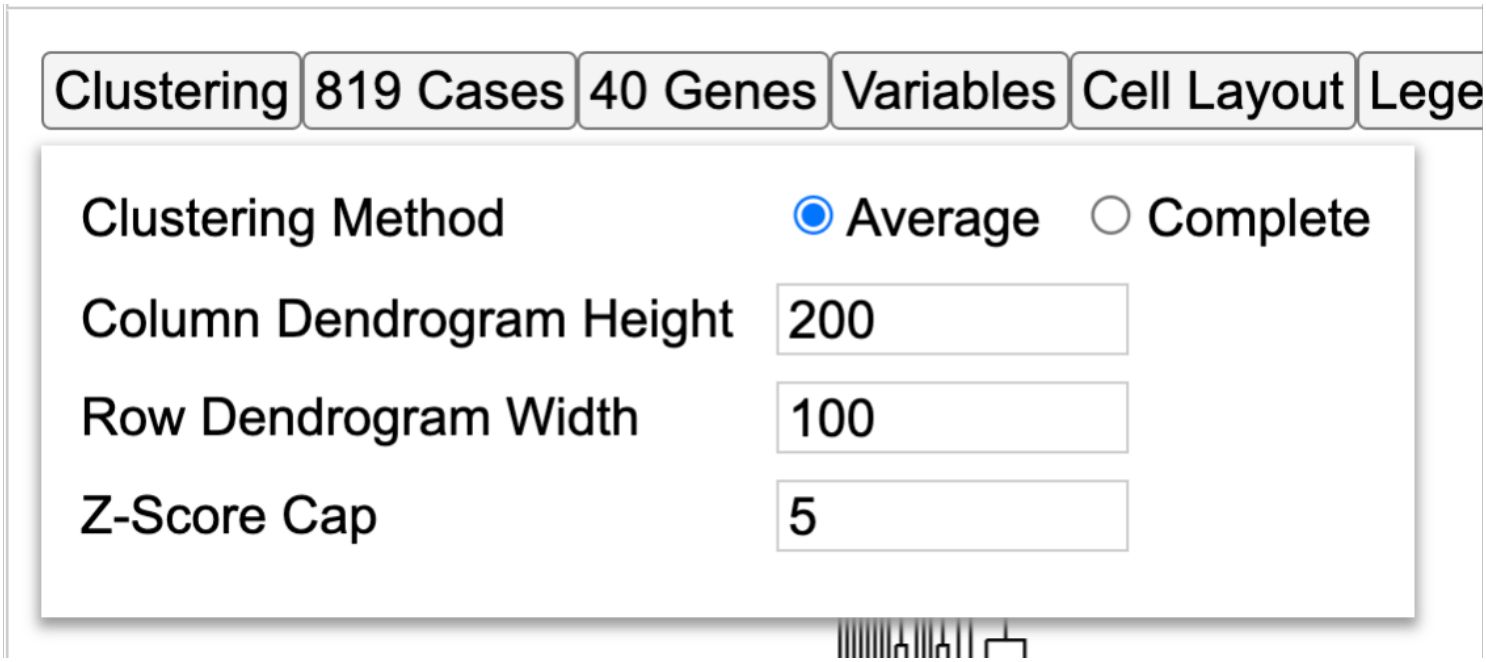
Controls

The control panel as shown has various functionalities with which users can change or modify the appearance of the matrix. The control panel provides flexibility and a wide range of options to maximize user control.



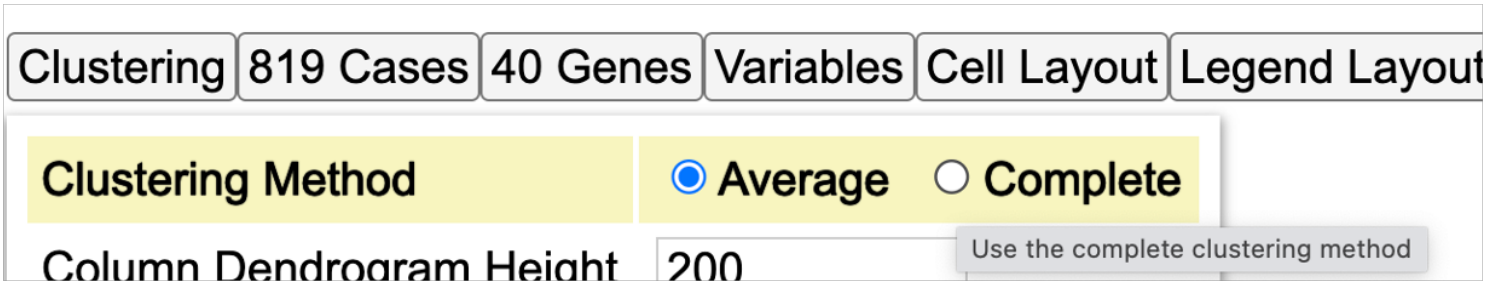
Clustering

The clustering control button provides several options to modify the default clustering of the heatmap. Click on the button labeled 'Clustering' to display a menu with options as shown.

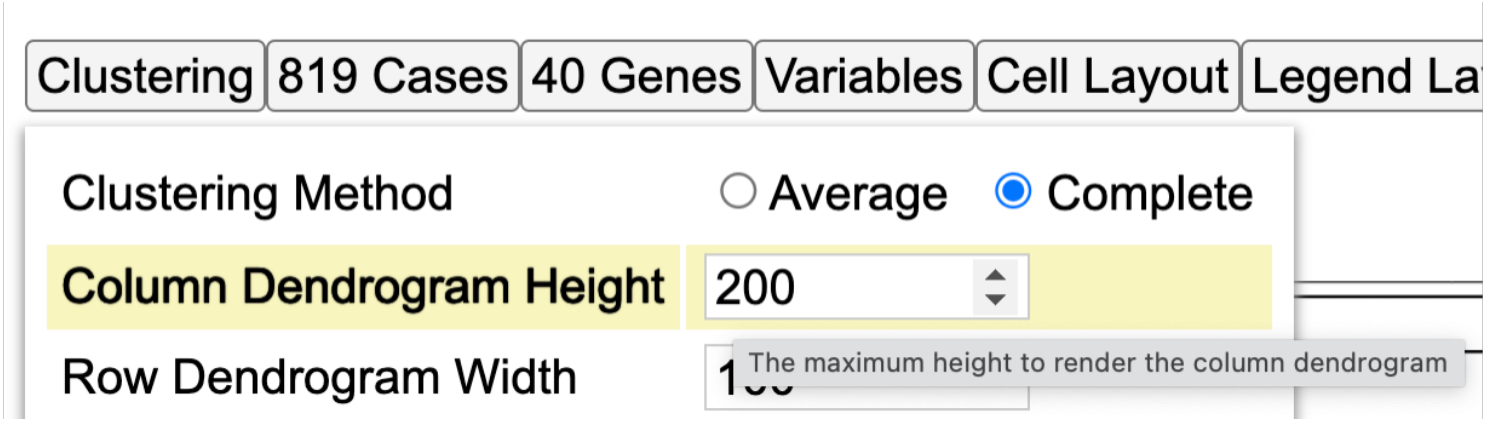


Clustering method

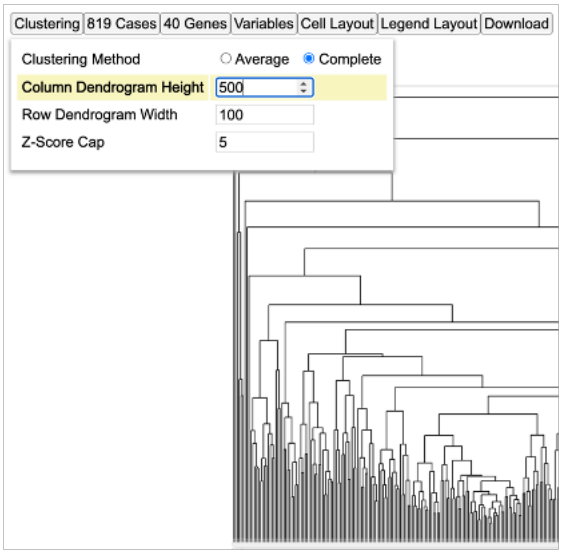
Click on the 'Complete' option as highlighted to change the method of clustering. The heatmap will render again to show the complete clustering method.



The maximum height of the column dendrogram is shown in the next highlighted option as shown.

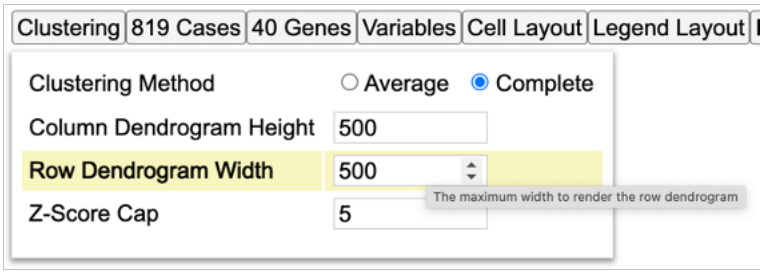


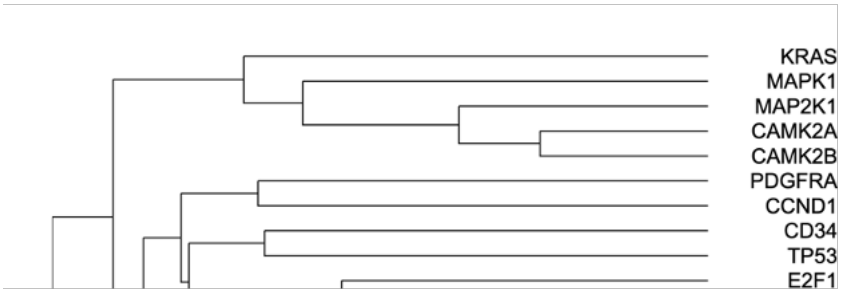
Click or edit the number in the input box to adjust the height of the column dendrograms as shown.



Row Dendrogram Width

Similary, row dendrogram width can also be modified as per user requirement as shown.

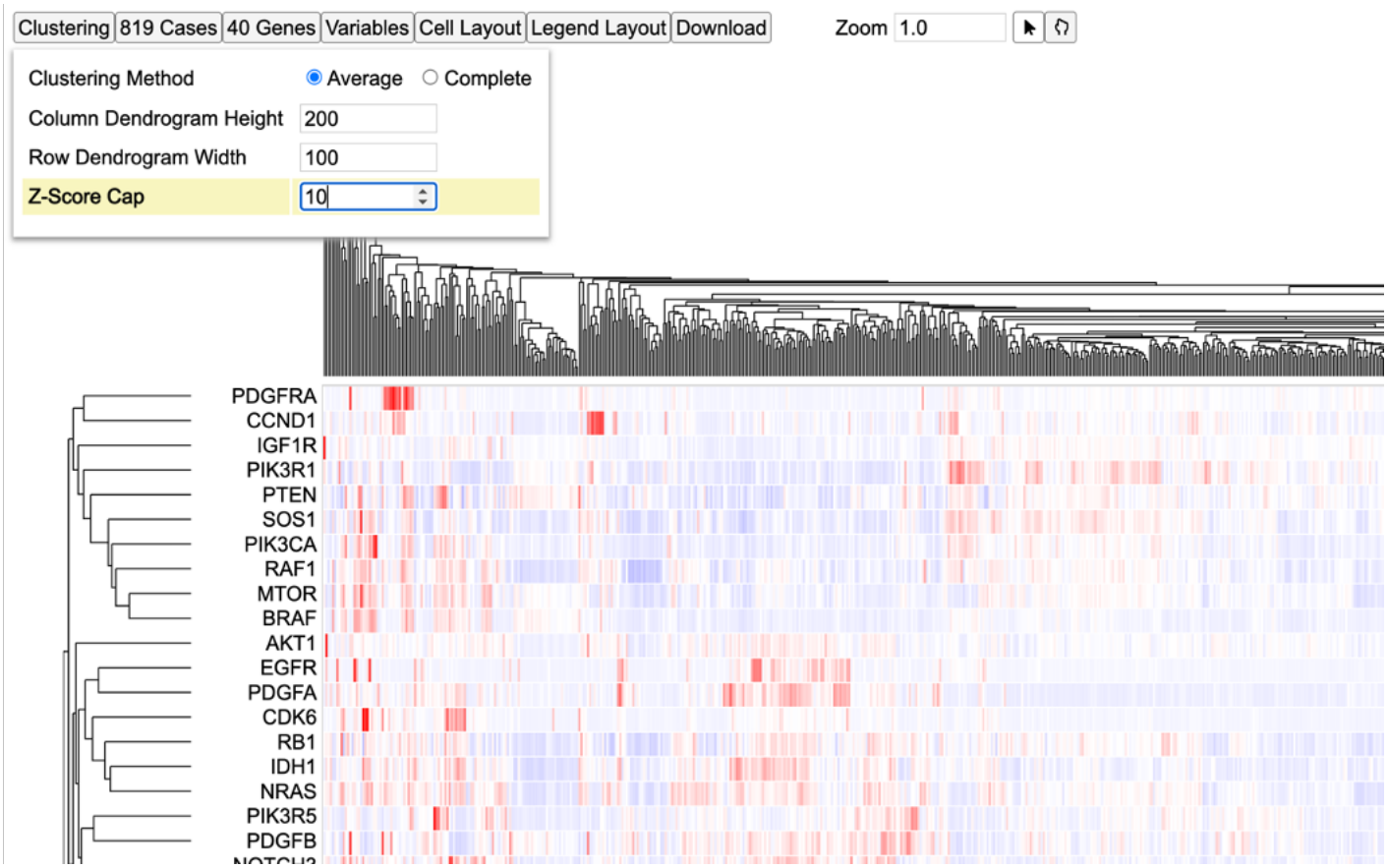




Z-score Cap

Z scores are used to compare gene expression across samples. A Z-score of zero indicates that the gene's expression level is the same as the mean expression level across all samples, while a positive Z-score indicates that the gene is expressed at a higher level than the mean, and a negative Z-score indicates that the gene is expressed at a lower level than the mean.

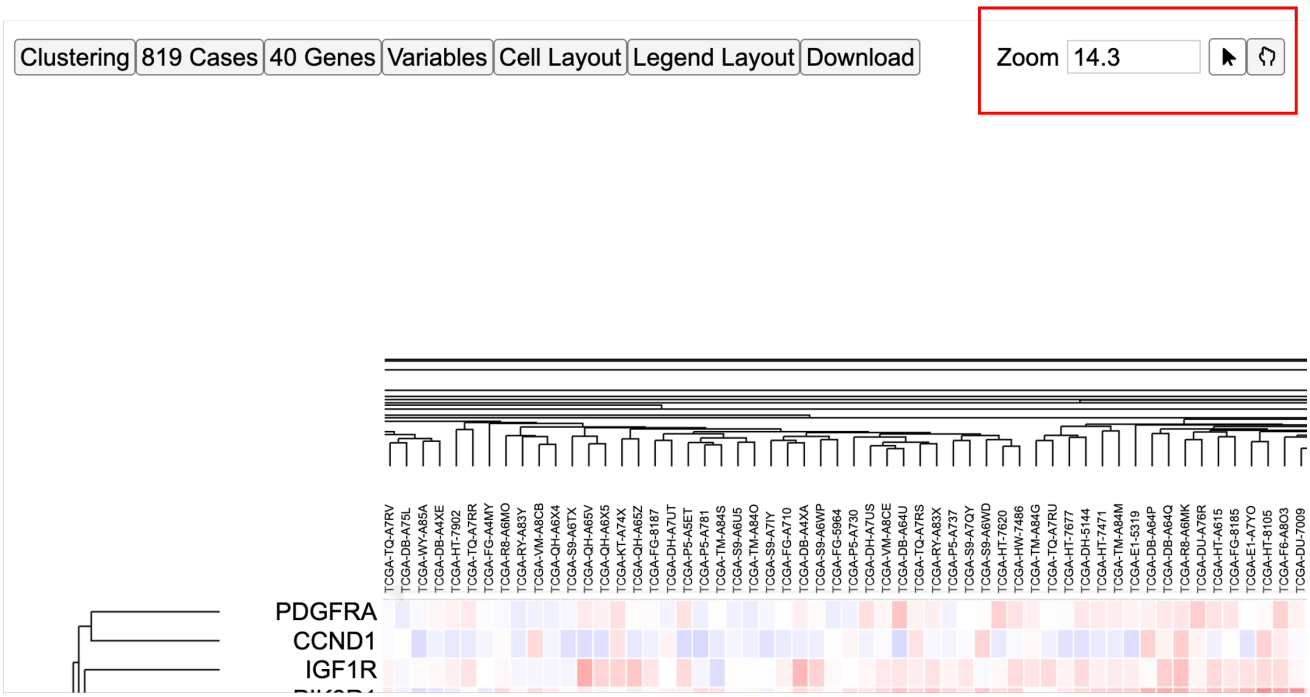
User can increase or decrease the Z-score Capping. Increase the Z-score cap from 5 to 10 as shown. Samples with lower gene expression gets lighter to allow highlighting of clusters with higher expression values as shown in red in the heatmap.



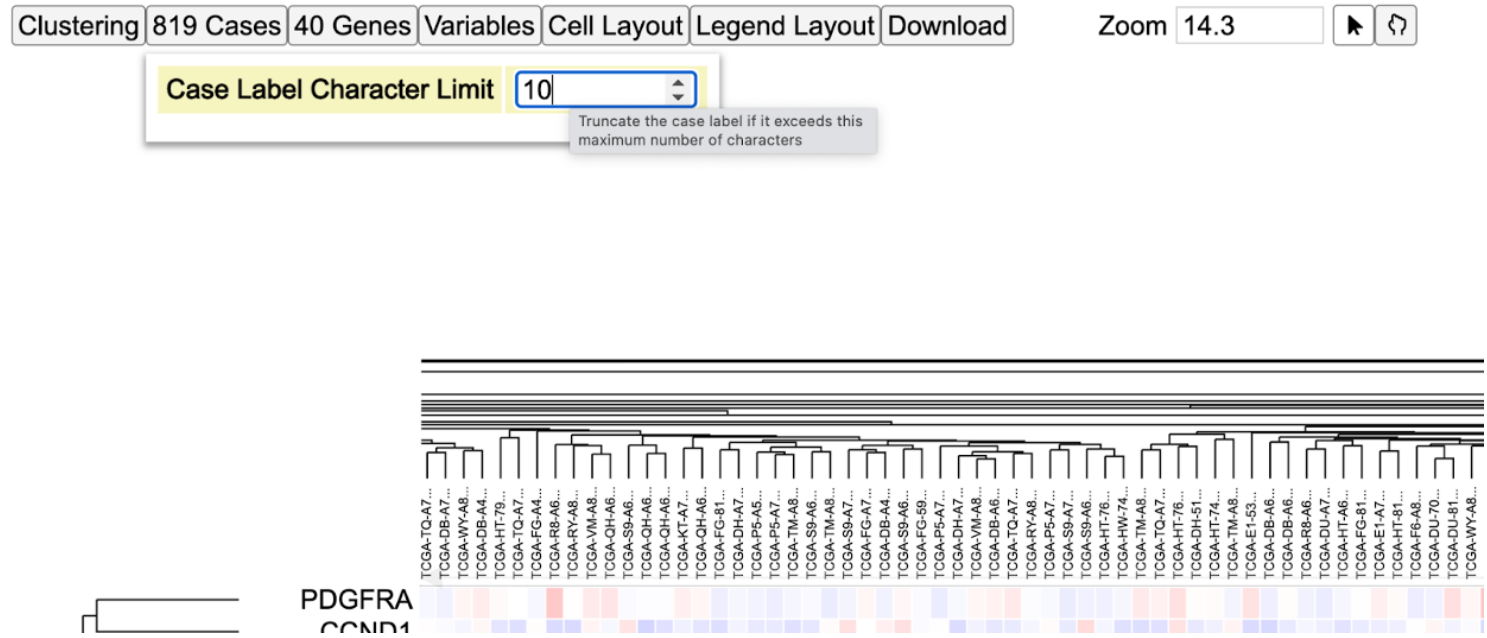
Cases

Adjusting the zoom using the zoom buttons

Adjust the zoom level by using arrows on the input box or entering a number to be able to view the sample lables as shown.

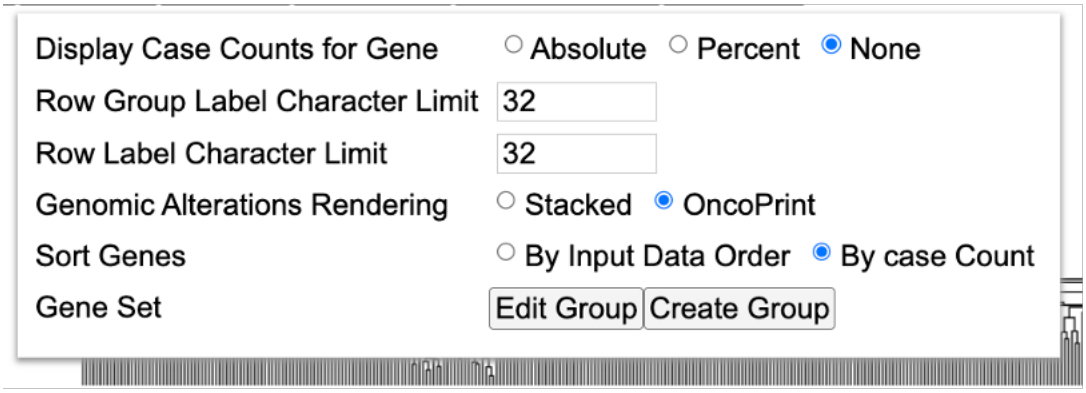


The 'Cases' control has the option 'Case Label Character Limit' to adjust the visible characters of these sample labels. The default is '32'. Change that to '10' to see the new limit applied to the sample labels as shown. Note that reducing the character limit truncates the labels.



Genes

User can modify the existing default gene set by clicking the 'Genes' button in the controls as shown. This displays the option to edit genes as well as variables from the dropdown as shown.

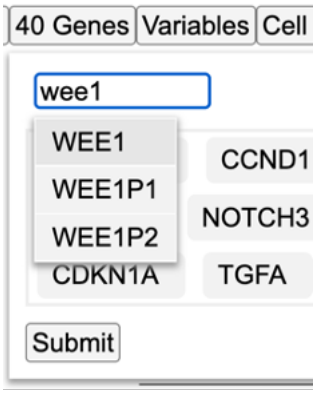


Modifying Genes

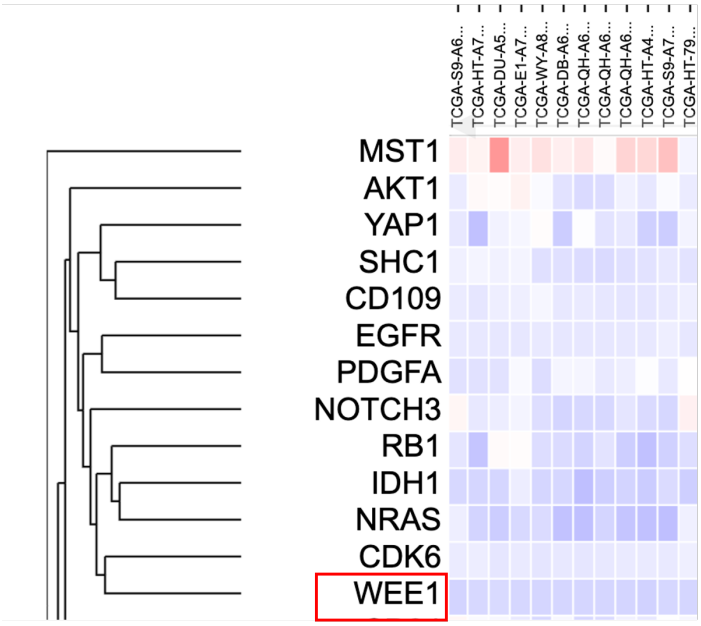
Click the 'Edit Group' button as shown in the 'Gene set' to display a panel of current selected genes.

Add/Delete a gene

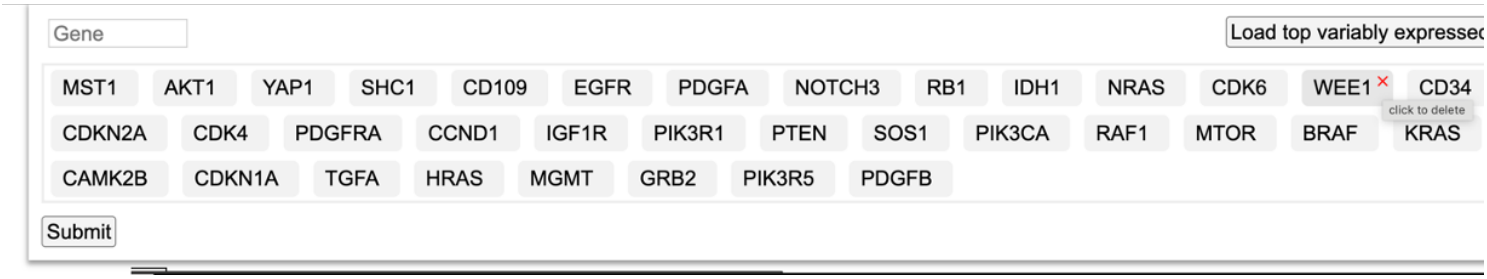
In the search box, type in any gene name for example 'Wee1' as shown and click submit.



The heatmap loads again after performing a clustering that includes 'WEE1' as shown.



Click on the 'Edit' functionality again within the 'Gene set' menu option. To delete a gene, hover over the gene as shown. A red cross mark will appear as shown.



Click on the gene 'Wee1' to delete the gene from the gene set. Click submit to redo the clustering.

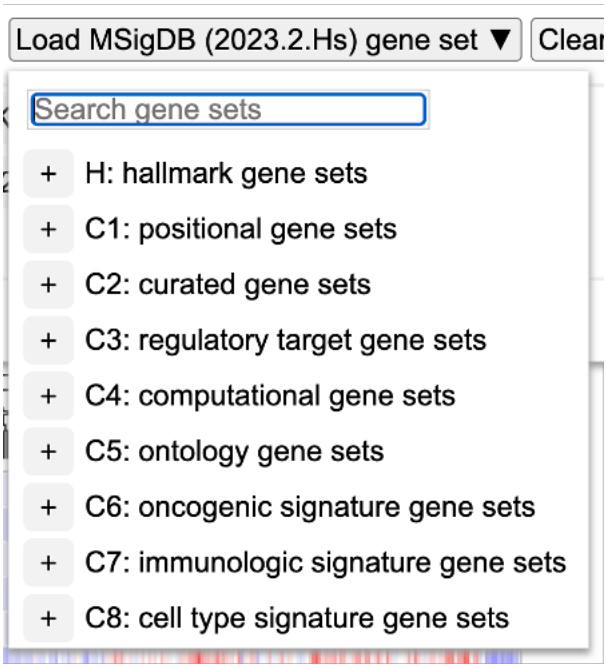
Load top variably expressed genes

User has the option to load the top genes that are variably expressed. To do so, click on the 'Edit Group' button under the 'Genes' controls. Click on the button that reads 'Load top variably expressed genes'. The genes will change to the top most variable genes as shown in this selected cohort.

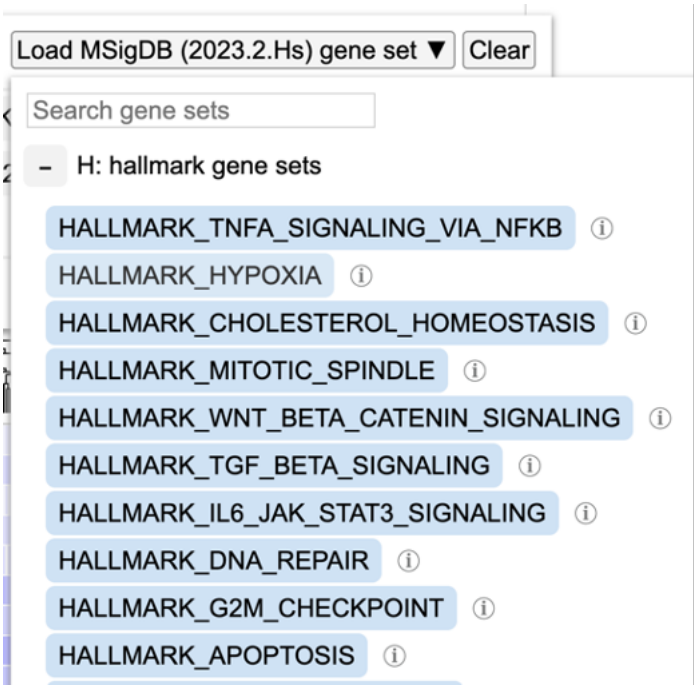
Click submit to reload the heatmap.

Load MSigDB gene set

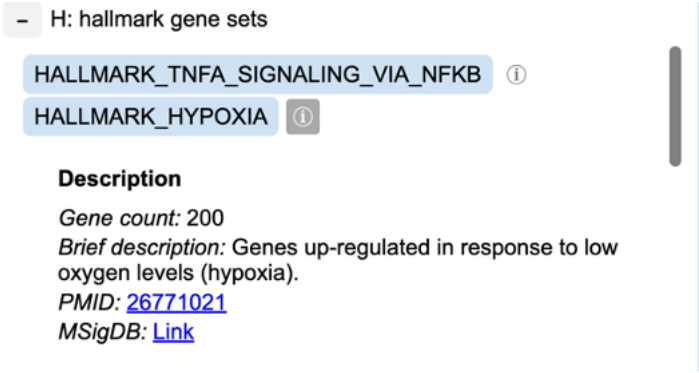
The gene expression clustering tool also enables users to load a pre-defined gene set provided by the MSigDB database. The current version enabled is the latest. Click on the dropdown button 'Load MSigDB (2023.2.Hs) gene set' and choose one of the following gene sets as shown.



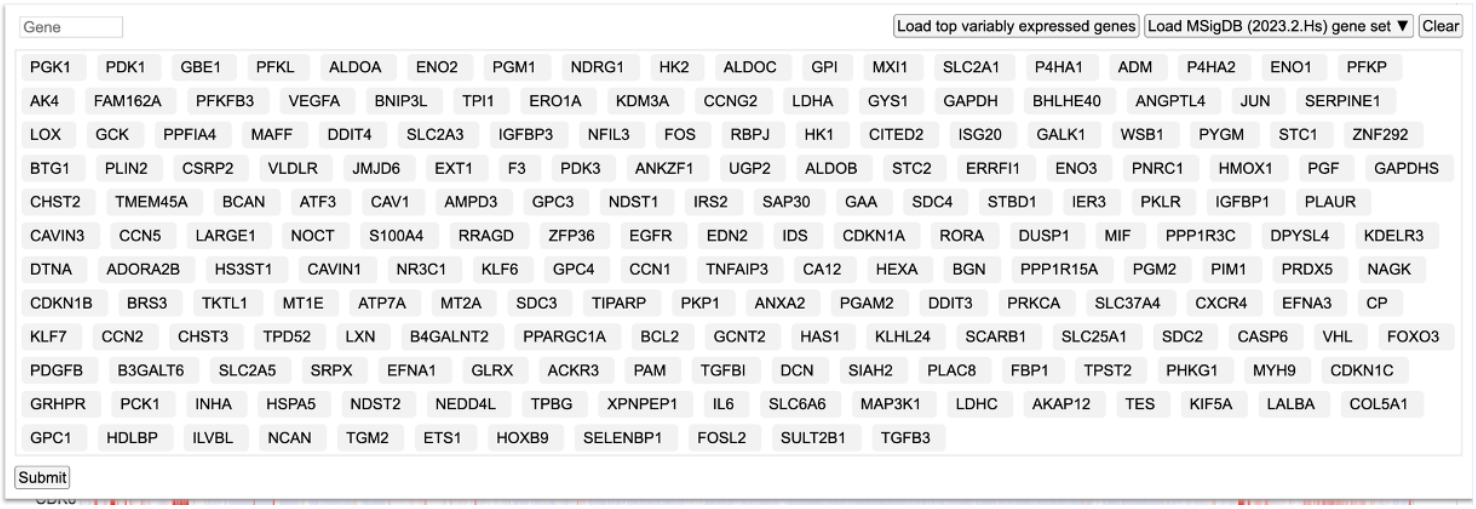
For example, select a hallmark gene set for 'Hypoxia' as shown.



Note the info icon next to the gene set that provides additional information about this gene set as well as a link to the database and the original publication PMID as shown.



Upon selecting a MSigDB gene set, the genes get updated as shown.

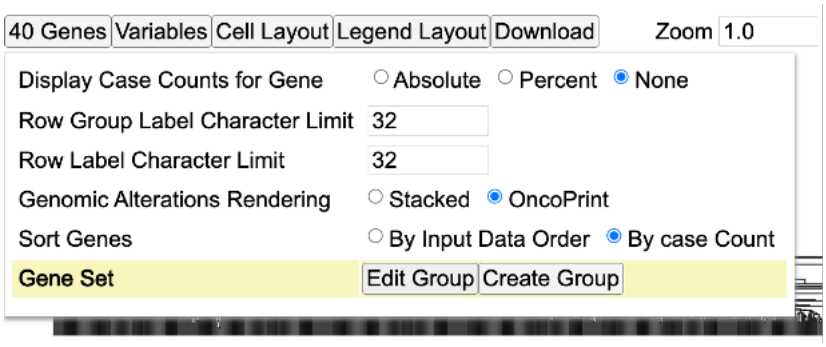


Click 'Submit' to reload the heatmap with the new gene set from MSigDB.

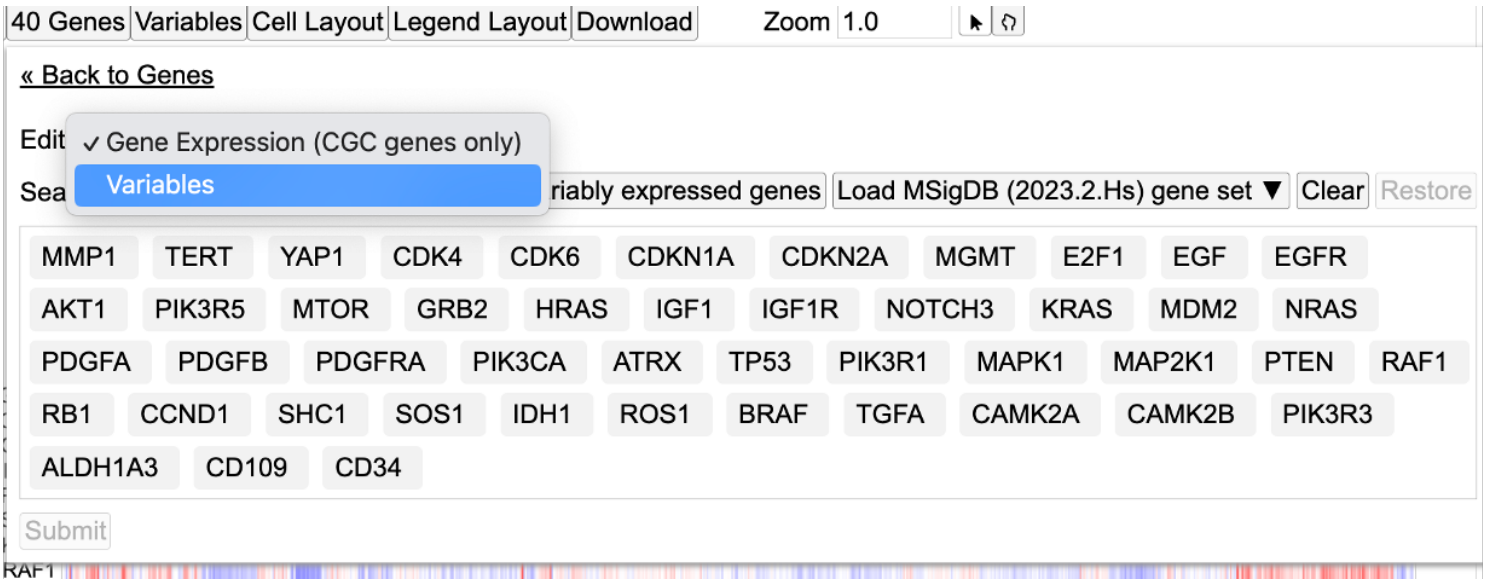
Adding gene as a variable

User also has the option to add gene variant terms as variable to line up mutation consequences with clustered gene expression data.

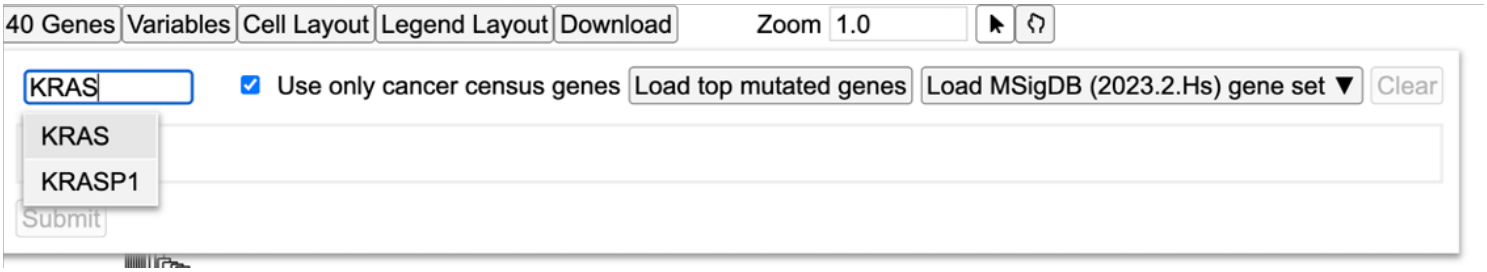
To do so, click the button 'Genes' and click 'Edit Group'.



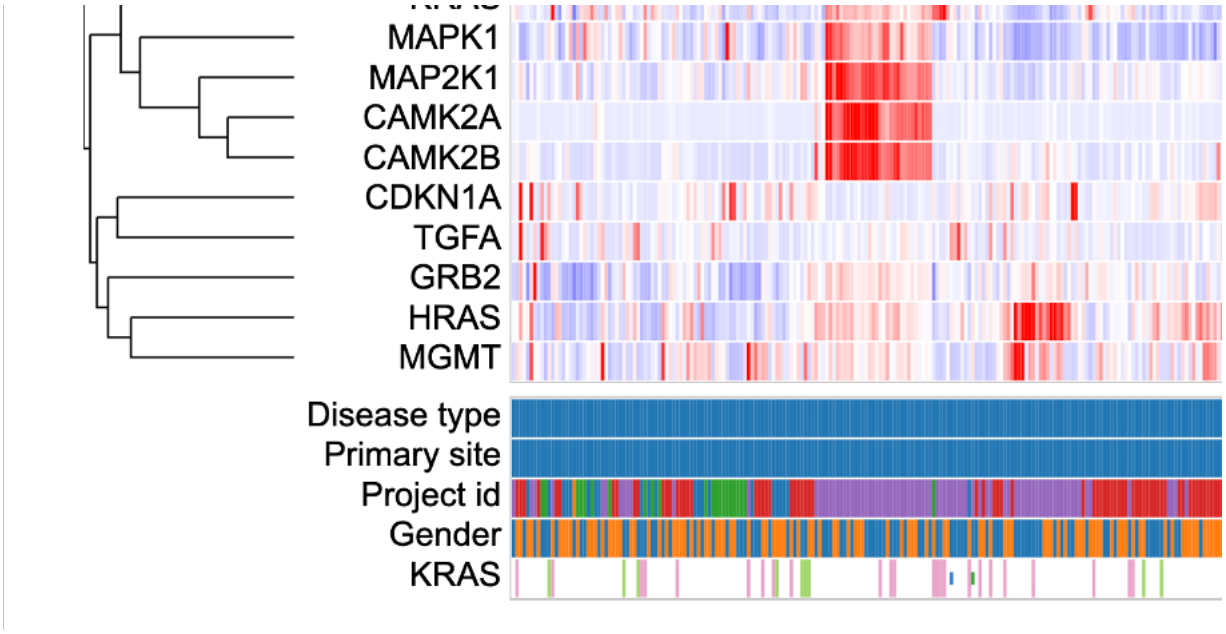
From the dropdown, select 'Variables' as shown.



Search and select 'KRAS'.



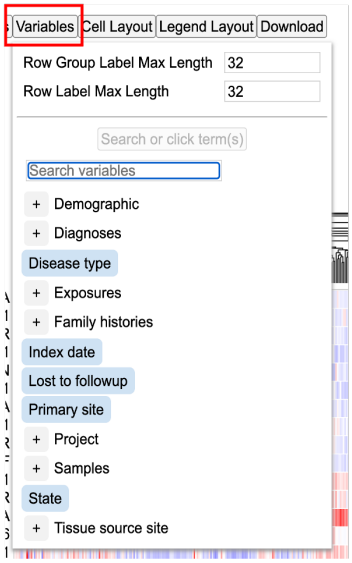
Click 'Submit' to reload the heatmap with the newly added KRAS gene as a variable. This displays the consequence type for the clustered samples for which KRAS has both the mutation calls and the gene expression data as shown.



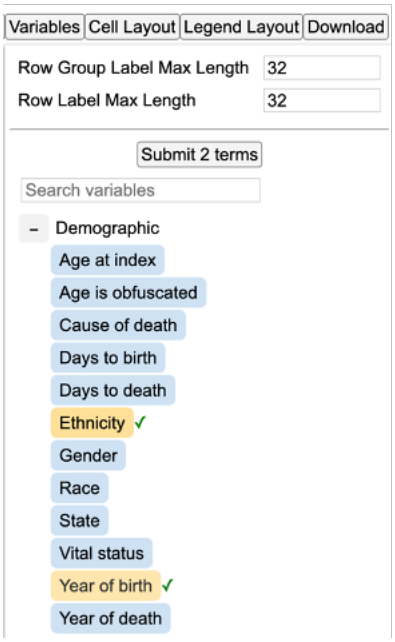
Variables

The button 'Variables' in the controls allows the user to search and select variables that get added below the heatmap.

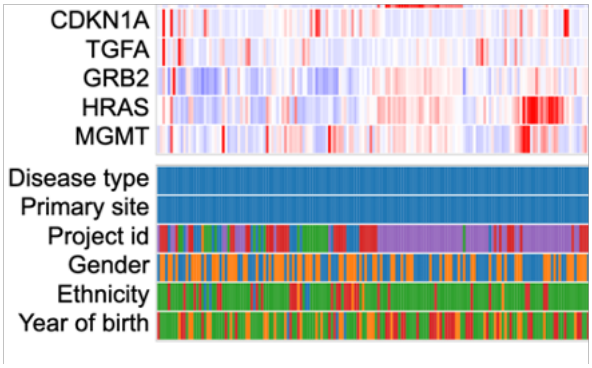
Click the button 'Variables' to show the following dictionary tree.



Click the '+' button on the 'Demographic' to display all the terms under the parent term as shown. Select terms 'Ethnicity' and 'Year of birth' and click 'Submit 2 terms'.



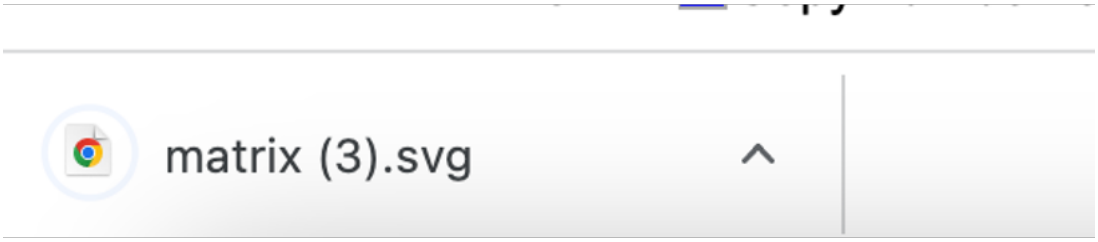
Once the variable terms are submitted, the heatmap will display the added variables as shown.



Download

The control panel shows an option to download the plot as an svg after user has specified their customizations. Select the 'Download' button as shown below to save the svg.

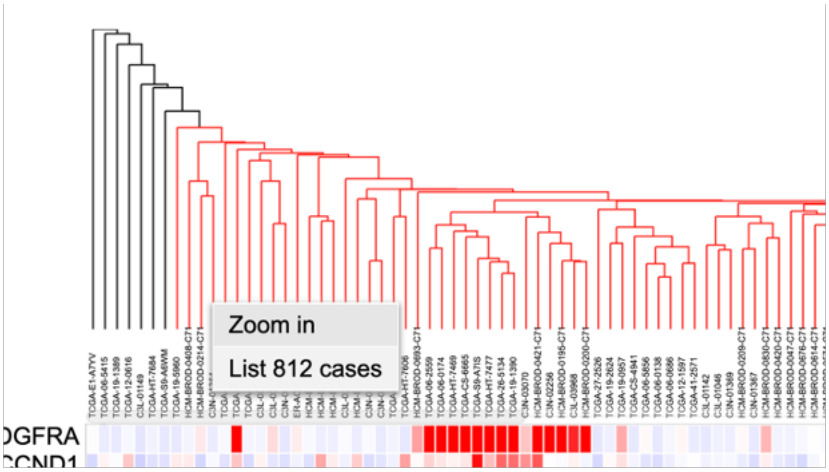
The download will get saved to the default download folder as shown at the bottom of the browser window.



Heatmap

Selecting cases on the cluster

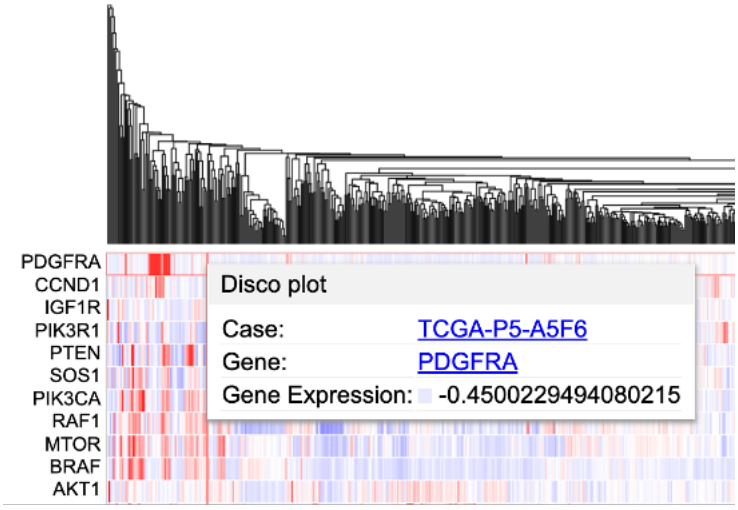
Cases on the cluster can be selected interactively by clicking on the column dendrograms. Click on the dendrograms above the heatmap as shown. The dendrograms get highlighted in red.



Once the dendrograms are selected, two options are displayed. A user can choose to zoom in the cases or list all the cases highlighted in the dendrograms.

Clicking a case column

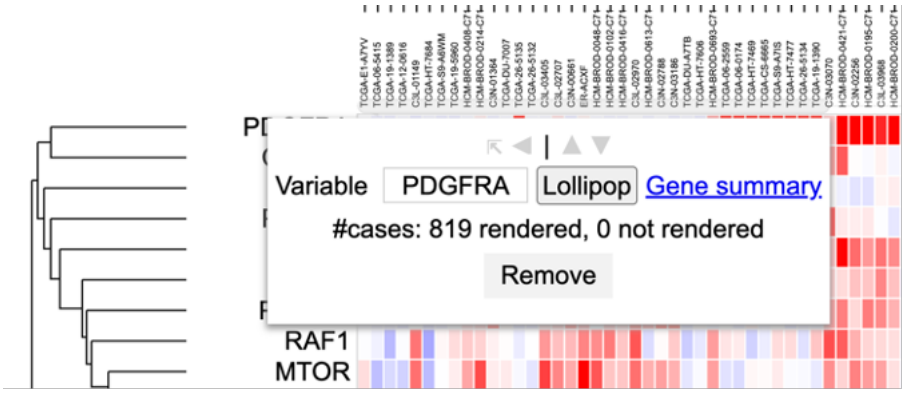
Click on a case label to display the options as shown.



User may choose to launch: - a circos plot by clicking 'Disco plot' button, - a webpage containing information about the case by clicking the case id - Gene summary page by clicking on the gene name 'PDGFRA'

Clicking a gene label

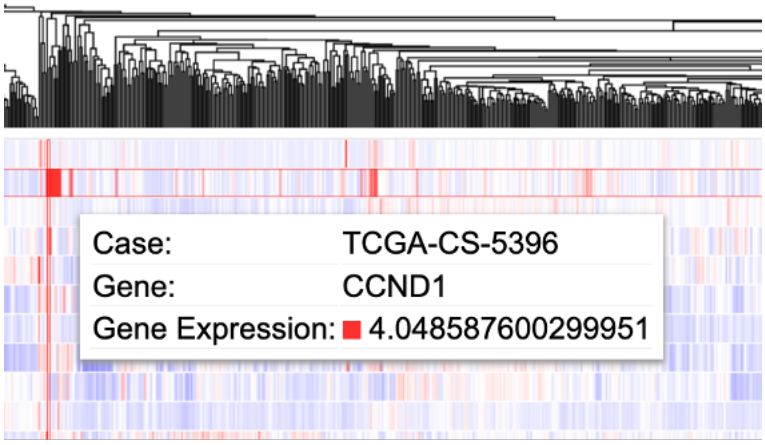
Click on a gene row label to display the following options



User can choose to change variable name by deleting and typing in a new name in the box where 'PDGFRA' is currently applied. User may also choose to launch the lollipop plot or gene summary page or remove this row entirely.

Hovering over/Clicking a cell

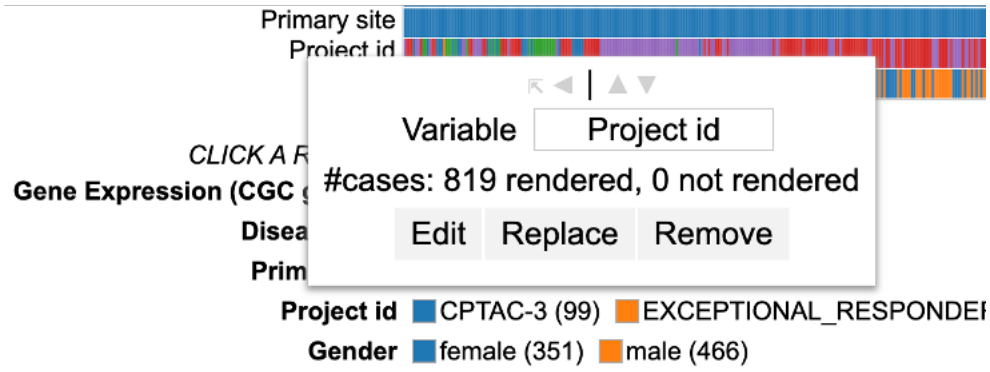
Hover over a cell of the heatmap to show information about the case. The information displayed shows the case id, the gene name (CCND1) and the z-score transformed value (4.04..)



Variables

Clicking a Variable

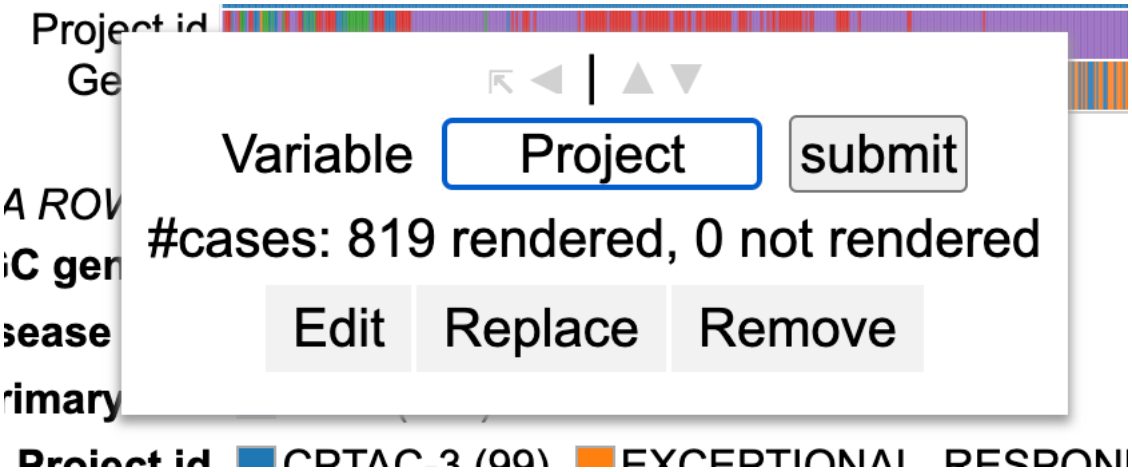
Click on a variable (for example 'Project id' here) row label to display the options as shown.



User can change the variable name (input box), edit the variable to exclude categories ('Edit' button), replace the variable by another one ('Replace' button) or remove the row containing the variable entirely by clicking the 'Remove' button.

Renaming a variable

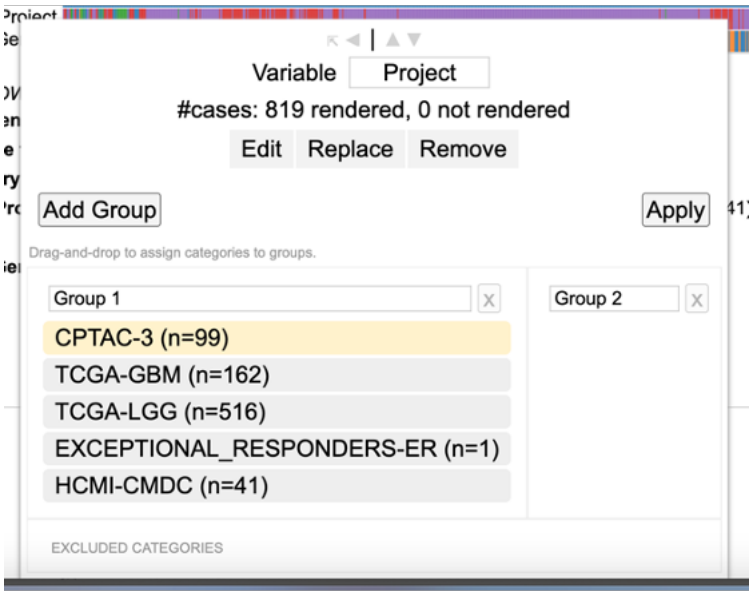
To rename a variable, edit the default name of the variable in the input box as shown.



After renaming the variable as per user preference, click ‘submit’. The row now shows a new variable name.

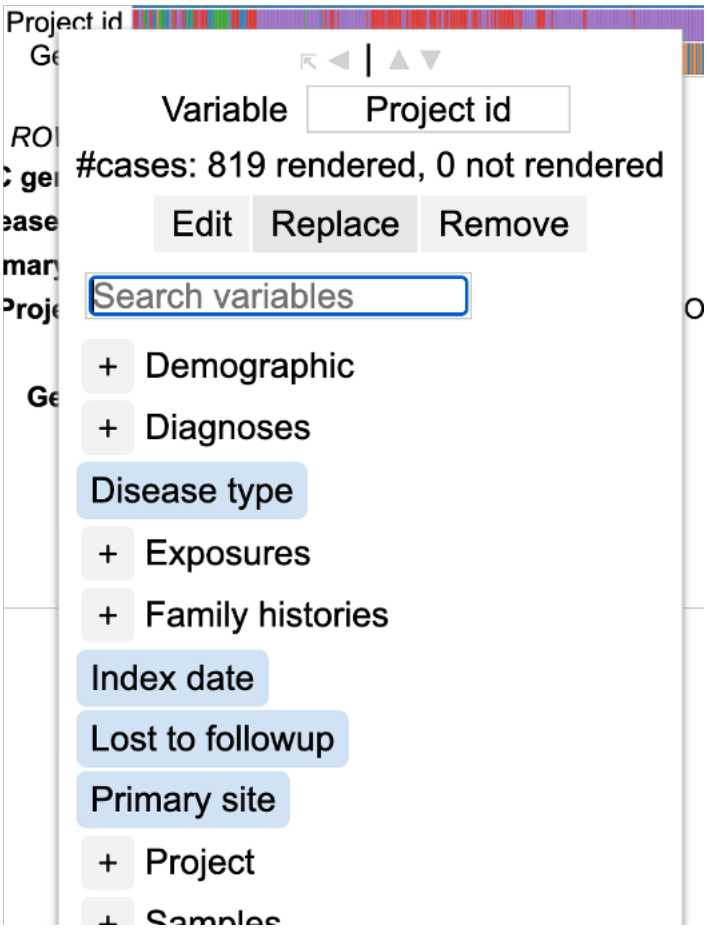
Editing a variable

To edit groups within the variable, click the ‘Edit’ button. Now, user can drag the categories from group 1 into group 2 to create two separate groups and also have the option to exclude a category. After making the choice, click ‘Apply’ to reload the chart.



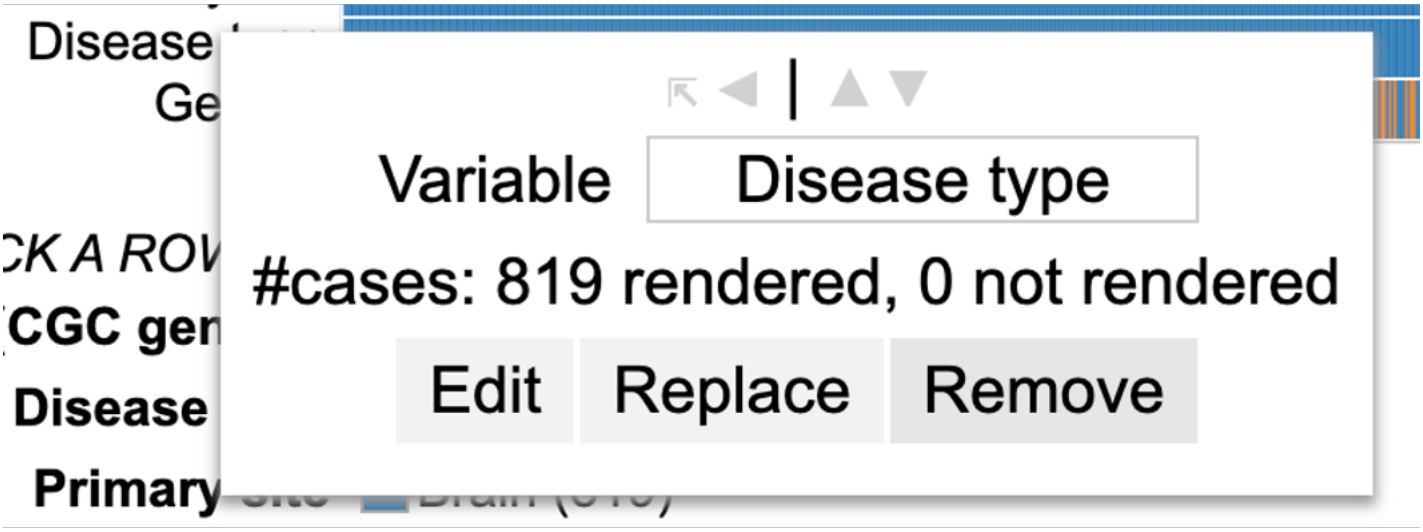
Replacing a variable

To replace a variable, click on the row label for that variable and click ‘Replace’. This shows the GDC dictionary from which a user can select a variable of choice as shown.



Removing a variable

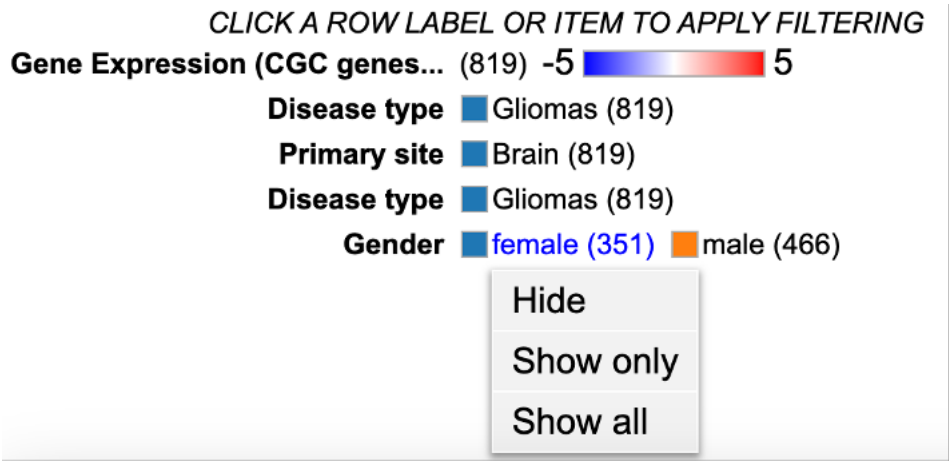
To remove a row containing a variable entirely, click on the row label for that variable and click 'Remove'. This removes the entire row from the heatmap.



Legend

Interacting with legend filters

Variables can be filtered upon via the legend. Click a legend item to display the following options. User may choose to 'Hide', 'Show only', or 'Show all' categories from a selected variable. This would allow the user to filter down on the category of choice.

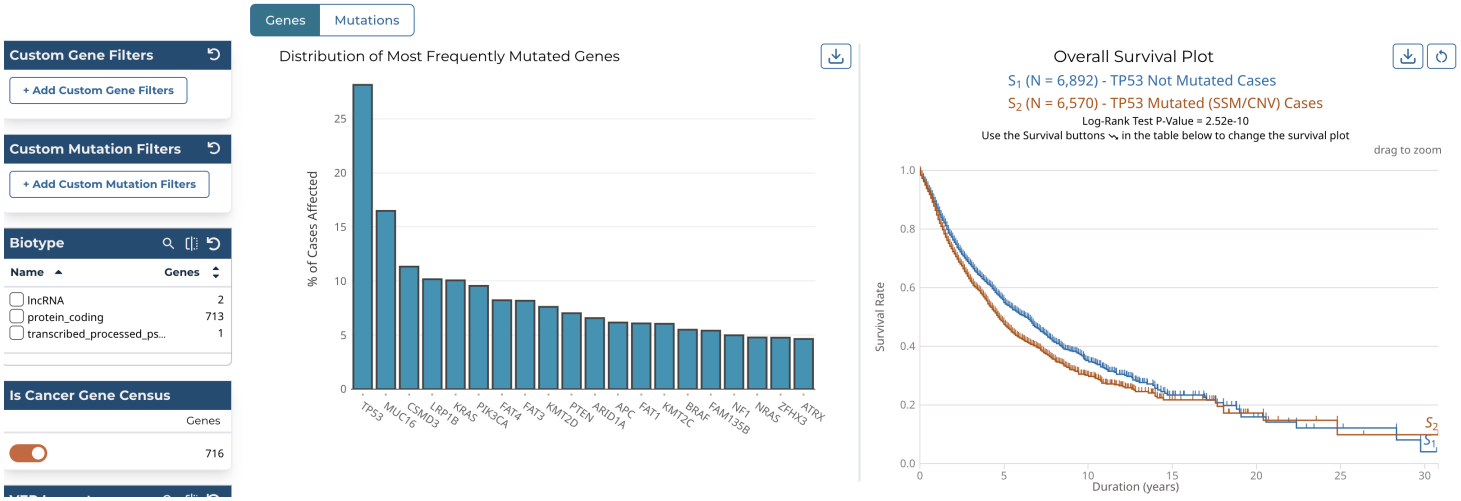


Chapter 11

Mutation Frequency

Mutation Frequency

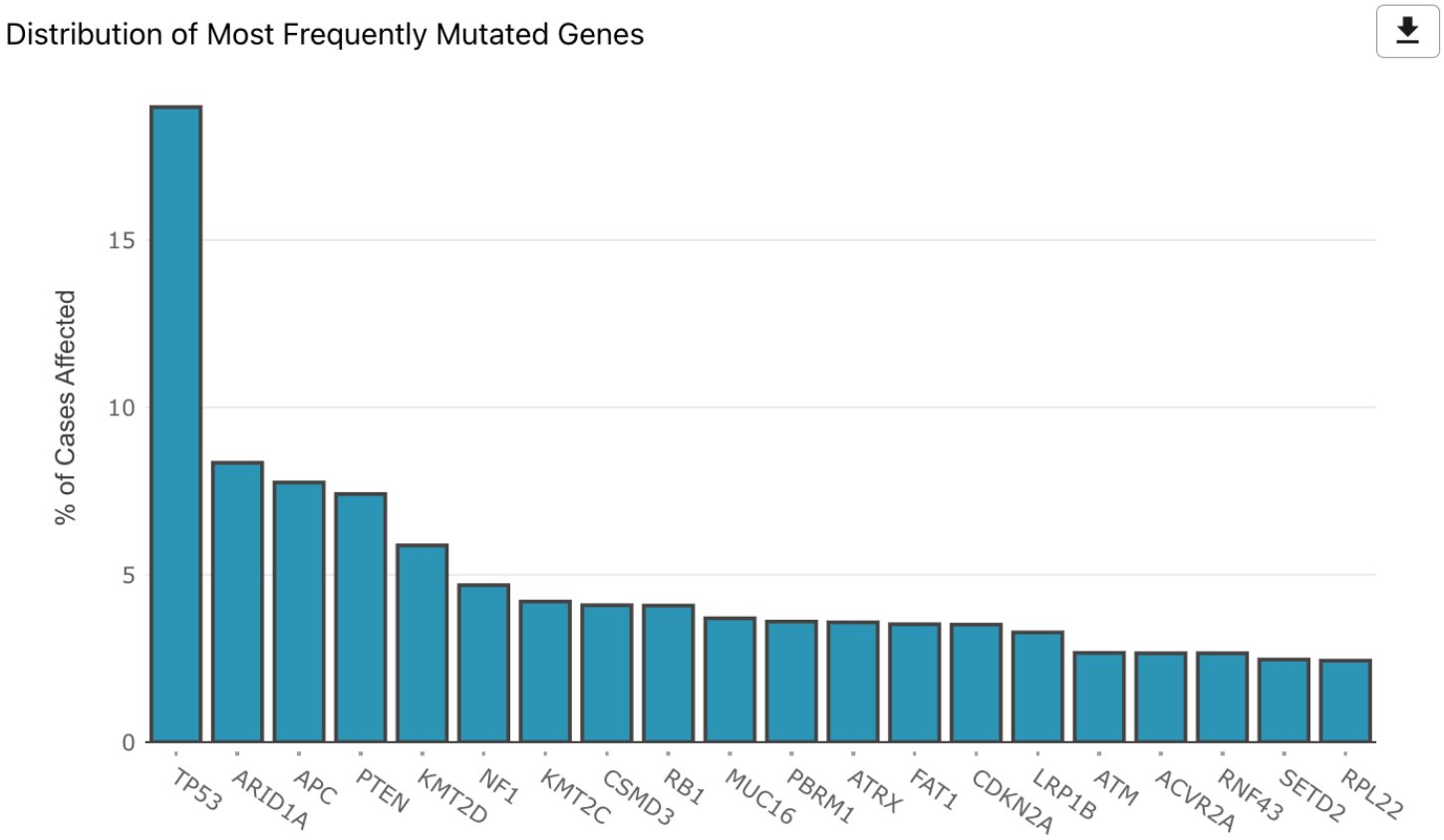
The Mutation Frequency tool visualizes the most frequently mutated genes and the most frequent somatic mutations for the active cohort. To launch the Mutation Frequency tool, click on its card from the Tools section of the Analysis Center.



This tool includes the following visualizations:

Mutated Genes Histogram

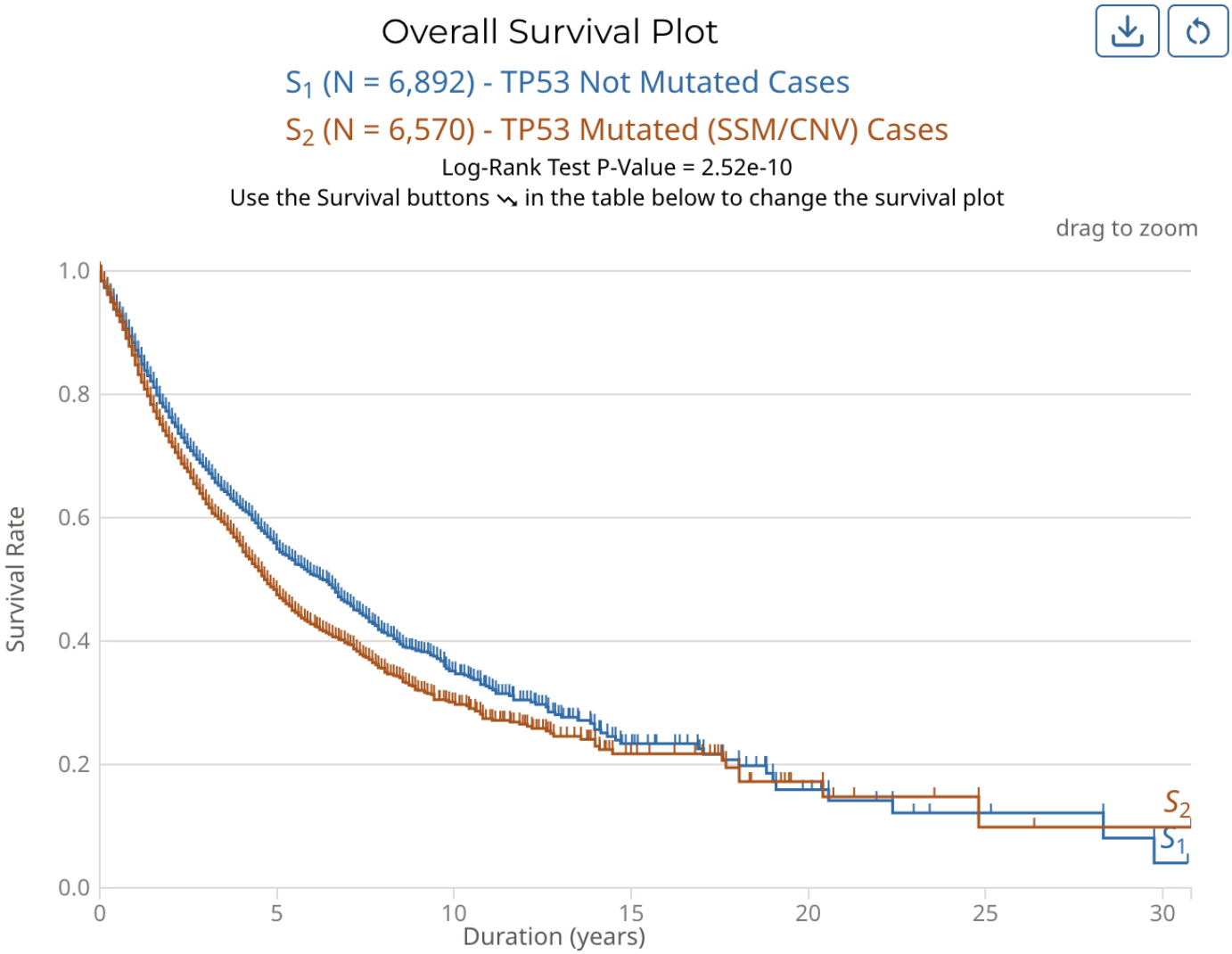
The most frequently mutated genes are represented with a histogram that shows the percentage of cases affected within the active cohort. The histogram can be downloaded as an image (SVG/PNG) or raw data (JSON) using the button at the top right of the graphic.



Survival Plot for Mutated Genes and Mutations

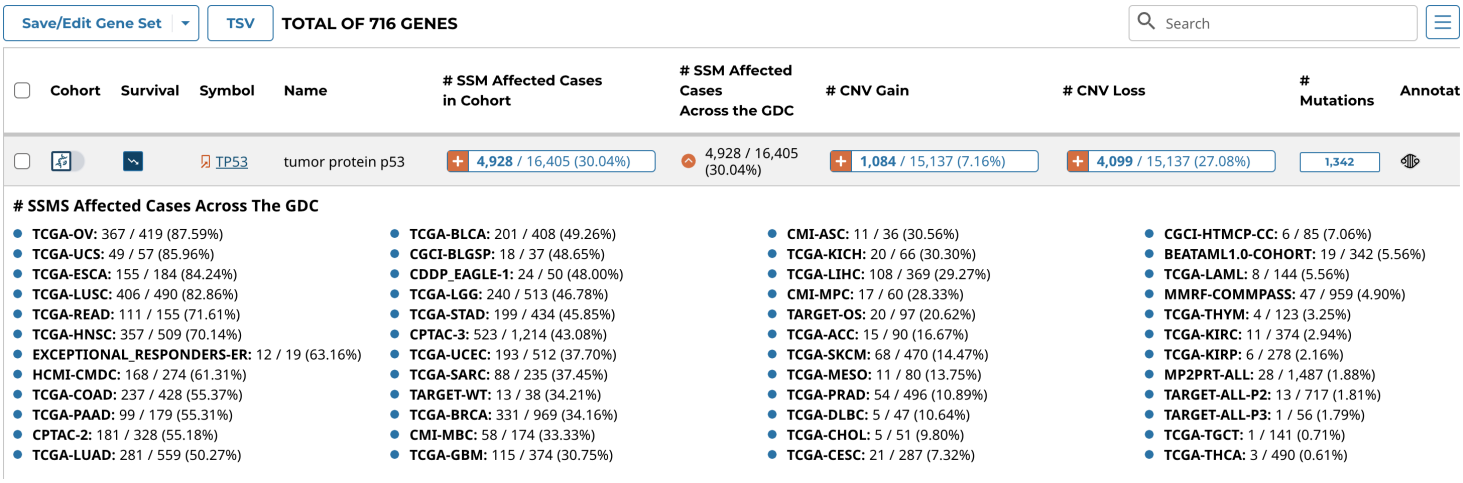
The mutation frequency survival plot is represented with two Kaplan-Meier curves based on cases with and without a specific mutation or mutated gene. Cases for both curves can be further filtered using the various filters available in the left panel of the Mutation Frequency tool. For example, selecting “high” for the VEP impact filter will limit the cases in both curves to those whose mutations have a high VEP impact.

The Log-Rank Test p-value is also displayed here. The survival plot can be downloaded as an image (SVG/PNG) or raw data (JSON/TSV) and the view can be reset using the buttons at the top right of the graphic.

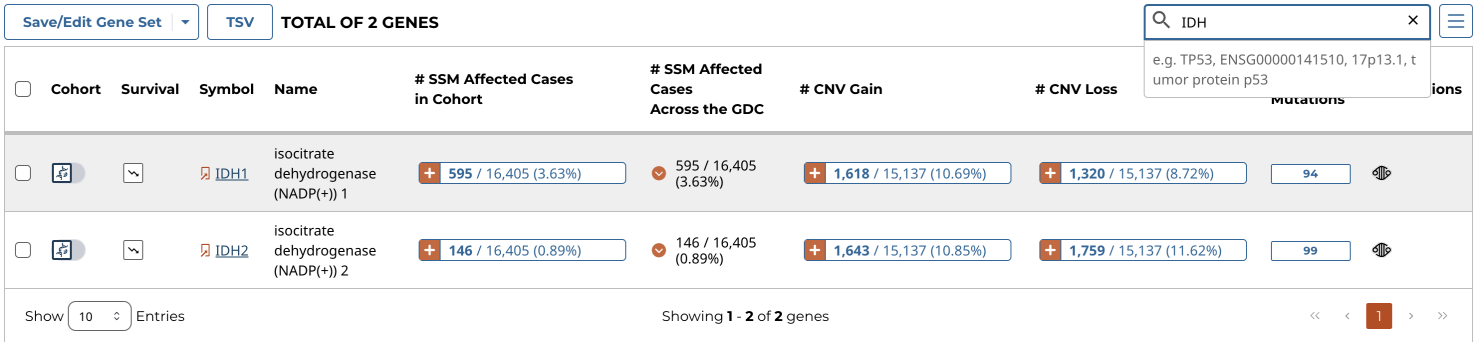


Genes/Mutations Table

The genes/mutations table displays the most frequently mutated genes or the most frequent mutations in the active cohort by percent frequency in descending order. Additional columns show CNV information as well as the number of affected cases. The "Cohort" toggle can be used to filter the current cohort by a specific gene or mutation, and the "Survival" button allows the user to modify the survival plot. The red arrow button allows for the percentage of affected cases to be displayed on a project-level. The data displayed in the table can be exported as a TSV using the button at the top left of the table. Additional cohorts can be created using buttons located within the table.



The table can be searched using the field at the top right of the table.



Additionally, clicking the button in the " Mutations" column within the genes table will automatically apply a search for the corresponding gene in the mutations table. This is a convenient way to view the specific mutations in a given gene.

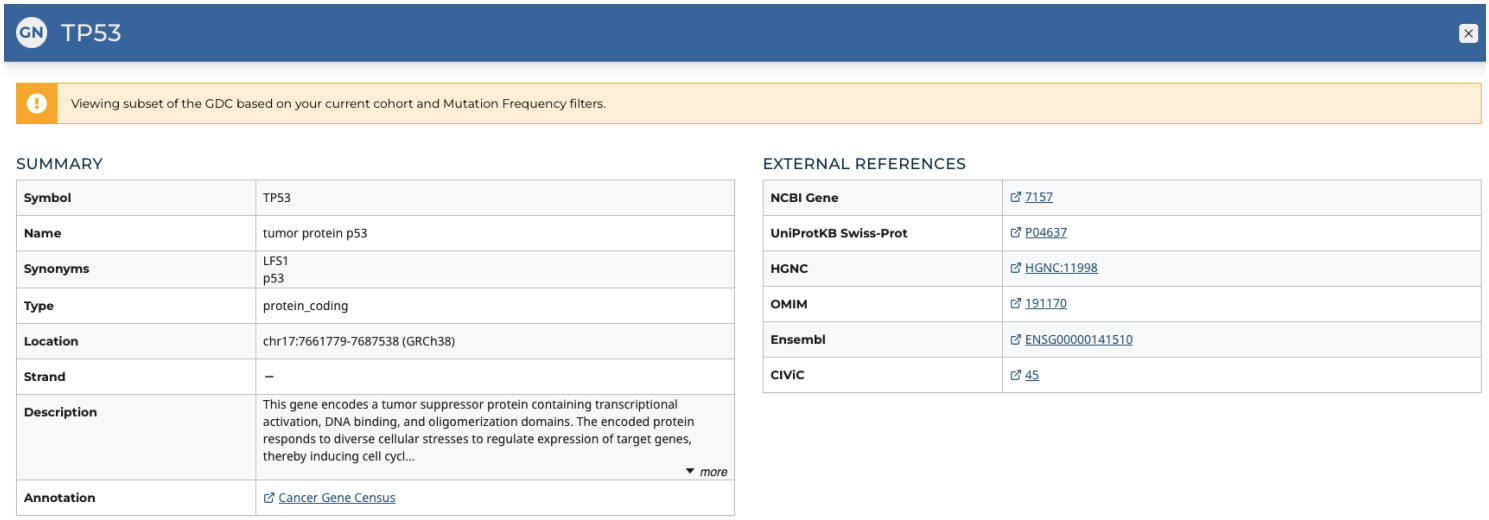
Gene and Mutation Summary Pages

Users can click on the Symbol links in the Genes Table and the DNA Change links in the Mutations Table to view the Gene and Mutation Summary Pages, respectively. These pages display information about specific genes and mutations, along with visualizations and data showcasing the relationship between themselves, the projects, and cases within the GDC. The gene and mutation data that is visualized on these pages are produced from the Open-Access MAF files available for download on the GDC Portal.

Gene Summary Pages describe each gene with mutation data and provide results related to the analyses that are performed on these genes.

Summary

The summary section of the gene page contains the following information:



- **Symbol:** The gene symbol
- **Name:** Full name of the gene
- **Synonyms:** Synonyms of the gene name or symbol, if available
- **Type:** A broad classification of the gene
- **Location:** The chromosome on which the gene is located and its coordinates
- **Strand:** If the gene is located on the forward (+) or reverse (-) strand
- **Description:** A description of gene function and downstream consequences of gene alteration
- **Annotation:** A notation/link that states whether the gene is part of The Cancer Gene Census

External References

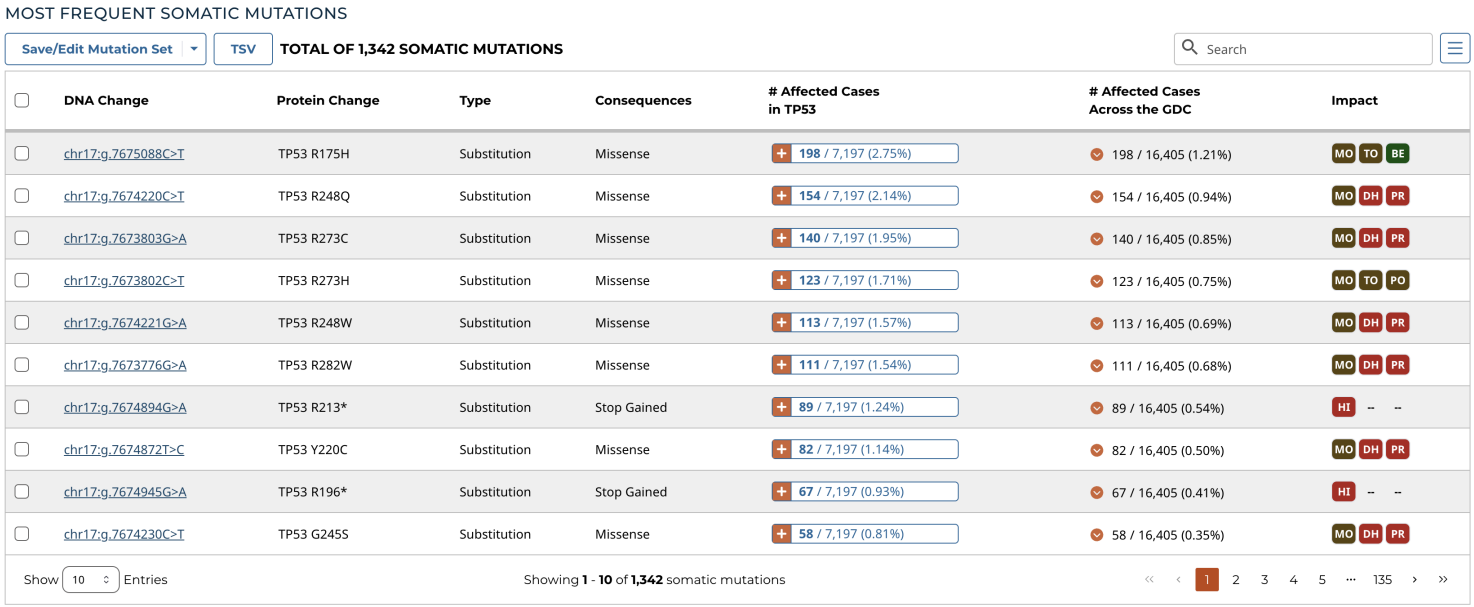
A list with links that lead to external databases with additional information about each gene is displayed here. These external databases include: Entrez, Uniprot, Hugo Gene Nomenclature Committee, Online Mendelian Inheritance in Man, Ensembl, and CIViC.

Cancer Distribution

A table and two bar graphs (one for mutations, one for CNV events) show how many cases are affected by mutations and CNV events within the gene as a ratio and percentage. Each row/bar represents the number of cases for each project. The final column in the table lists the number of unique mutations observed on the gene for each project.

Most Frequent Mutations

The 20 most frequent mutations in the gene are displayed as a bar graph that indicates the number of cases that share each mutation.



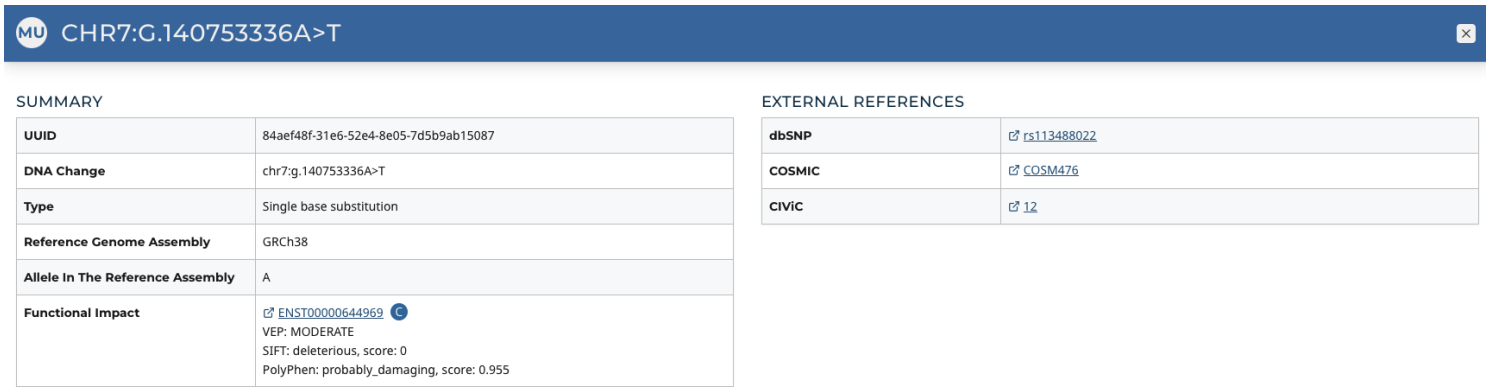
A table is displayed below that lists information about each mutation including:

- **DNA Change:** The chromosome and starting coordinates of the mutation are displayed along with the nucleotide differences between the reference and tumor allele
- **Protein Change:** The gene and amino acid change
- **Type:** A general classification of the mutation
- **Consequences:** The effects the mutation has on the gene coding for a protein (i.e. synonymous, missense, non-coding transcript)
- ___ Affected Cases in Gene:___ The number of affected cases, expressed as number across all mutations within the Gene
- ___ Affected Cases Across GDC:___ The number of affected cases, expressed as number across all projects. Choosing the arrow next to the percentage will expand the selection with a breakdown of each affected project.
- **Impact:** A subjective classification of the severity of the variant consequence. This is determined using Ensembl VEP, PolyPhen, and SIFT. The categories are outlined here.

Note: The Mutation UUID can be displayed in this table by selecting it from the Customize Columns button, represented by three parallel lines

The Mutation Summary Page contains information about one somatic mutation and how it affects the associated gene. Each mutation is identified by its chromosomal position and nucleotide-level change.

Summary



- **UUID:** A unique identifier (UUID) for this mutation
- **DNA Change:** Denotes the chromosome number, position, and nucleotide change of the mutation

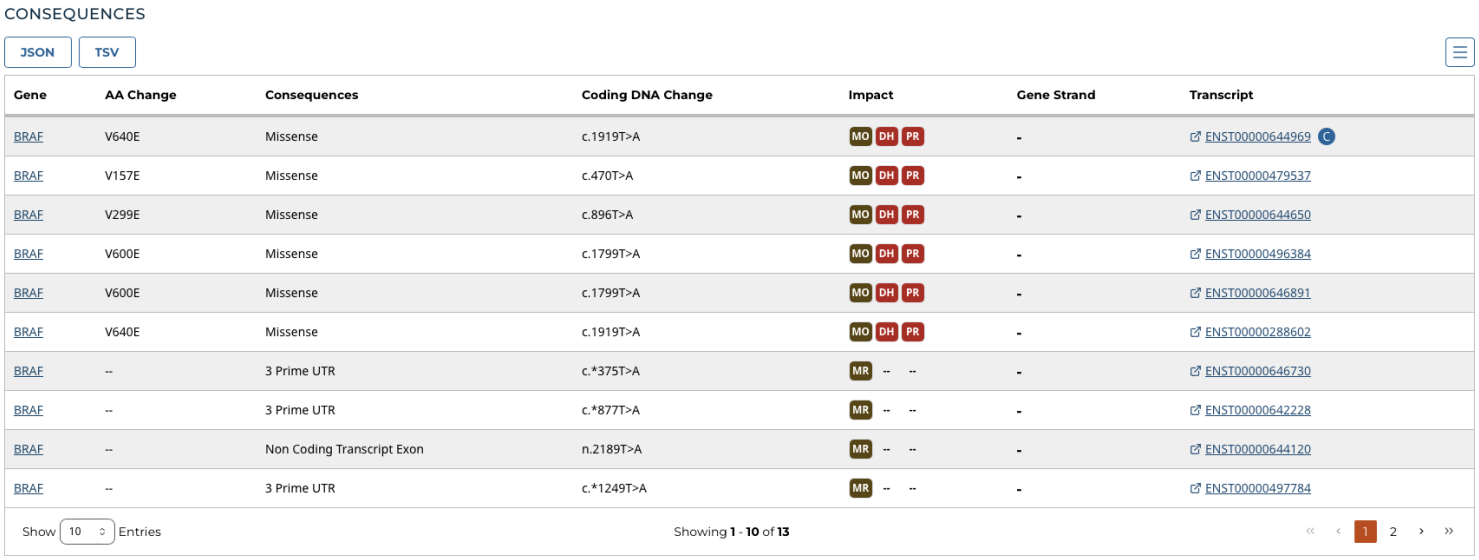
- **Type:** A broad categorization of the mutation
- **Reference Genome Assembly:** The reference genome in which the chromosomal position refers to
- **Allele in the Reference Assembly:** The nucleotide(s) that compose the site in the reference assembly
- **Functional Impact:** A subjective classification of the severity of the variant consequence.

External References

A separate panel contains links to databases that contain information about the specific mutation. These include dbSNP, COSMIC, and CIViC.

Consequences

The consequences of the mutation are displayed in a table. The set of consequence terms, defined by the Sequence Ontology.


CONSEQUENCES

JSON TSV

Gene	AA Change	Consequences	Coding DNA Change	Impact	Gene Strand	Transcript
BRAF	V640E	Missense	c.1919T>A	MO DH PR	-	ENST00000644969 C
BRAF	V157E	Missense	c.470T>A	MO DH PR	-	ENST00000479537
BRAF	V299E	Missense	c.896T>A	MO DH PR	-	ENST00000644650
BRAF	V600E	Missense	c.1799T>A	MO DH PR	-	ENST00000496384
BRAF	V600E	Missense	c.1799T>A	MO DH PR	-	ENST00000646891
BRAF	V640E	Missense	c.1919T>A	MO DH PR	-	ENST00000288602
BRAF	--	3 Prime UTR	c.*375T>A	MR -- --	-	ENST00000646730
BRAF	--	3 Prime UTR	c.*877T>A	MR -- --	-	ENST00000642228
BRAF	--	Non Coding Transcript Exon	n.2189T>A	MR -- --	-	ENST00000644120
BRAF	--	3 Prime UTR	c.*1249T>A	MR -- --	-	ENST00000497784

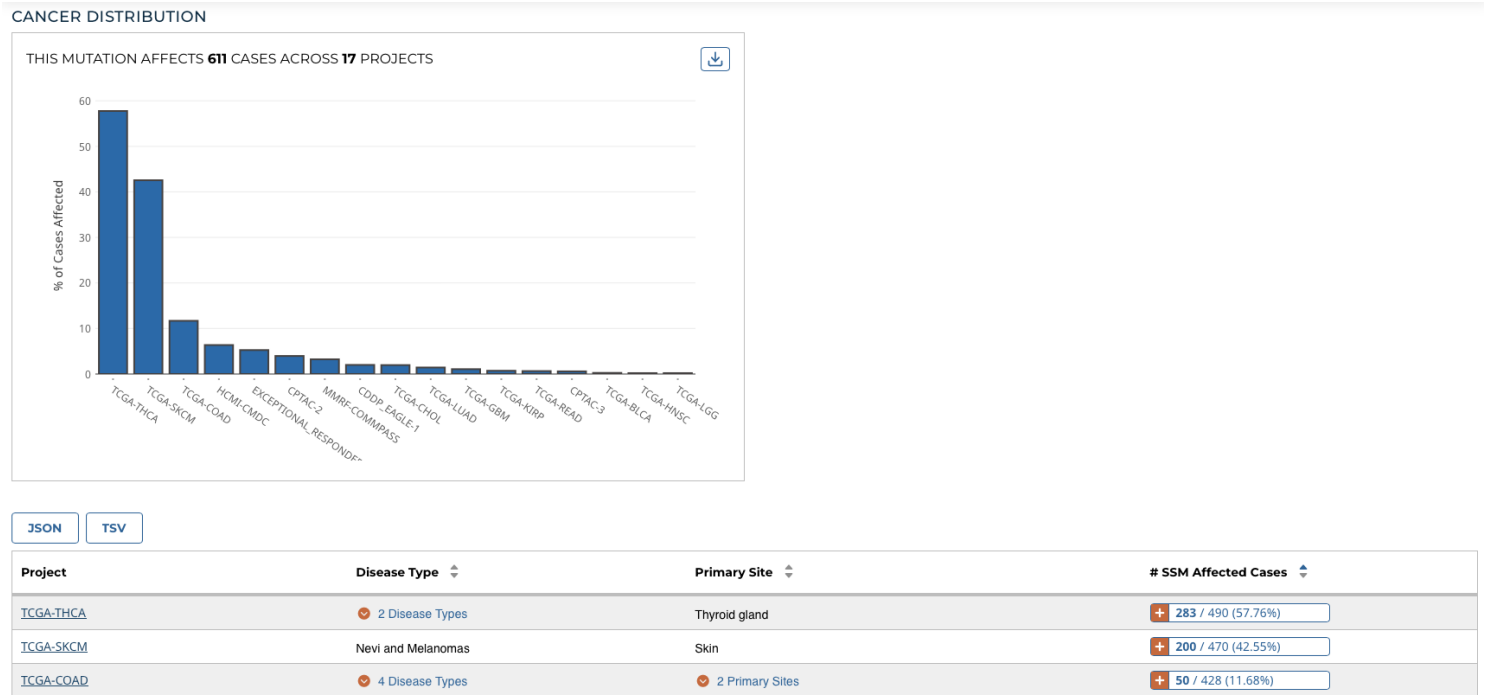
Show 10 Entries Showing 1 - 10 of 13 « ‹ 1 2 › »

The fields that describe each consequence are listed below:

- **Gene:** The symbol for the affected gene
- **AA Change:** Details on the amino acid change, including compounds and position, if applicable
- **Consequences:** The biological consequence of each mutation
- **Coding DNA Change:** The specific nucleotide change and position of the mutation within the gene
- **Impact:** VEP, SIFT, and/or PolyPhen Impact ratings
- **Gene Strand:** If the gene is located on the forward (+) or reverse (-) strand
- **Transcript:** The transcript(s) affected by the mutation. Each contains a link to the Ensembl entry for the transcript

Cancer Distribution

A table and bar graph shows how many cases are affected by the particular mutation. Each row/bar represents the number of cases for each project.



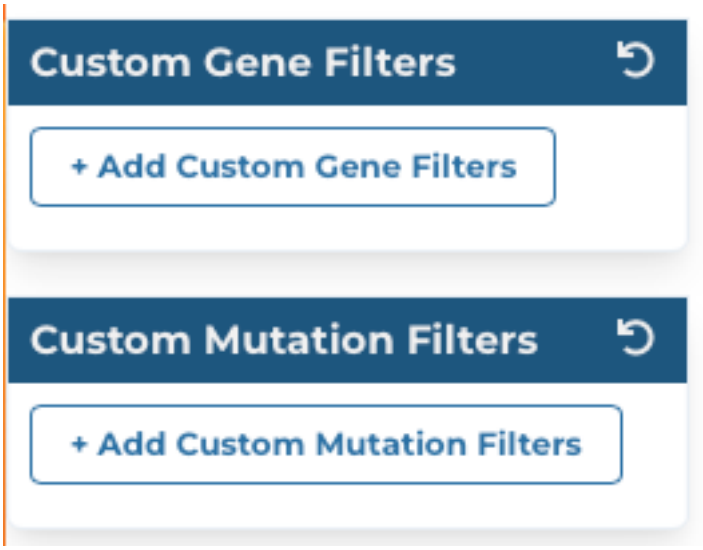
JSON TSV

Project	Disease Type	Primary Site	# SSM Affected Cases
TCGA-THCA	2 Disease Types	Thyroid gland	283 / 490 (57.76%)
TCGA-SKCM	Nevi and Melanomas	Skin	200 / 470 (42.55%)
TCGA-COAD	4 Disease Types	2 Primary Sites	50 / 428 (11.68%)

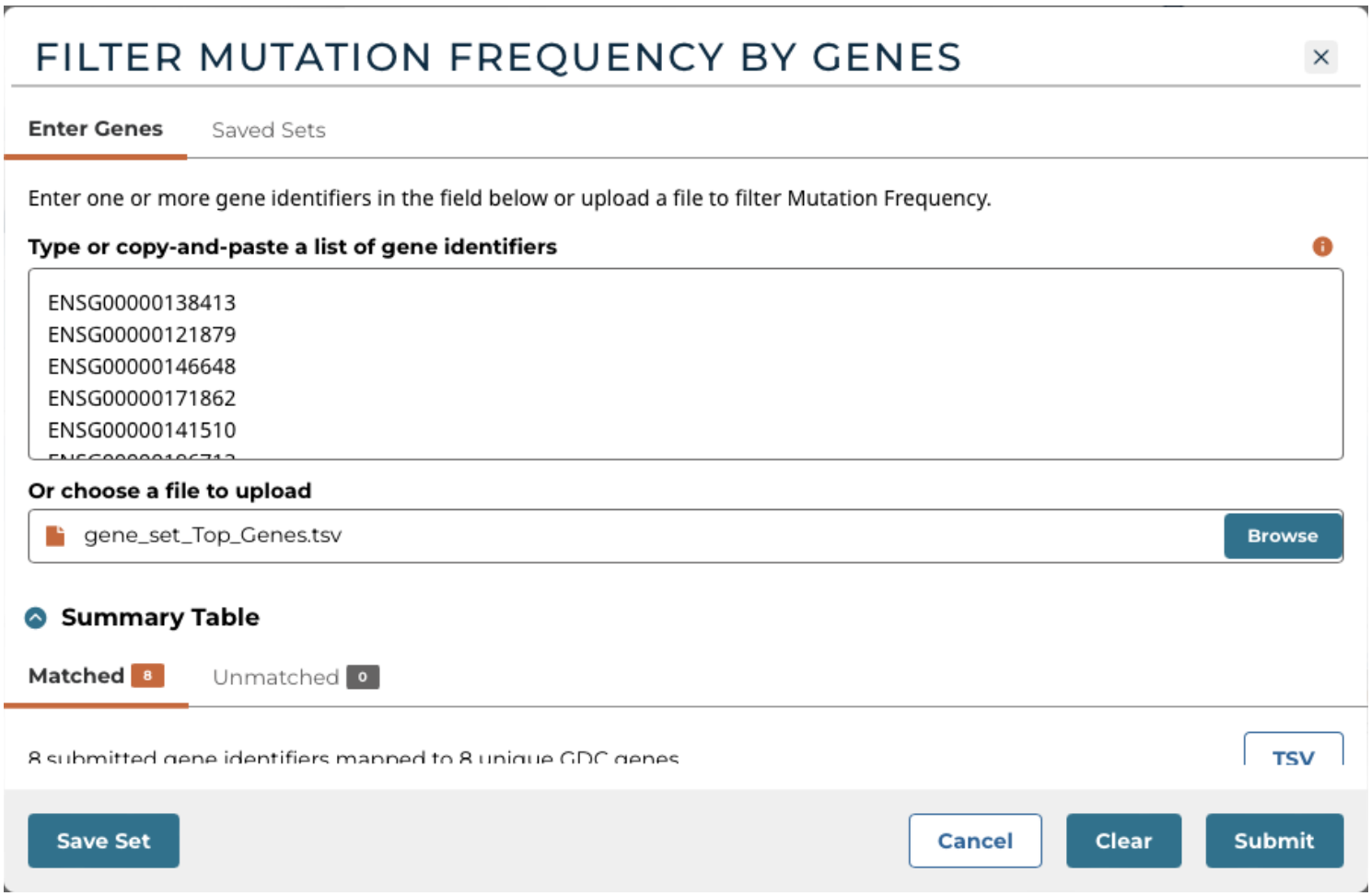
The table contains the following fields:

- **Project**: The ID for a specific project
- **Disease Type**: The disease associated with the project
- **Primary Site**: The anatomical site affected by the disease
- ___ SSM Affected Cases___: The number of affected cases and total number of cases displayed as a fraction and percentage

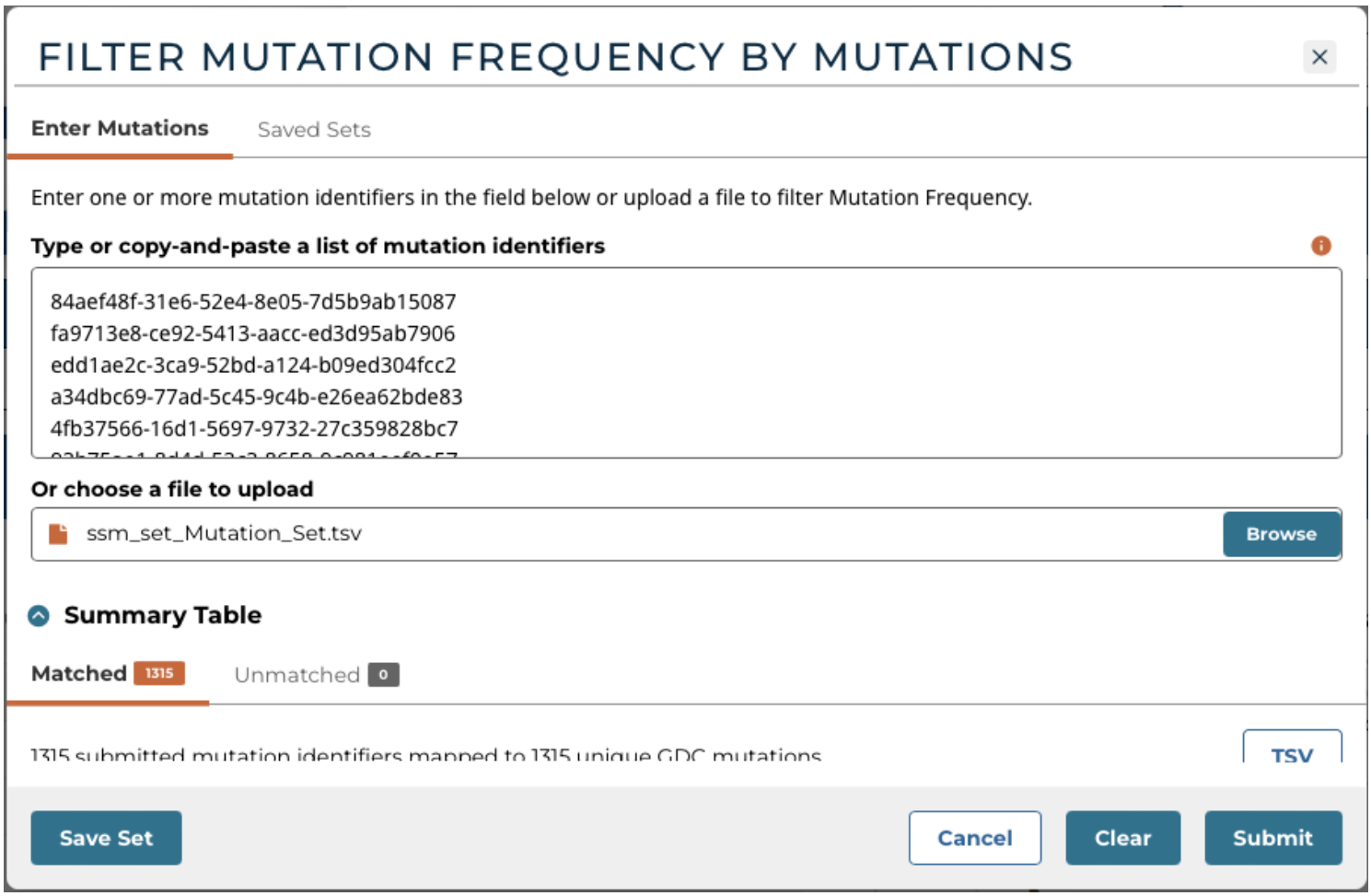
Custom Gene and Mutation Filters



The `+ Add Custom Gene Filters` button in the left panel of the Mutation Frequency tool allows users to filter mutation frequency by genes. Users can enter unique identifiers (i.e. gene symbols, gene IDs, etc.) directly into the text box as a plain text list or upload a list of unique identifiers as a CSV, TSV, or TXT file. Users can hover over the orange (i) to verify accepted gene identifiers, delimiters, and file formats.



The `+ Add Custom Mutation Filters` button allows users to filter mutation frequency by mutations. Users can enter unique identifiers (i.e. mutation UUIDs, etc.) directly into the text box as a plain text list or upload a list of unique identifiers as a CSV, TSV, or TXT file. Users can hover over the orange (i) to verify accepted mutation identifiers, delimiters, and file formats.



Mutation Frequency Facet Filters

A set of frequently-used properties are available to filter genes and mutations in the left panel of the Mutation Frequency tool. Using each of these filters will dynamically change the graphics and table to represent the filtered data.

- **Biotype:** Classification of the type of gene according to Ensembl. The biotypes can be grouped into protein coding, pseudogene, long noncoding and short noncoding. Examples of biotypes in each group are as follows:
  - **Protein coding:** IGC gene, IGD gene, IG gene, IGJ gene, IGLV gene, IGM gene, IGV gene, IGZ gene, nonsense mediated decay, nontranslating CDS, non stop decay, polymorphic pseudogene, TRC gene, TRD gene, TRJ gene, TRV gene.
  - **Pseudogene:** Disrupted domain, IGC pseudogene, IGJ pseudogene, IG pseudogene, IGV pseudogene, processed pseudogene, transcribed processed pseudogene, transcribed unitary pseudogene, transcribed unprocessed pseudogene, translated processed pseudogene, translated unprocessed pseudogene, TRJ pseudogene, TRV pseudogene, unprocessed pseudogene.
  - **Long noncoding:** 3 prime overlapping ncrna, ambiguous orf, antisense, antisense RNA, lincRNA, macro lincRNA, ncrna host, processed transcript, sense intronic, sense overlapping.
  - **Short noncoding:** miRNA, miRNA pseudogene, miscRNA, miscRNA pseudogene, Mt rRNA, Mt tRNA, rRNA, scRNA, snlRNA, snoRNA, snRNA, tRNA, tRNA pseudogene, vaultRNA.
- **Is Cancer Gene Census:** Whether or not a gene is part of The Cancer Gene Census. Note that this is switched on as a default.
- **Impact:** A subjective classification of the severity of the variant consequence. These scores are determined using the following three tools:
  - **VEP:**
    - **HIGH (H):** The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay
    - **MODERATE (M):** A non-disruptive variant that might change protein effectiveness

    - **LOW (L):** Assumed to be mostly harmless or unlikely to change protein behavior
    - **MODIFIER (MO):** Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact
  - **PolyPhen:**
    - **probably damaging (PR):** It is with high confidence supposed to affect protein function or structure
    - **possibly damaging (PO):** It is supposed to affect protein function or structure
    - **benign (BE):** Most likely lacking any phenotypic effect
    - **unknown (UN):** When in some rare cases, the lack of data does not allow PolyPhen to make a prediction
  - **SIFT:**
    - **tolerated:** Not likely to have a phenotypic effect
    - **tolerated_low_confidence:** More likely to have a phenotypic effect than 'tolerated'
    - **deleterious:** Likely to have a phenotypic effect
    - **deleterious_low_confidence:** Less likely to have a phenotypic effect than 'deleterious'
- **Consequence Type:** Consequence type of this variation; sequence ontology terms
- **Type:** A general classification of the mutation

Saving a Gene or Mutation Set

After filtration, a set of genes or mutations can be saved by choosing the "Save/Edit Gene Set" or "Save/Edit Mutation Set" button at the top left of the table.

Chapter 12

OncoMatrix

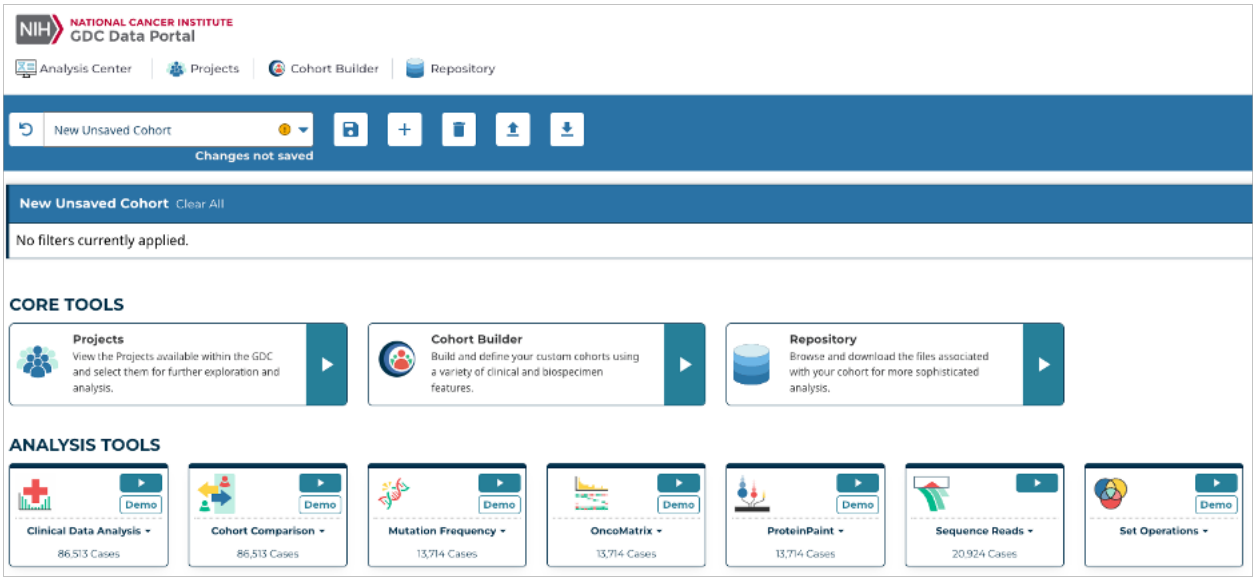
OncoMatrix

Introduction to OncoMatrix

The OncoMatrix tool is a web-based tool for visualizing coding mutations such as Simple Somatic Mutations (SSM) and Copy Number Variations (CNV) from the NCI Genomic Data Commons (GDC).

Accessing the OncoMatrix Chart

At the Analysis Center, click on the 'OncoMatrix' card to launch the app.



Users can view publicly available genes as well as login with credentials to access controlled data.

Quick Reference Guide

There are three main panels in the OncoMatrix tool: control panel, matrix plot, and legend panel.

Control Panel

The control panel has various functionalities with which users can change or modify the appearance of the matrix. The control panel provides flexibility and a wide range of options to maximize user control.

Control Panel:

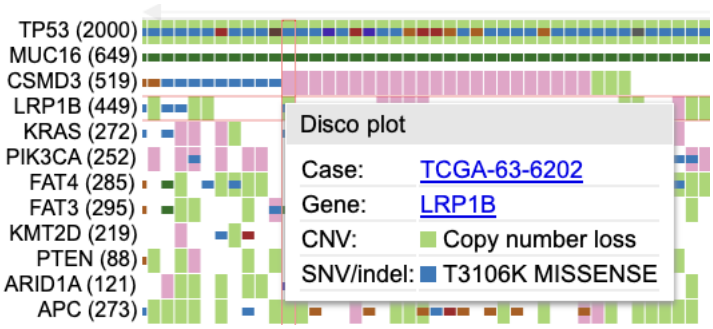
- **Cases:** Choose how to sort the cases, specify the maximum number of cases to display, group cases according to selected GDC variables, and adjust the visible characters of the case labels
- **Genes:** Modify how cases are represented for each gene (Absolute, Percent, or None), row group and label lengths, rendering style, how genes are sorted, the maximum number of genes displayed, and the existing gene set
  - **Edit Group:** Displays a panel of currently selected genes, which can be modified by clicking on a gene to remove it from the gene set, searching for a particular gene to add, loading top variably expressed genes, or loading a pre-defined gene set provided by the MSigDB database
  - **Create Group:** Create a new gene set by searching for a particular gene, loading top mutated genes, or loading a pre-defined gene set provided by the MSigDB database
- **Variables:** Search and select variables to add to the bottom of the matrix
- **Cell Layout:** Modify the format of the cells by changing colors, cell dimensions and spacing, and label formatting
- **Legend Layout:** Alter the legend by changing the font size, dimensions and spacing, and other formatting preferences
- **Download:** Download the matrix in svg format
- **Zoom:** Adjust the zoom level by using the up and down arrows on the input box, entering a number, or using the sliding scale to view the case labels.

Matrix Plot

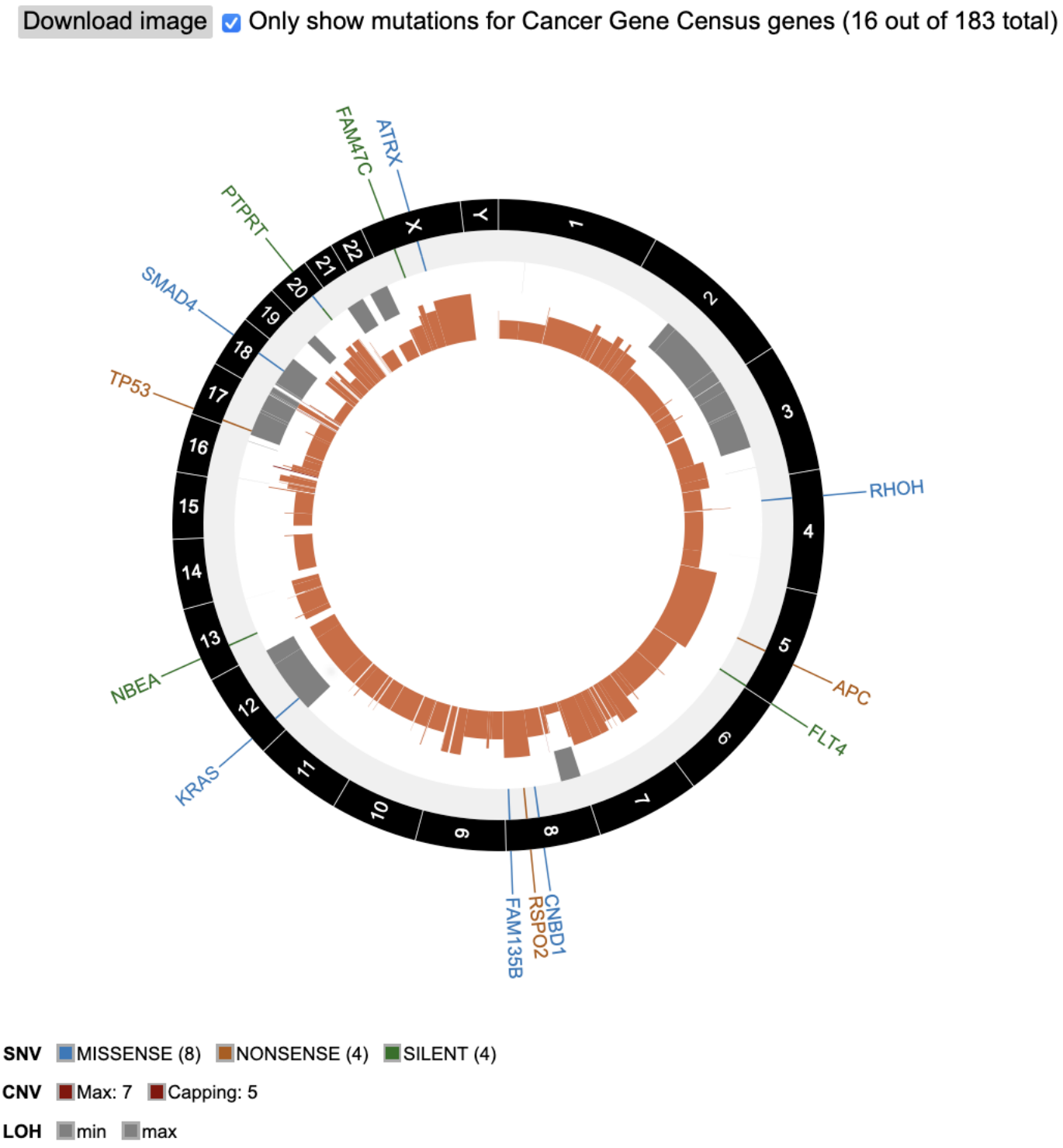
The OncoMatrix plot displays the genes along the left panel with each column representing a case.

Matrix cells

Each column in the matrix represents a case. Hovering over a cell will display the corresponding case submitter_id, gene name, copy number information, and mutation class if any are provided. Clicking on a cell also gives users the option to launch the Disco Plot.

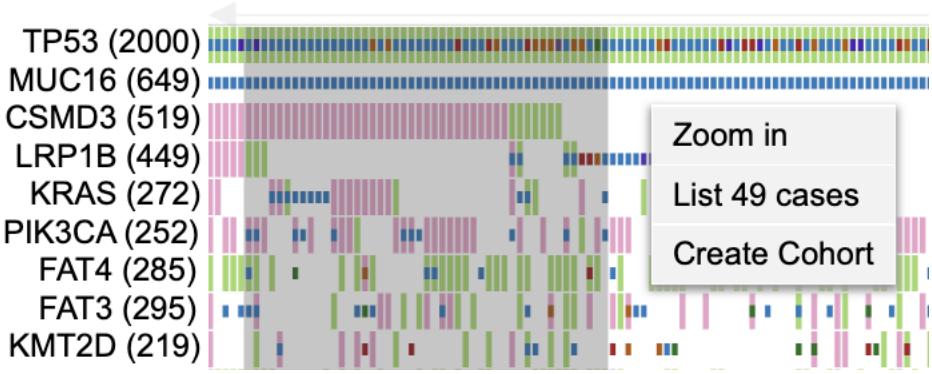


The Disco Plot is a circular plot that shows all the mutations and CNVs for a given case. The Disco Plot also displays the legend for the mutation class and the CNV.



Automatic Zoom

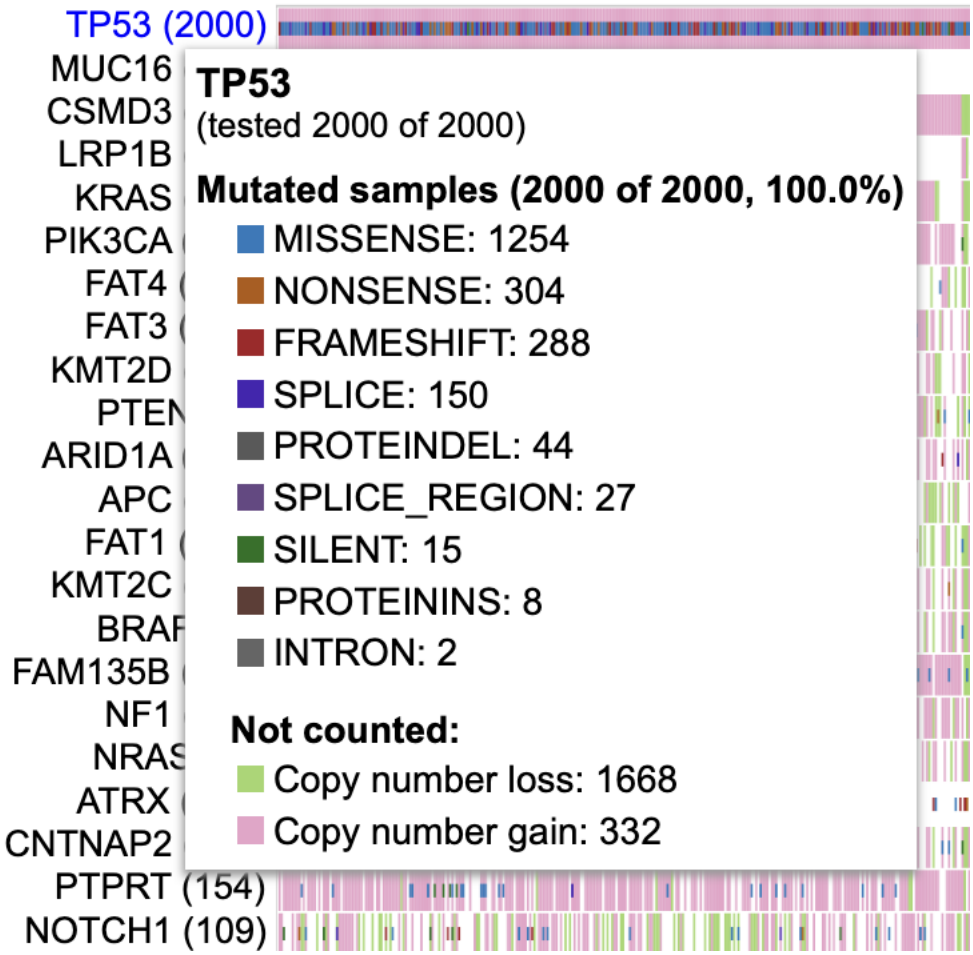
To perform an automatic zoom, users can click on and hold a case column then drag the mouse from left to right to form a zoom boundary. From the pop-up window, users can choose to zoom in to the cases, list all highlighted cases, or create a cohort of the selected cases.



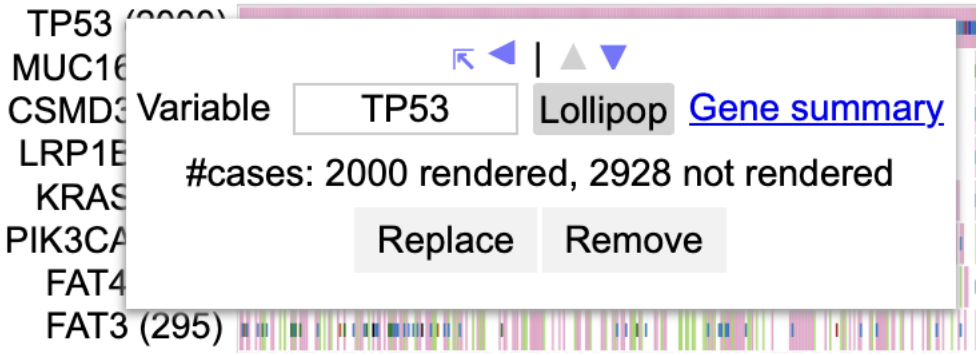
The individual case columns are now visible with a demarcated boundary. Above the cases, a slider has been provided for moving from one view to another to accommodate all cases.

Genes

In the panel of genes on the left, users can hover over a gene to view the number of mutated samples, a breakdown of consequence type, and copy number gain and loss counts.

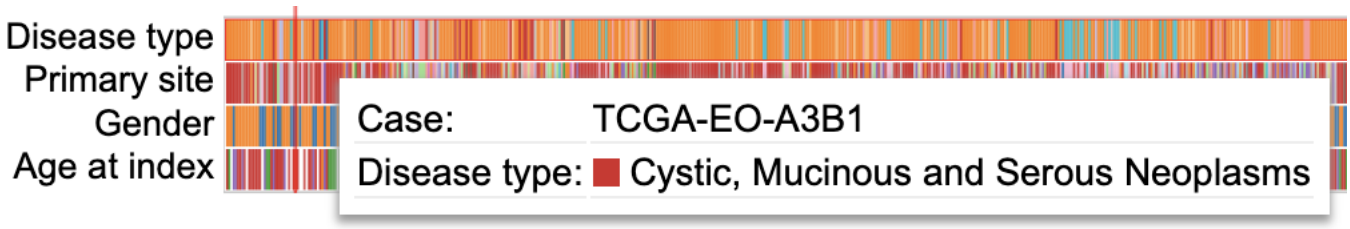


Clicking on a gene opens a pop-up window where users can rename it, launch the ProteinPaint Lollipop plot, display the Gene Summary Page, and replace or remove the gene. The lollipop plot displays all cases across the GDC affected by SSMs in the selected gene.



Variables

Any variables added to the matrix appear at the bottom of the plot. Users can hover over a cell in a variable row to display the case submitter_id and their value for the given variable.



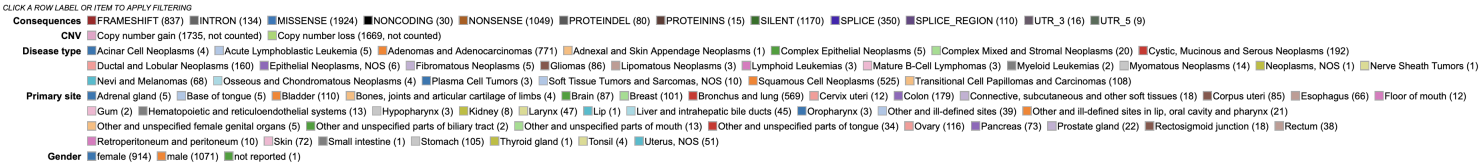
Clicking on a variable allows users to rename it, edit it by excluding categories, replace it with a different variable, or remove it entirely.

Drag and drop genes and variables

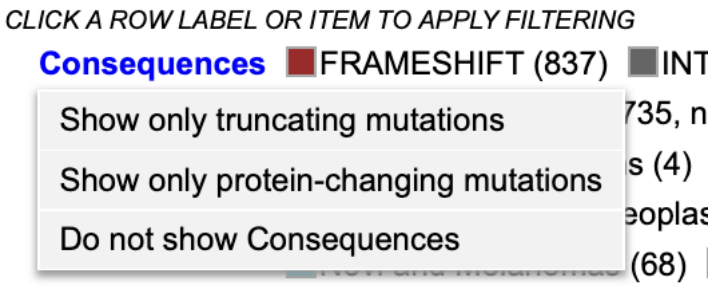
By default, the genes in the matrix are sorted in descending order according to which genes have the highest number of rendered cases. Users can override this by dragging and dropping gene and variable row labels to sort the rows manually.

Legend Panel

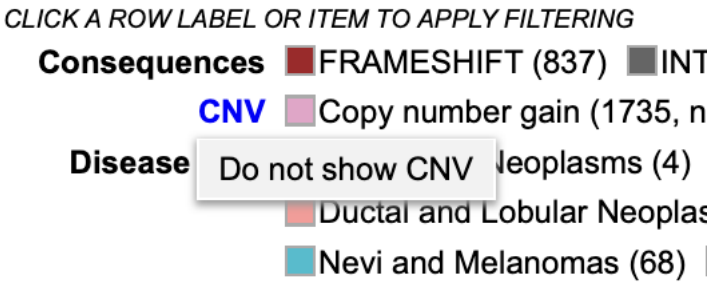
Below the matrix, the legend displays color coding for mutation classes, CNV, as well as each variable that is selected to appear in the plot.



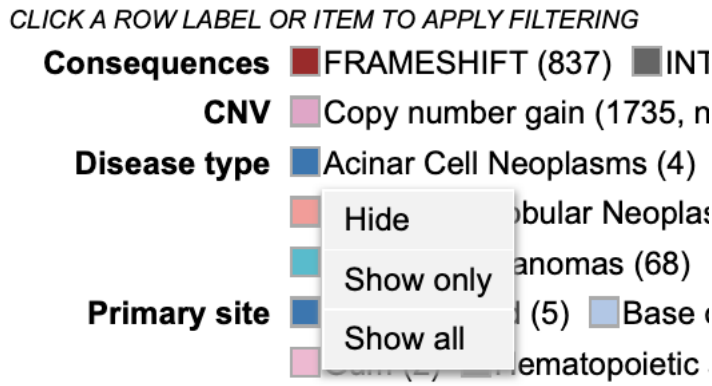
Clicking on `Consequences` offers options to show only truncating mutations, show only protein-changing mutations, or hide consequences.



Clicking on `CNV` allows users to hide CNV.



Additionally, users can click on a variable's category to hide a specific group, only show a specific group, or show all groups for the selected variable.



Features

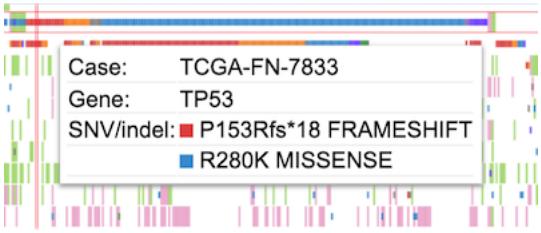
The following features are viewable once the matrix application is loaded. There are three main panels as outlined in the figure below i.e., the `Control panel`, `Matrix chart`, and the `Legend panel`.

Each of the features and functionalities are described in detail in the following sections.

Matrix plot

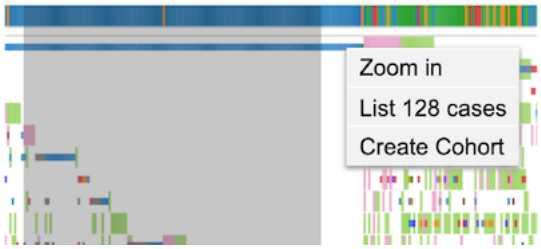
Hovering on sample columns

Each column in the matrix represents a sample. Hover over sample cells/columns to display information about the sample such as case id, gene name, Copy number information and mutation/mutation class (if any provided) as shown.

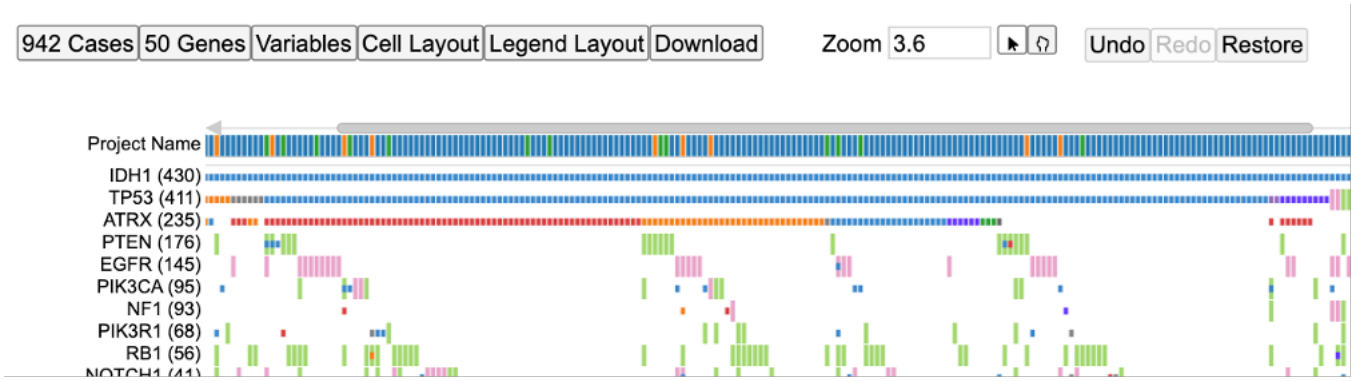


Drag to zoom

A user may click a row label and drag it while keeping the mouse button down, to sort the rows manually. Click and hold on a column of sample and drag the mouse from left to right to form a zoom boundary as shown in the image and leave the mouse.



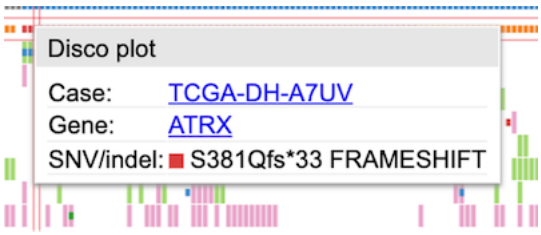
This allows for an automatic zoom as shown. The individual sample columns are now visible with a well demarcated boundary. Above the samples, a slider (as shown in gray) has been provided for moving from one view to another to accommodate all cases.



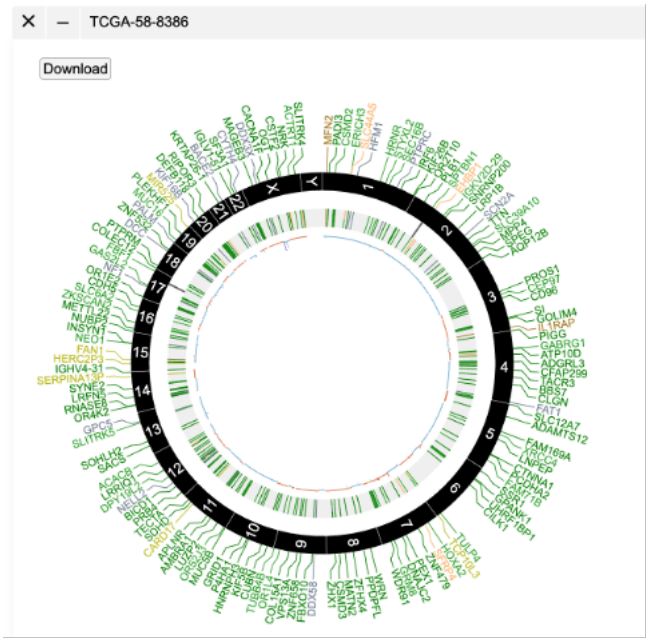
Additionally, to have a finer control on the zoom the user may follow the steps outlined in the section - Zooming

Clicking on Sample columns

In the same zoomed in view as shown above, click on any sample column for TP53. This displays a clickable button `Disco plot` as shown.



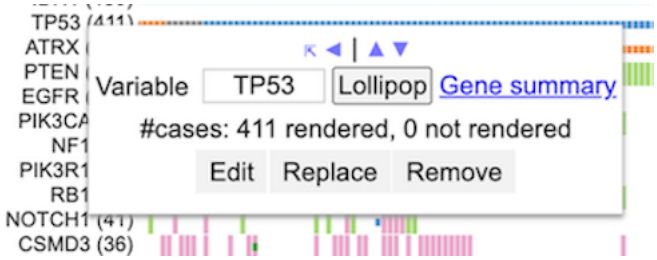
Click on the disco plot button to display a circular plot that shows all the mutations for a given sample as shown.



The disco plot can also be accessed by following steps outlined in the section - Disco Plot

Clicking on gene/variable labels

Click on TP53 gene label to display the following options.

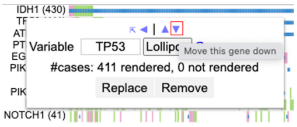


The first row in the options highlighted by a red box as shown in the image above allows the user to sort rows and move rows up and down (please note that rows can also be moved by dragging and dropping as outlined in section Drag and Drop Gene Label/Variable variable). Every time a sorting icon is clicked the chart will update and reload.

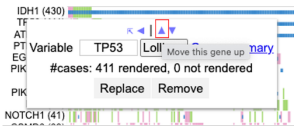
Click the first arrow as shown by clicking the gene label TP53. This will sort the samples against the gene at the top left corner which is TP53 in this example.

Next, click on the left arrow as shown. This allows for sorting samples against the gene.

Now click the down arrow as shown. The row with TP53 cases will move below ATRX.



Click the gene label `TP53` and click the up arrow as shown.

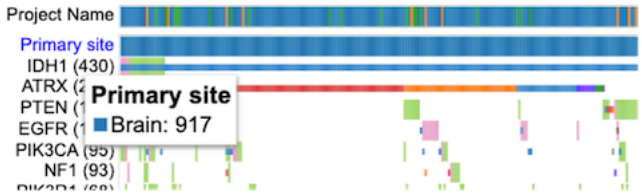


The row containing TP53 cases now moves back up in position 1 above ATRX.

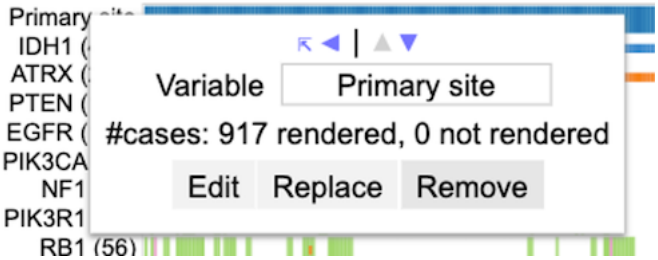
Click TP53 again to showcase the edit menu.

Click on 'Replace' as shown above to replace TP53 gene variable with 'Primary site' as shown below. The chart updates with the first row as 'Primary site' thereby replacing TP53 gene variable as shown below. User may choose to sort samples by clicking the 'Primary site' label.

Replace a gene row](/home/ubuntu/gdc-docs-qa/docs/Data_Portal/Users_Guide/./images/oncomatrix/13-replace_gene.png 'Click to see the full image.')



Click on the label 'Primary site' and click the option 'Remove' as shown to remove the row completely.



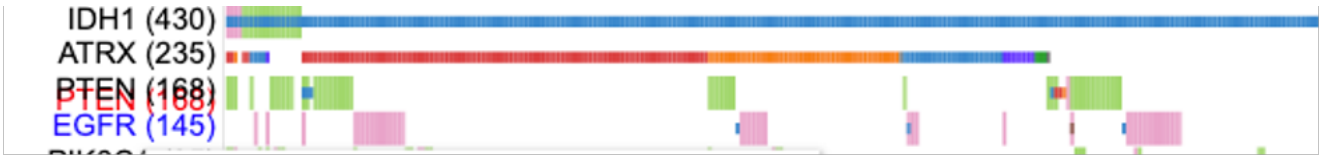
This updates the chart. User may choose to add back TP53 through the gene panel.

Click on `Replace` as shown above to replace TP53 gene variable with `Primary site` as shown below.

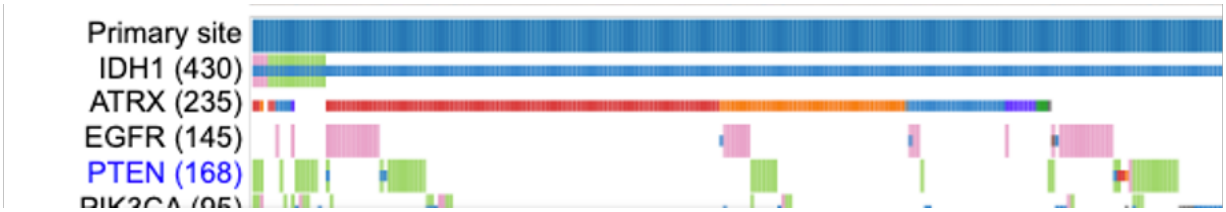
Drag and Drop Gene Label/Variable

The genes on the matrix are sorted by default on the number of cases with the gene having the highest number of cases at the top of the matrix. A user may choose to override this by dragging a gene label and dropping it above or below any other gene in order to customize their own gene groupings.

Select 'PTEN' gene label and drag it below the gene labeled 'EGFR' as shown. When dragging a gene label, hover over EGFR such that the EGFR gene label would appear blue.



When the EGFR gene label appears blue, then drop the PTEN gene label row. The display updates to show PTEN below EGFR as shown below.

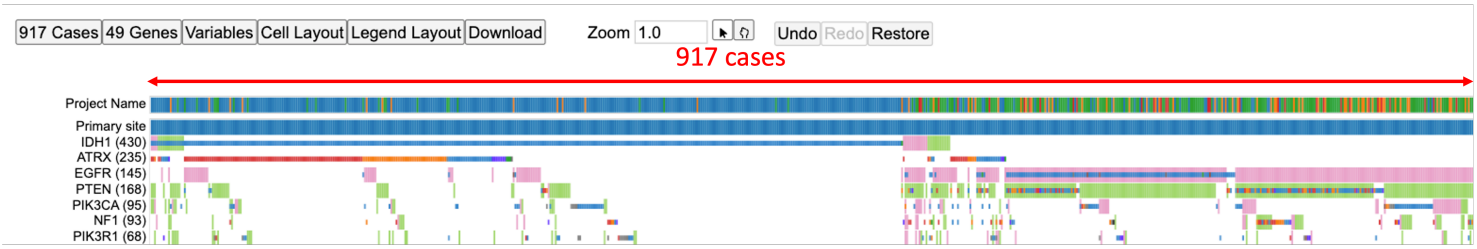


Control panel

The control panel as shown has various functionalities with which users can change or modify the appearance of the matrix. The control panel provides flexibility and a wide range of options to maximize user control.

Cases

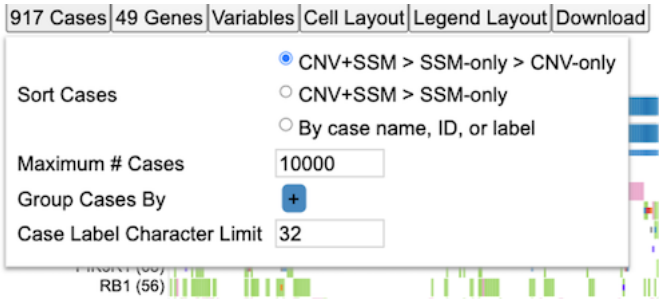
Within the control panel, the first button displays the number of cases that are shown as columns of the matrix. The default view is as shown.



Click on the `917 Cases` button to display the following options as shown.

1. Sort Cases
2. Maximum #cases
3. Group cases by
4. Sort Case Groups
5. Case Group Label Max Length
6. Case Label Max Length

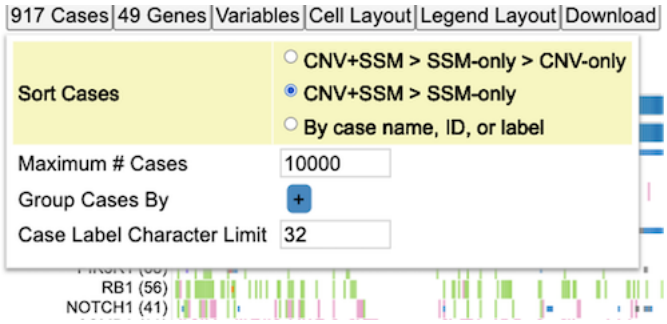
These sections are described below.



Sort Cases

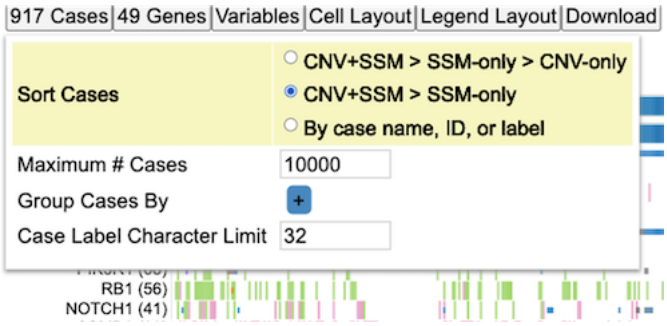
The default sort setting sorts the cases by row with first displaying samples with both CNV and SSM followed by SSM only and lastly CNV only.

Click the second option 'CNV+SSM > SSM only" to change the sorting as shown.



Maximum #cases

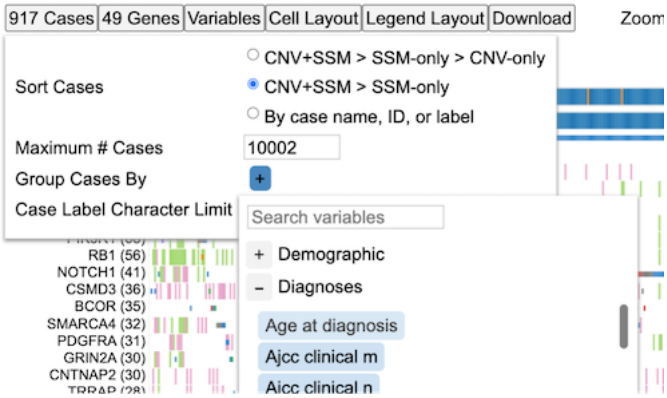
There is a default number of samples that are shown in the matrix chart. Users can choose to increase or decrease the number of samples. This allows the chart to re-render and display the number of columns based on the user's selection. Figure below shows increased cases to 10000. Please note that any high arbitrary number can be selected but the chart will only show the maximum cases that GDC has.



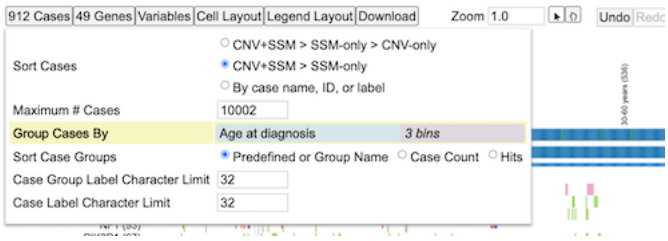
The chart will reload with new cases added.

Group cases by

This option allows users to group cases by different variables from the GDC dictionary. Click on the `+` icon shown in blue to display different variables such as demographics, diagnoses, Exposures etc. Users may also search for a variable from the search bar provided in the menu as shown by `Search Variables`.

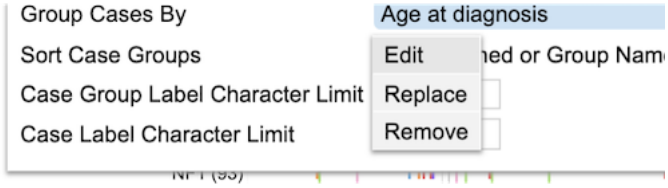


Click 'Age at diagnosis' from the options. The matrix reloads to show the following view.

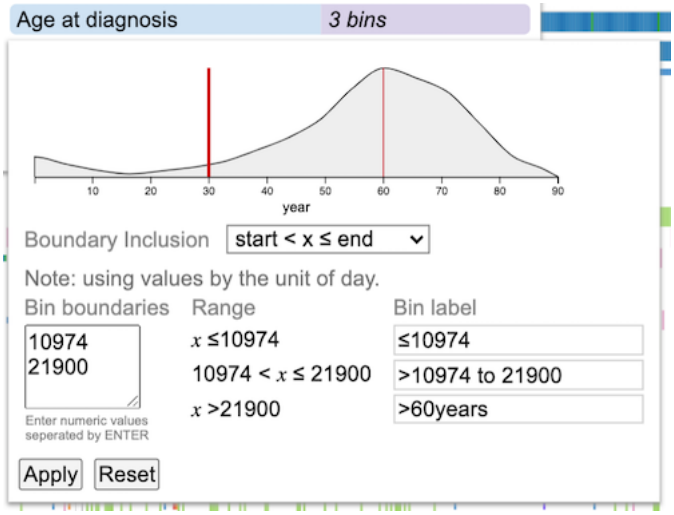


As shown above, labels for different age groups show up vertically and all cases get distributed with a clearcut separation according to the age bins.

Click on the 'Age at diagnosis' (blue pill, as shown). This opens a short menu with action items. Click on the first item 'Edit' as shown.

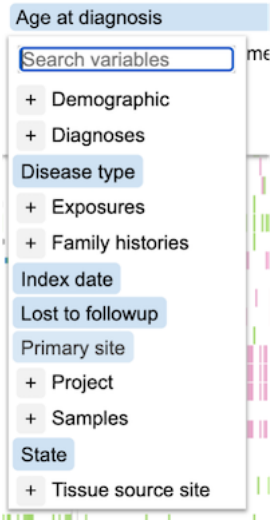


Drag the red lines on the density distribution to select binning or input numbers for custom binning and select 'Apply'.



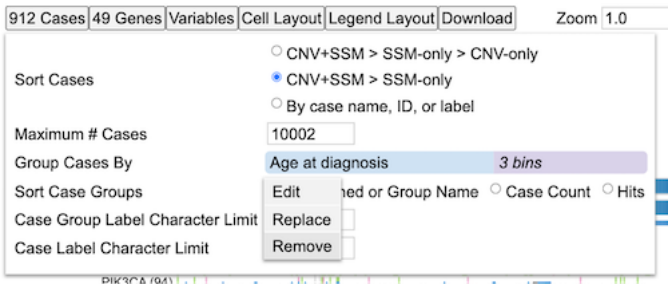
The matrix reloads with new bin groupings The labels for the groups are user controlled and hence can be modified according to user requirements.

Click on the blue pill for 'Age at diagnosis' again and click 'Replace'. Select 'Primary site' as shown.



The matrix reloads with the new variable distribution.

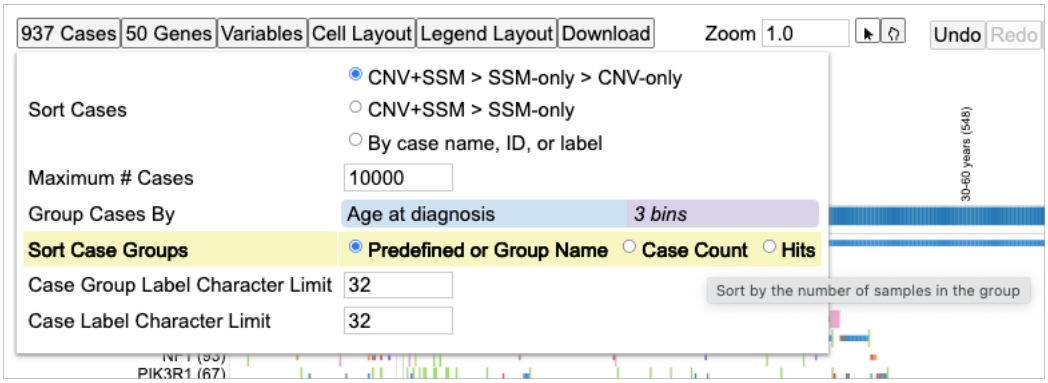
The last option on the menu is 'Remove'. Click on the '917 Cases' button, followed by 'Age at diagnosis' shown in blue to reveal the menu option. Click 'Remove' to completely get rid of any groups.



This will remove all and any groupings and show the default view again.

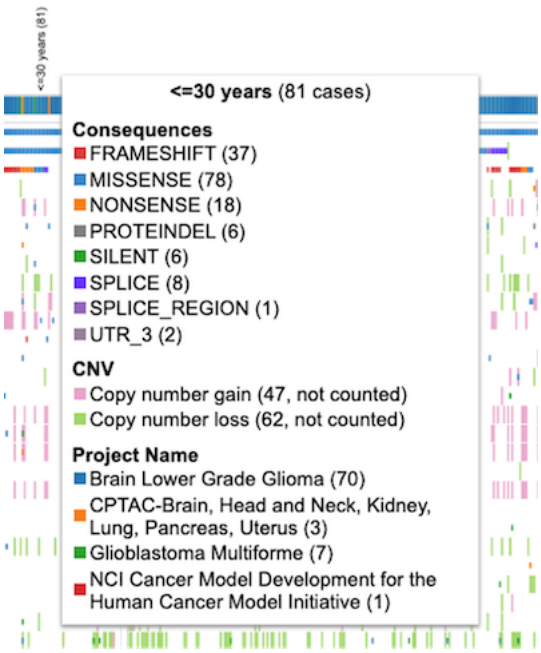
Sort Case Groups

Add the variable 'Age at diagnosis' again using the 'Group Cases by' button as shown in the previous section. By default, groups are loaded ordered by their name. Change the selection to 'Case count' as shown below.



The third selection option 'Hits' orders the groupings based on the number of gene variants for a particular case or case group for the genes in display. Click 'Hits' under 'Sort Case Groups' to change the order of groupings.

Next, hover over the first group label '<=30 years (81)' as shown below.

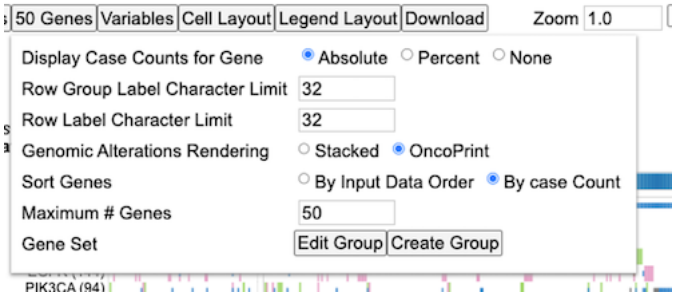


This shows the number of cases in parenthesis of the group label and the breakdown for the number of variants and CNV for all the samples within that group for the genes in display.

Genes

The gene panel as shown below has several options as listed below for modifying the genes visible on the plot as well as their appearance/style.

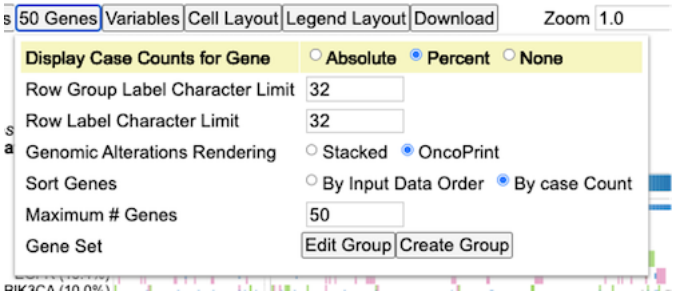
- Display Case Counts for Gene
- Rendering Style
- Sort Genes
- Maximum Genes
- Gene Set



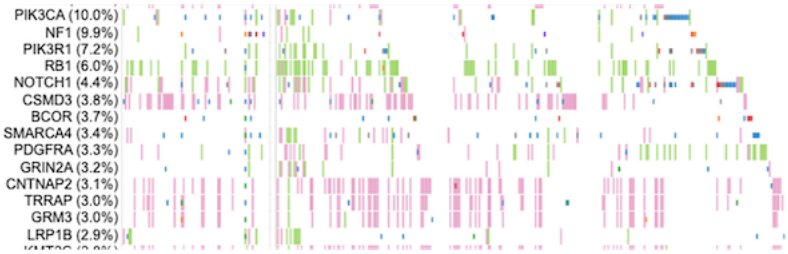
Display Case Counts for Gene

This option allows change in the number of cases that is represented in parentheses next to the gene variable label as shown below. By default, the number of cases for each gene is an `Absolute`.

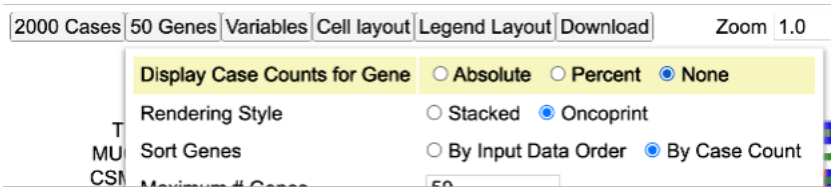
Click on the button `50 Genes` to display the menu and select `Percent` as shown below.



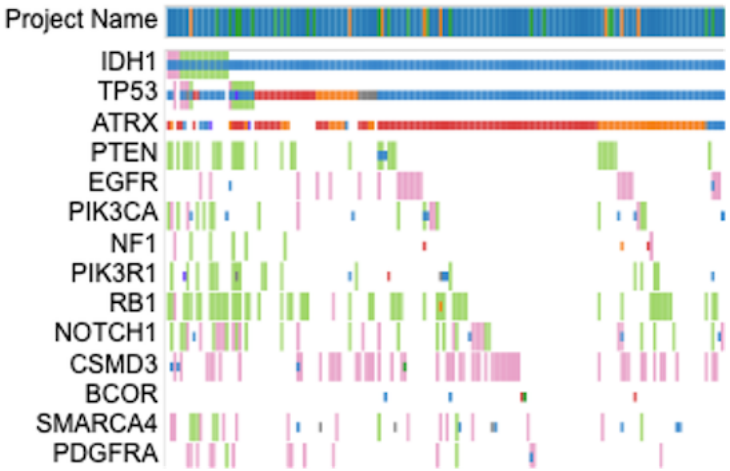
This shows the case counts as a percentage of the absolute values as shown.



User has the option to hide the display of case counts. Click `Genes` button again and select `None` for `Display Case Counts for Gene` as shown below

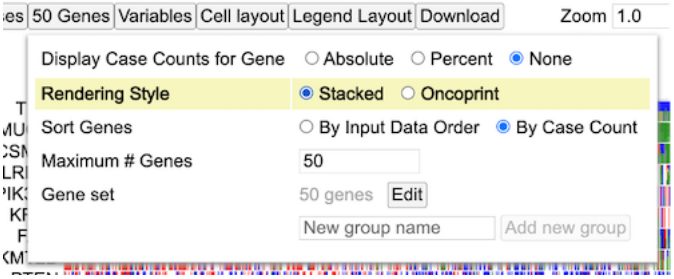


This hides all the case counts as shown.

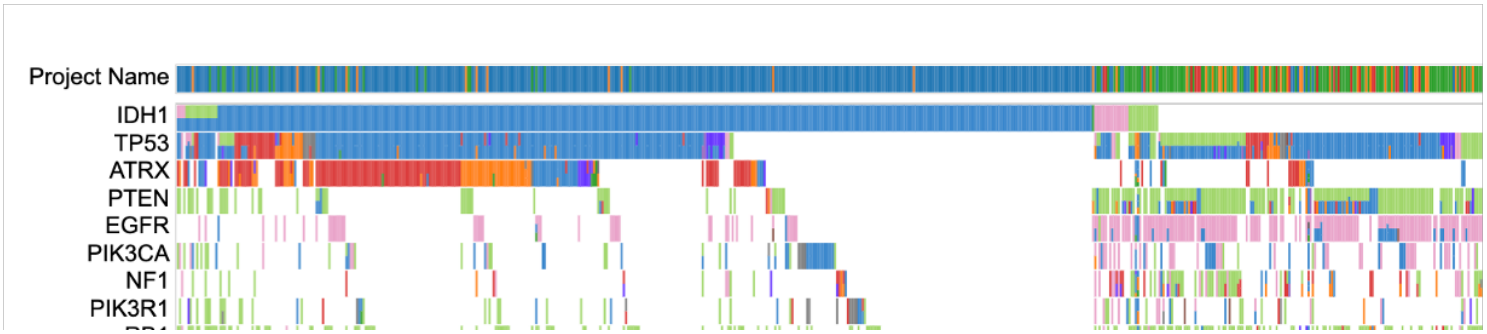


Rendering Style

The style of rendering for the sample cells/columns is an Oncoprint style by default. Click on `Stacked` option via `50 Genes` button as shown below.

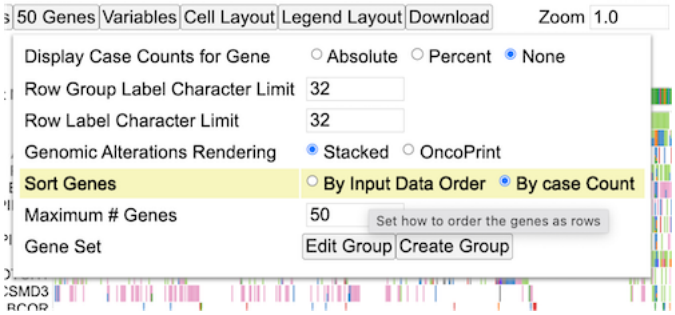


The mutations and CNV are now stacked on top of each other as shown below.



Sort Genes

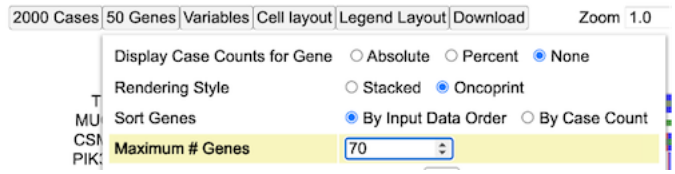
The default sorting option for genes is `By Case Count`. This means the genes are sorted by the number of cases from increasing to decreasing order. Click `50 Genes` button on the control panel, and select `By Input Data Order` under the `Sort Genes` as shown below.



The genes will now sort according to the order that is stored in the dataset and queried. However, please note that the sorting order can be overridden by the users choice as described in the section - Drag and Drop Gene Label/Variable.

Maximum Genes

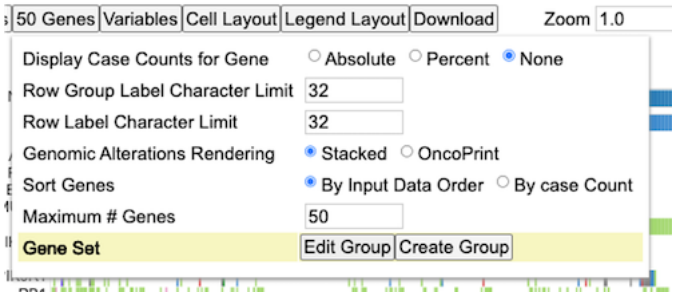
The number of genes to display on the matrix plot can be modified by the input option as shown below. Click `50 Genes` button and change input number for `Maximum Genes` to 70.



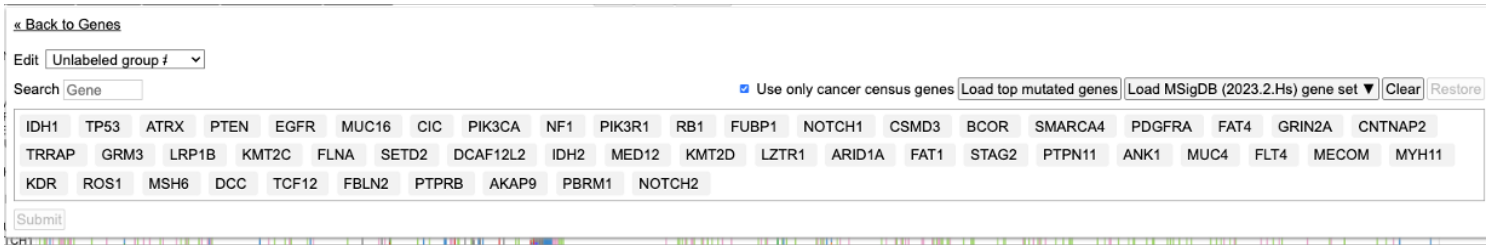
The chart updates and loads the extra 20 genes. User can modify the set of genes by using the `Gene set` option next.

Editing gene set

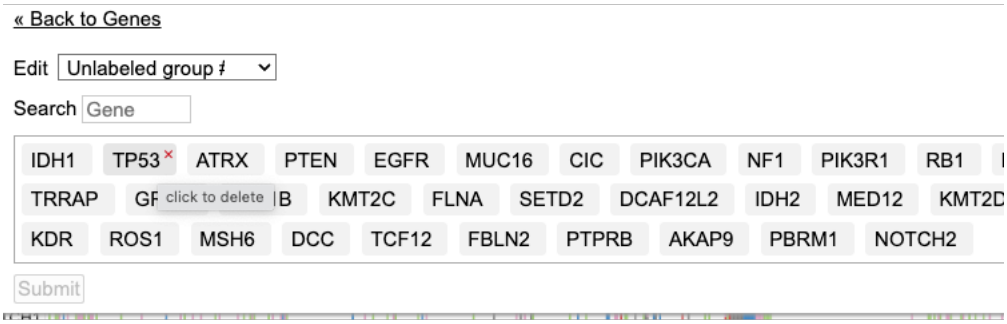
Gene groups can be edited using the `Gene set` option as shown below. Click `50 Genes` button to display this option and then click the `Edit` button in the `Gene set` as shown.



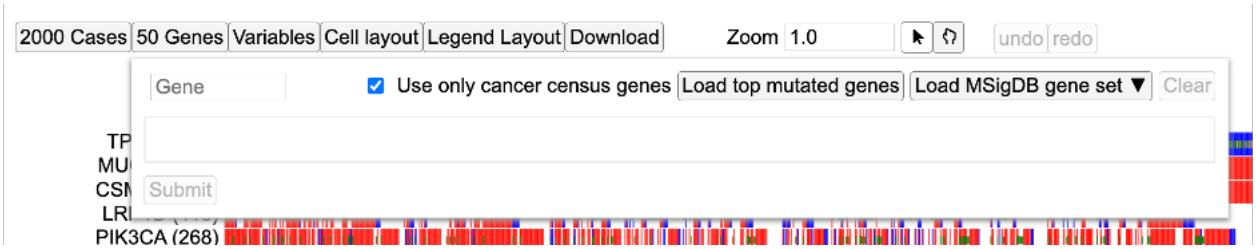
Select or deselect the blue checkbox to change the display to show `Cancer Gene Census` (CGC) genes only as shown below.



More information about CGC can be found at https://cancer.sanger.ac.uk/census. Figure displays the top 50 CGC genes. User may choose to remove single genes one at a time by clicking over the genes.To do so, hover over TP53 as shown in the image below. A red cross mark appears with a description box. Click TP53 to delete the gene as shown below.



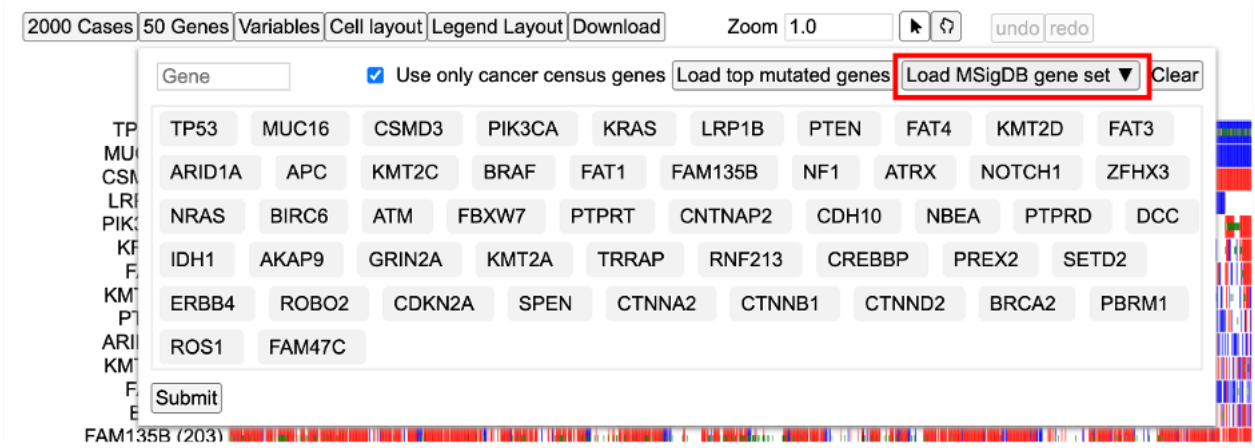
User may choose to delete all genes from view by clicking the `Clear` button as shown below. However, a gene/variable selection is mandatory for the chart to load.



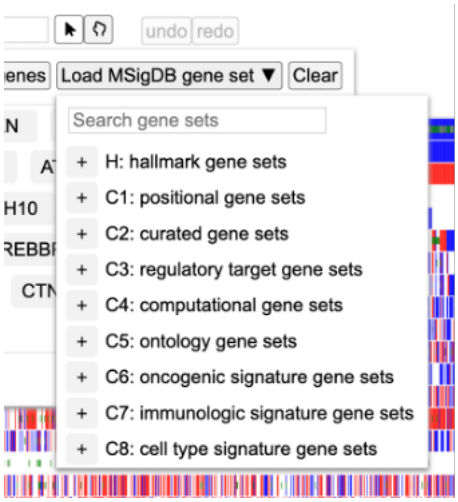
MSigDB genes

The MSigDB database (Human Molecular Signatures Database) has 33591 gene sets divided into 9 major collections and several subcollections. Users can choose to view the gene sets on the matrix plot.

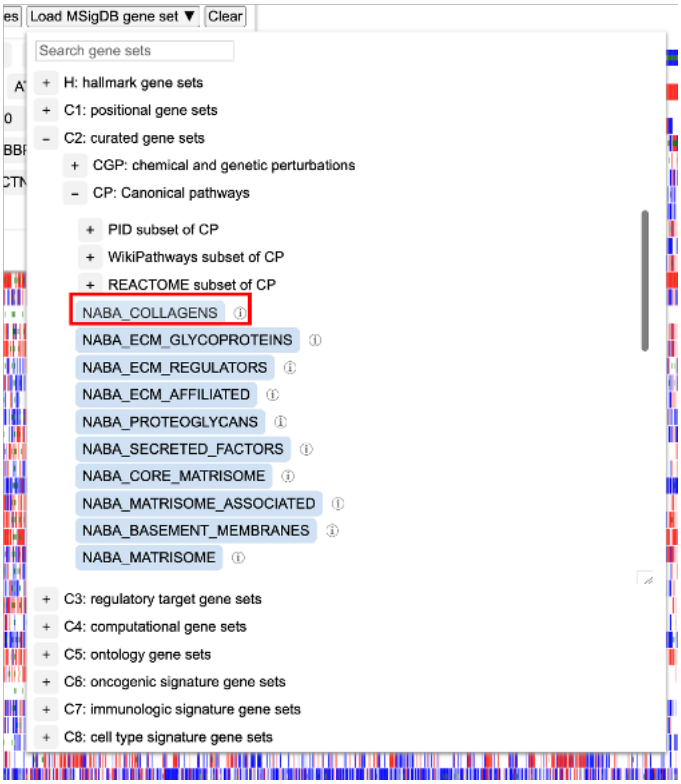
Click on the `50 Genes` button. Then click on the `Gene set - Edit`. Here user can see a button with a dropdown for loading MSigDB genes.



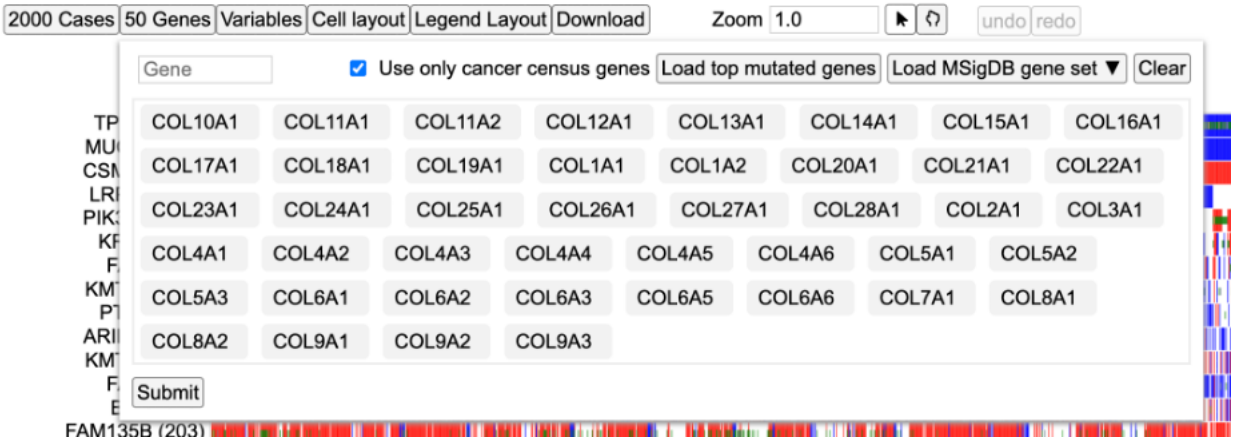
Click on this dropdown to display a tree for the different gene sets.



Select `C2: curated gene sets` and select `NABA_COLLAGENS` as shown below.



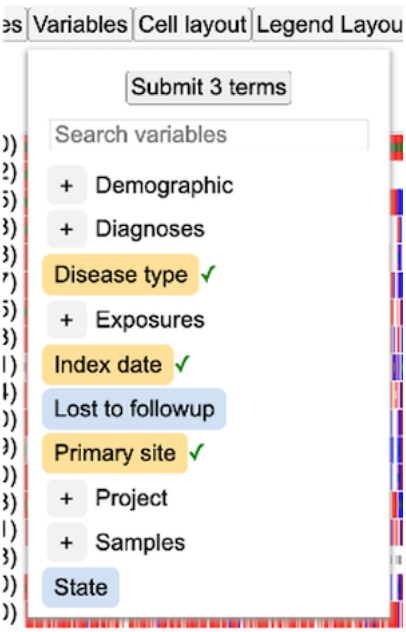
This loads the following genes as shown below.



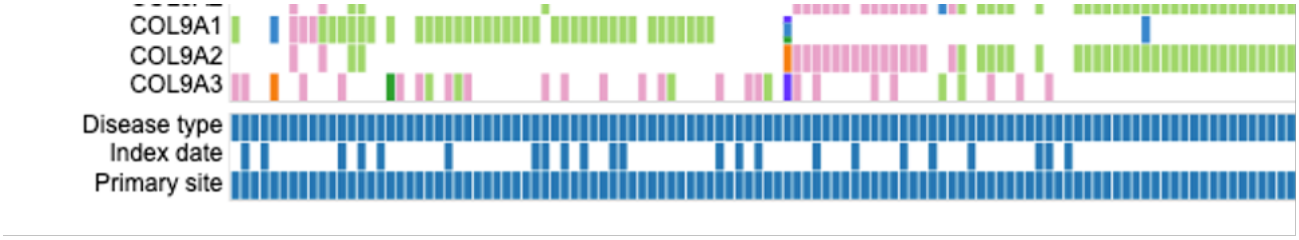
Click `Submit` and the matrix will update to reflect the selected MSigDB gene set as shown.

Variables

The third button from the left called `Variables` allows user to add in additional variables in the form of rows on the matrix. Click `Variables` to display a tree of variables and select `Disease type`, `Index date` and `Primary site`. Click the button `Submit 3 variables` as shown.

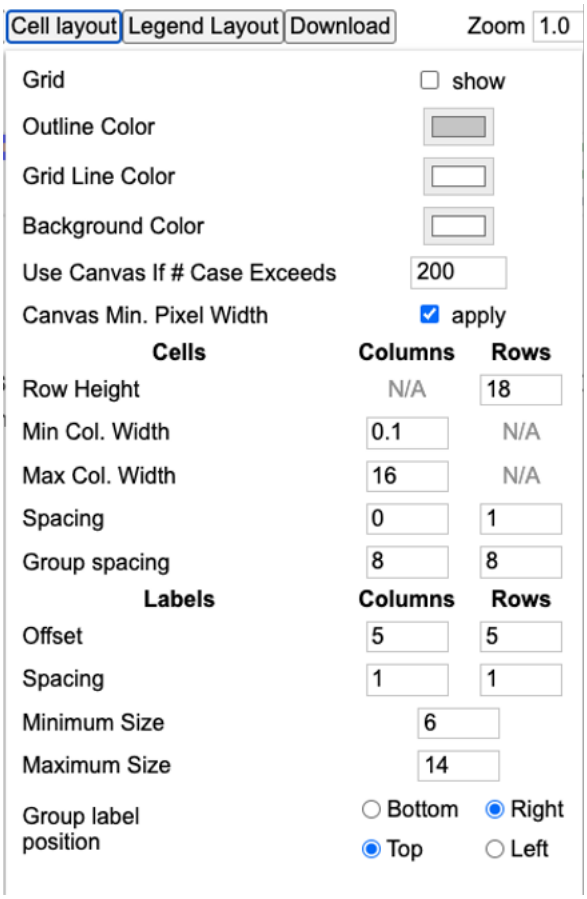


This updates the chart to display the selected variables on the very top of the matrix as shown below. User may choose to configure these rows by following steps outlined in section Clicking on gene/variable labels.



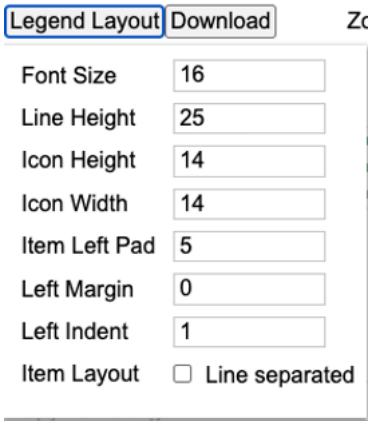
Cell Layout

The cell layout menu enables customization of the appearance such as cell dimensions, spacing, font sizes, and borders. You may mouseover an input to see the description for that input, or try checking or editing inputs to test the effects of the control input and undo/redo as needed.



Legend Layout

The legend layout menu enables customization of the appearance of the legend, such as dimensions, spacing, and font sizes. These customizations can help avoid or minimize the need for post-download edits when generating figures.

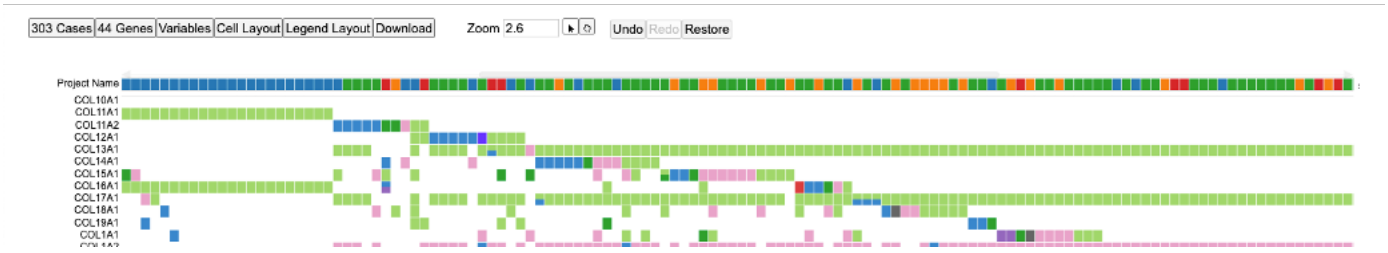


Zooming

The matrix plot offers an interactive zoom panel as shown below with which a user can zoom in to view individual samples. There are two ways to use this panel. One by changing the input number and second by sliding the zoom bar to a desired zoom level as shown.

Change zoom level to 10+ as shown.

Scroll down to view individual samples at the bottom of the plot as shown below.

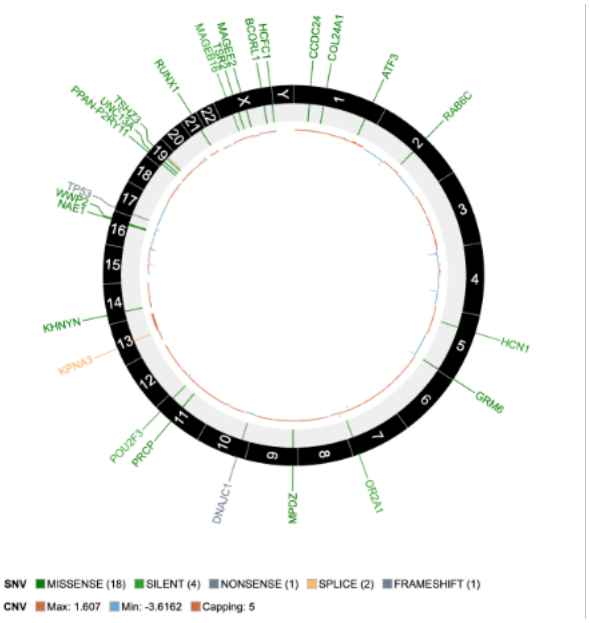


The zoom action can also be implemented by following steps as outlined in section - Drag to zoom.

Disco Plot

Click on any sample to reveal a second type of plot called as the `Disco Plot` as shown.

Click on `Disco plot` as shown above in gray. This loads a new chart above the matrix plot as shown below.



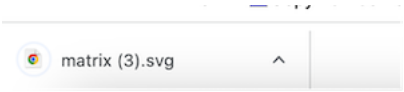
This plot shows all the mutations and CNV associated with that sample id as shown above. The plot also displays the legend for the mutation class and the CNV.

To reset the zoom level to default, click on the `Reset` button as shown. This will reset the zoom level to a default of 1.0

Download

The control panel shows an option to download the plot as an svg after user has specified their customizations. Select the `Download` button as shown below to save the svg.

The download will get saved to the default download folder as shown at the bottom of the browser window.



Legend

The legend for the matrix is below the plot and shows color coding for different mutation classes as well as color codes for CNV as shown here. This legend is interactive and user may choose to hide or show features such as mutation classes or copy number changes.

Click on the legend icons to hide anything.

Chapter 13

ProteinPaint

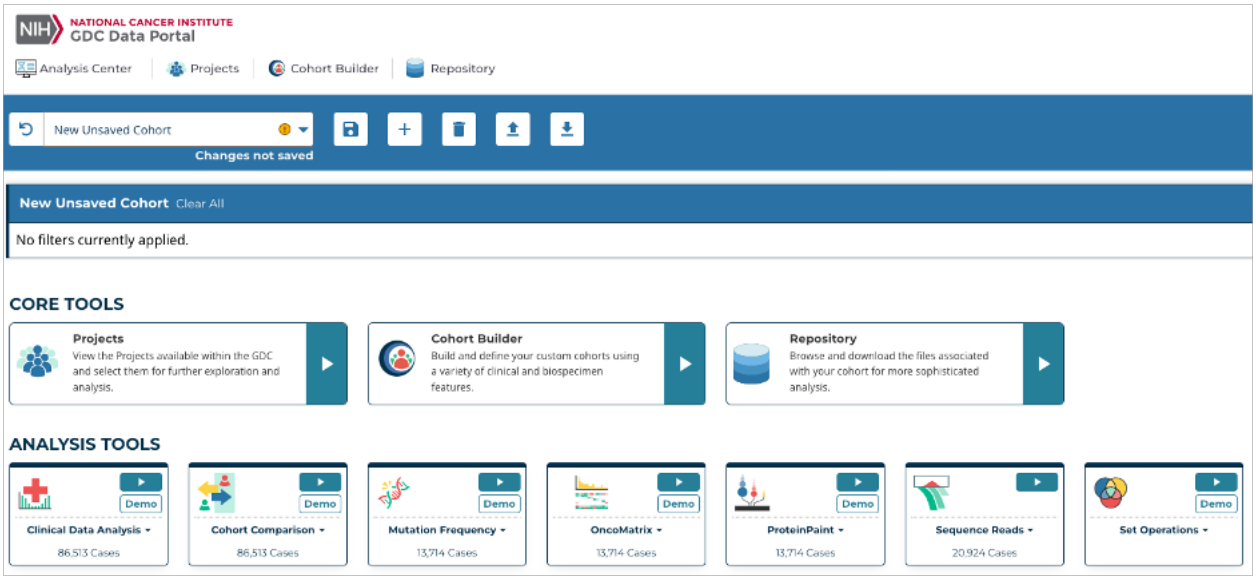
ProteinPaint Tool

Introduction to ProteinPaint

ProteinPaint is a web-based, dynamic visualization tool that displays a lollipop chart based on the multidimensional skewer version 3 (mds3 track). This tool utilizes variant annotations from GDC datasets. Given a particular gene, it displays variants associated with that gene as well as the occurrence, disease type, and demographic information of the associated case given a case.

Quick Reference Guide

At the Analysis Center, click on the 'ProteinPaint' card to launch the app.



Users can view publicly available variants as well as login with credentials in order to access controlled data.

When launched, ProteinPaint will display a search box where users can enter a gene symbol, alias, or GENCODE accession. Once a gene is entered, a lollipop frame is displayed with the name of the chart in the header.

In addition to the search box, there are two other main panels in the ProteinPaint tool: Lollipop chart, and legend.

Search box

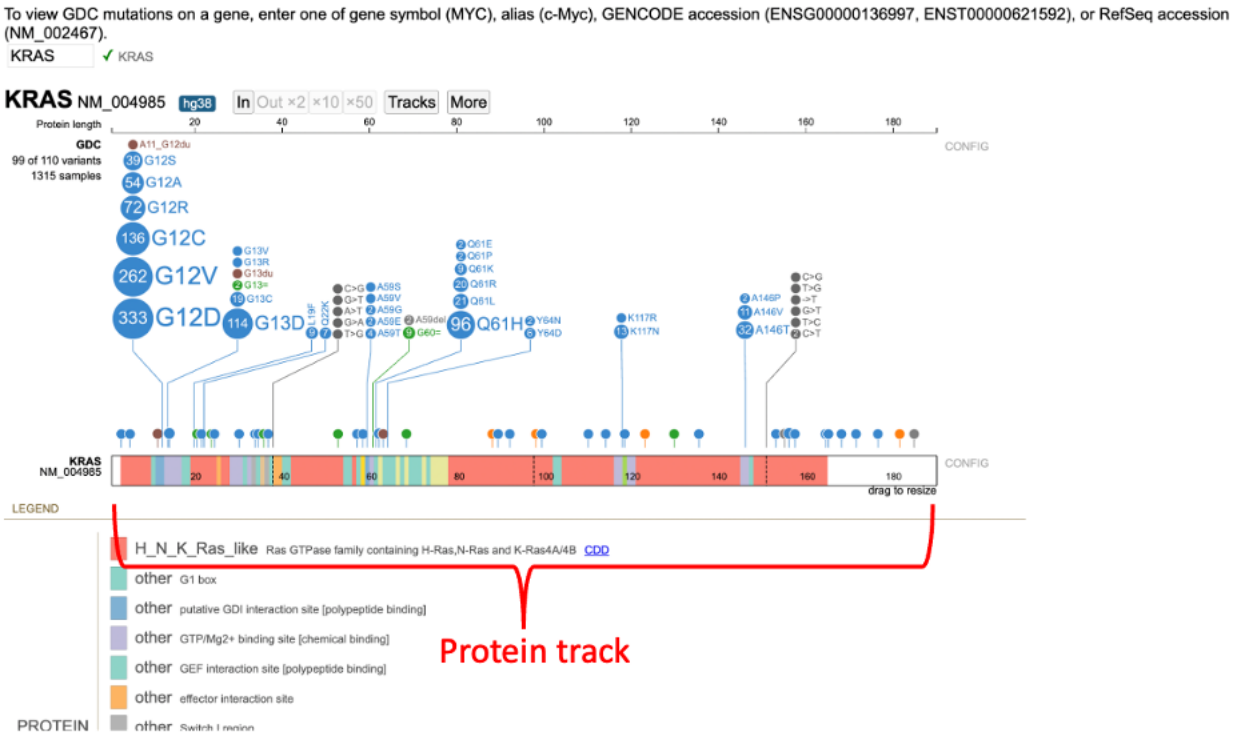
Lollipop Chart panel

Legend Panel

Lollipop Chart Panel

After entering a gene, the tool will display a Lollipop chart for the GDC variants as well as a Protein View for the default isoform.

In the Lollipop chart, the circular discs for each variant are color coded per GDC mutation classes and are proportional in size to the number of occurrences. Variants in the same position are arranged in descending order of occurrences.



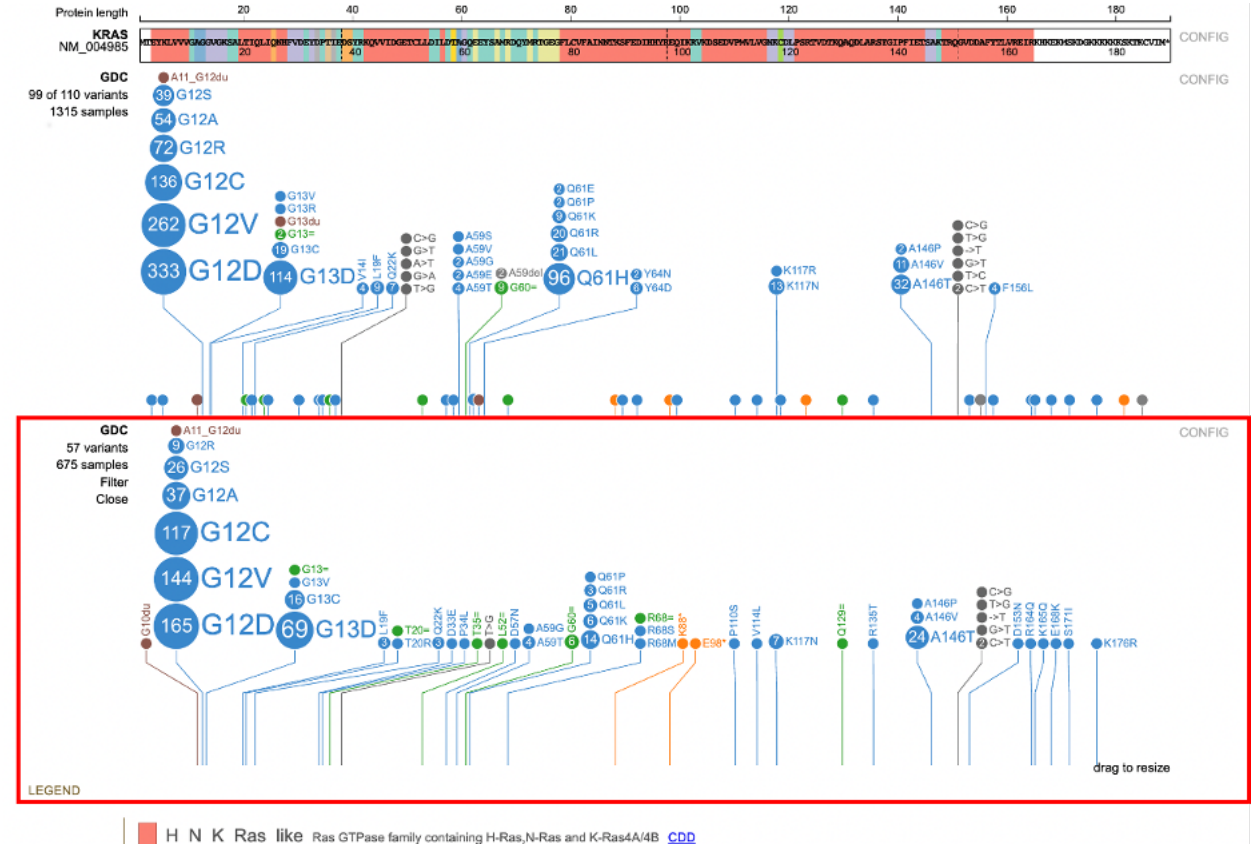
Exon variants report the amino acid change at the referenced codon. For example, G12D is a G > D substitution at the 12th codon of the protein.

The default isoform will appear directly to the right of the gene name. Clicking on the isoform number will open a display to view/select other isoforms and switch the display track.

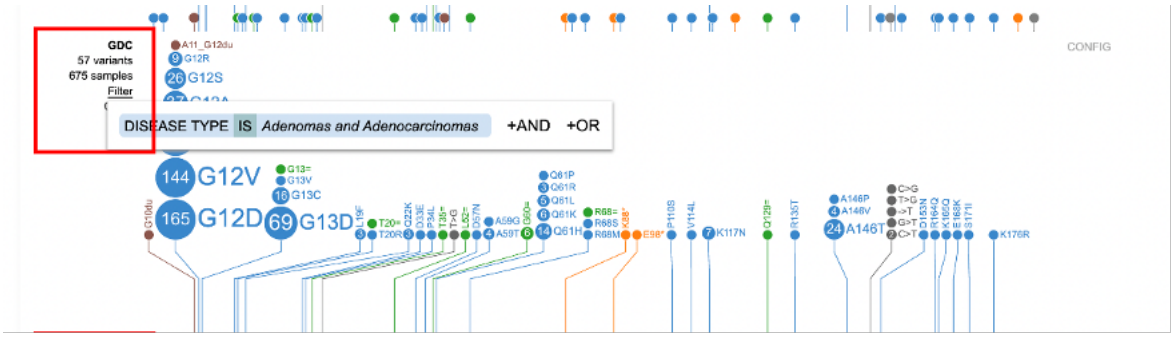
Clicking on the number of variants link, to the left of the plot, opens a menu where users can view annotations and manipulate the Lollipop:

- **List:** Displays all variants, each of which can be selected to launch the annotation table which displays consequence, mutation, sample submitter_id, and other data related to the sample
  - **Mutation:** Launches the GDC Mutation Summary Page
  - **Sample:** Launches the GDC Case Summary Page. Users also have the ability to create a new cohort or launch the Disco plot.
- **Collapse/Expand:** Collapses or expands all skewers in the lollipop
- **Download:** Downloads the mutations in a TXT file
- **As lollipops:** Displays variants via circular discs proportional to the number of occurrences
- **Occurrence as Y axis:** Sorts variants on the y-axis by number of occurrences

Clicking on the number of samples opens a window to view annotations grouped by GDC case properties such as disease type and primary site. Selecting a value adds a new Lollipop subtrack that displays only the samples with the given value. This side-by-side view allows for a comparison between the mutations in the main track versus the subtrack.

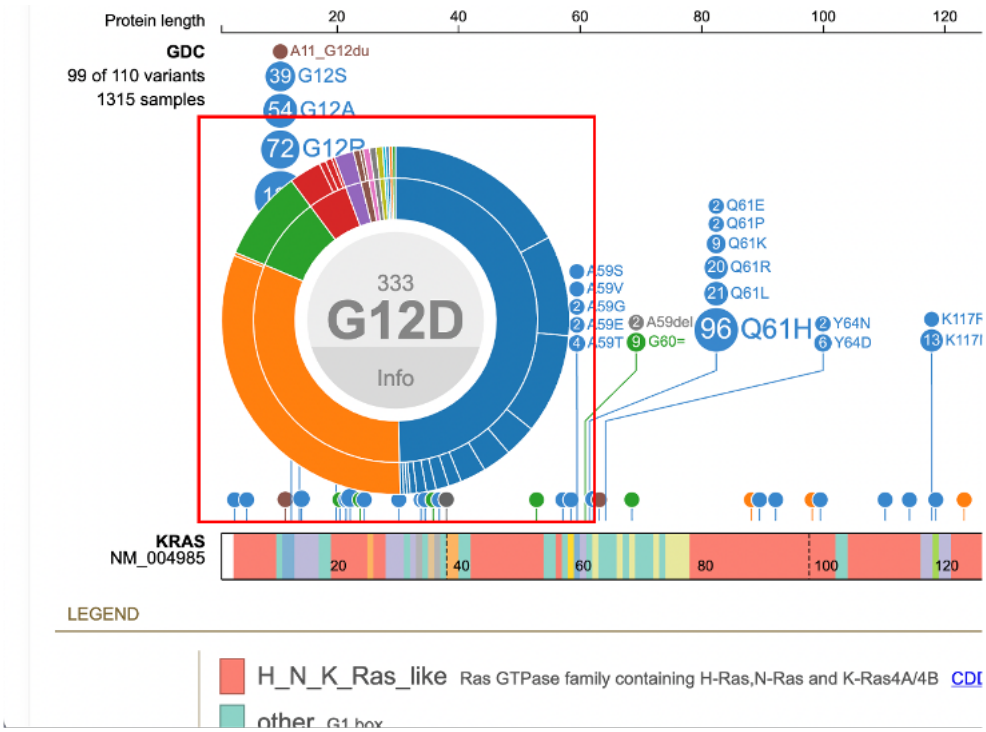


Each subtrack offers advanced filtering for users to narrow down particular features. Clicking on the value to the right of the Lollipop launches a pop-up window where users can add subsequent filters using the `+AND` or `+OR` options.

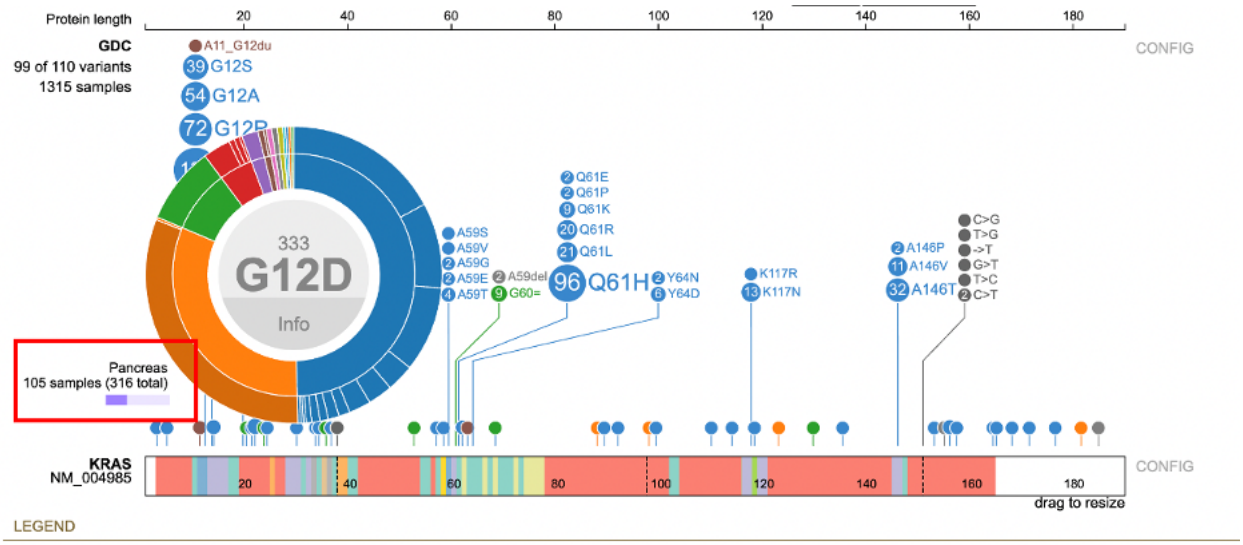


Detailed variant annotation is viewable by clicking on the disc next to the variant label. The sunburst chart is composed of a ring

hierarchy, arranged by disease types then broken down by primary sites.



Hovering over the inner and outer rings displays the disease type or primary site, number of samples, and cohort size.



An aggregate table displaying all the samples associated with that variant is available by clicking the ‘Info’ button in the center of the sunburst.

Consequence G12D MISSENSE
Mutation chr12:25245350 C>T
Occurrence 439

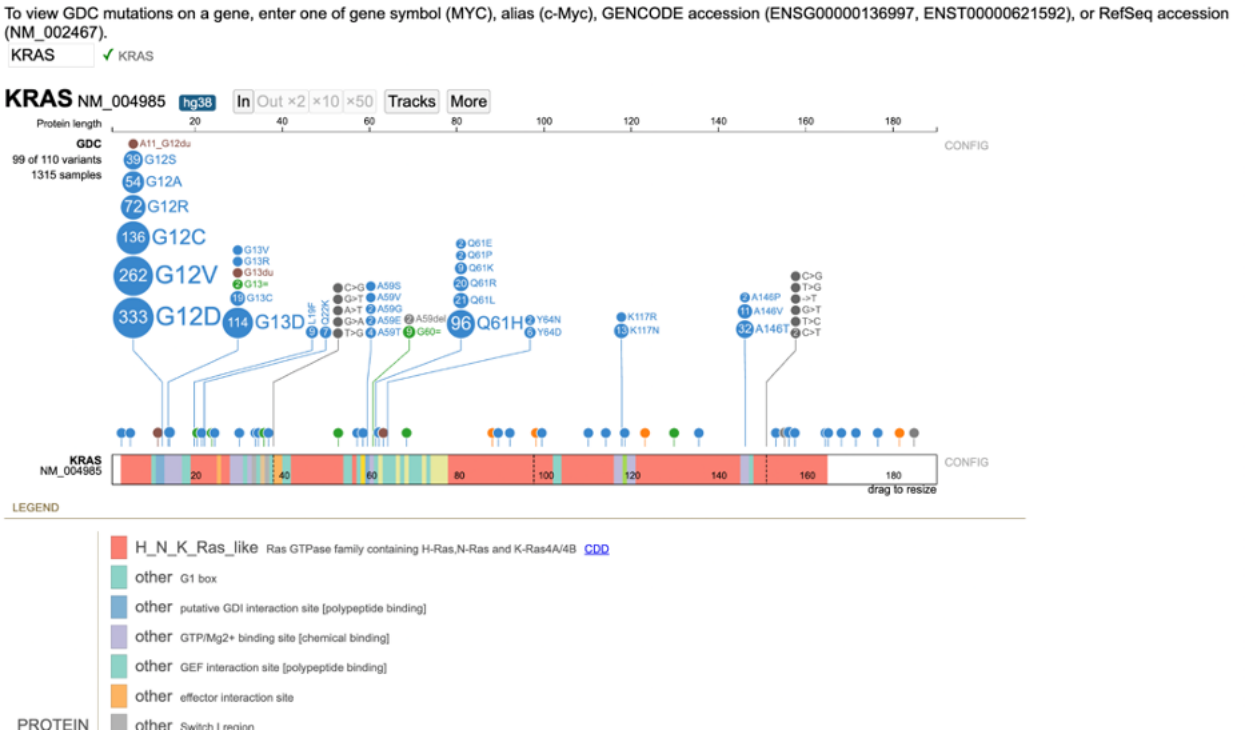
	Actions	Sample	Access	Disease type	Primary diagnosis	Primary site	Project id	Gender	Age at diagnosis	Race	Ethnicity	Mutations	Tumor MAF	Normal depth
1	Disco	TCGA-BR-4184-01A	Open	Adenomas and Adenocarcinomas	Adenocarcinoma, NOS	Stomach	TCGA-STAD	male	70 years 266 days	white	not hispanic or latino	G12D MISSENSE	6/26	30
2	Disco	C3N-01381-01	Open	Ductal and Lobular Neoplasms	Infiltrating duct carcinoma, NOS	Pancreas	CPTAC-3	female	43 years 49 days	white	not reported	G12D MISSENSE	42/343	360
3	Disco	TCGA-49-4510-01A	Open	Adenomas and Adenocarcinomas	Adenocarcinoma, NOS	Bronchus and lung	TCGA-LUAD	female	51 years 147 days	black or african american	not hispanic or latino	G12D MISSENSE	15/30	18
4	Disco	C3L-01036-01	Open	Ductal and Lobular Neoplasms	Infiltrating duct carcinoma, NOS	Pancreas	CPTAC-3	male	64 years 97 days	other	not reported	G12D MISSENSE	25/347	434
5	Disco	MP2PRT-PARFIZ-TMP1-A	Open	Lymphoid Leukemias	Acute lymphoblastic leukemia, NOS	Hematopoietic and reticuloendothelial systems	MP2PRT-ALL	female	7 years 66 days	unknown	hispanic or latino	G12D MISSENSE	94/312	317
6	Disco	MMRF_2378_1_BM_CD138pos	Open	Plasma Cell Tumors	Multiple myeloma	Hematopoietic and reticuloendothelial systems	MMRF-COMMPASS	male	66 years 351 days	white	not hispanic or latino	G12D MISSENSE	34/230	261
7	Disco	C3N-00859-05	Open	Adenomas and Adenocarcinomas	Endometrioid adenocarcinoma, NOS	Uterus, NOS	CPTAC-3	female	60 years 265 days	white	not reported	G12D MISSENSE	72/207	353

The top of the table displays consequence, mutation, and occurrence count with a link to the GDC Mutation Summary Page.

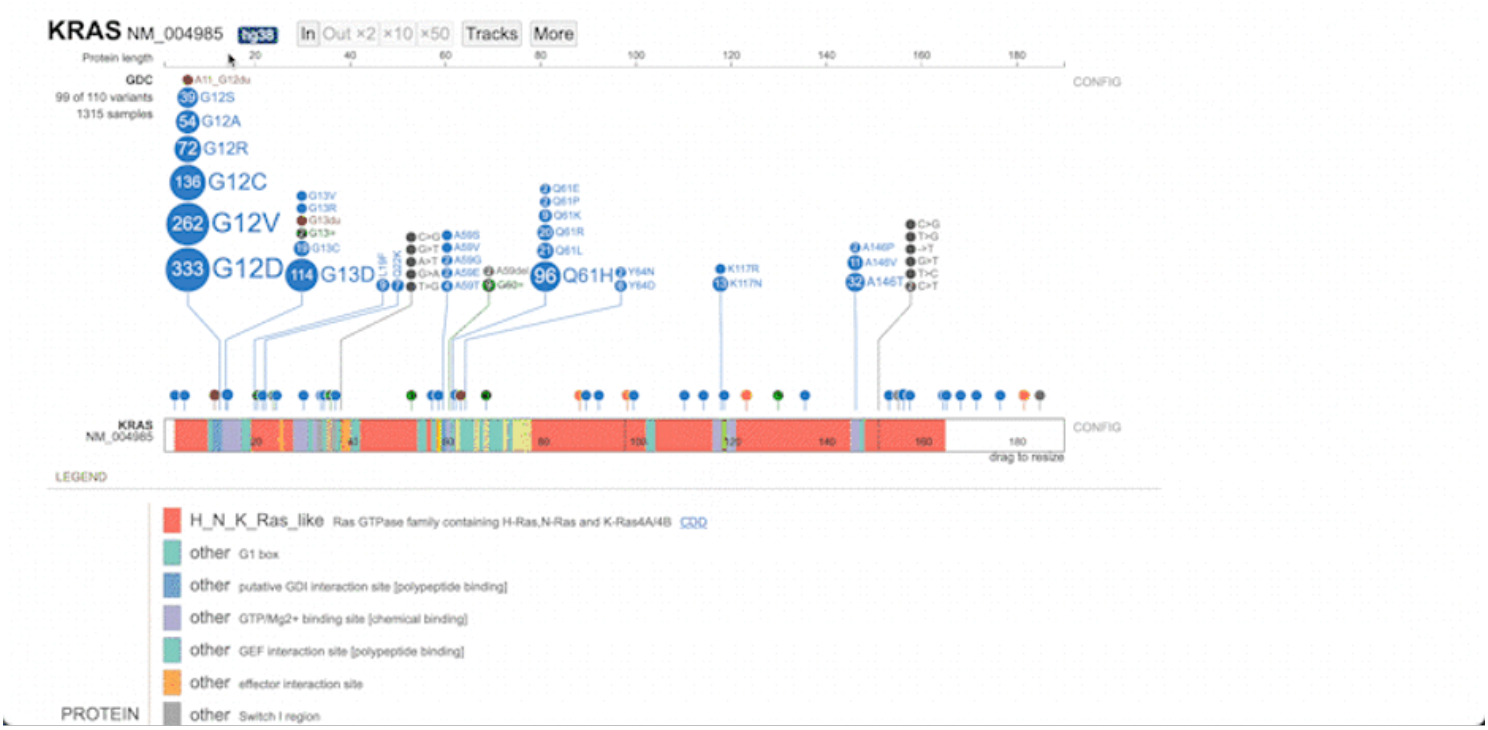
The sample table contains a number of columns for various associated features per sample such as Disease type, Mutations, and Tumor DNA Mutant Allele Frequency. Users can create a new cohort by selecting the checkboxes in the first column then clicking 'Create Cohort' in the bottom right corner of the table. The table also includes options to launch the Disco plot and the GDC Case Summary Page for each sample.

Protein View

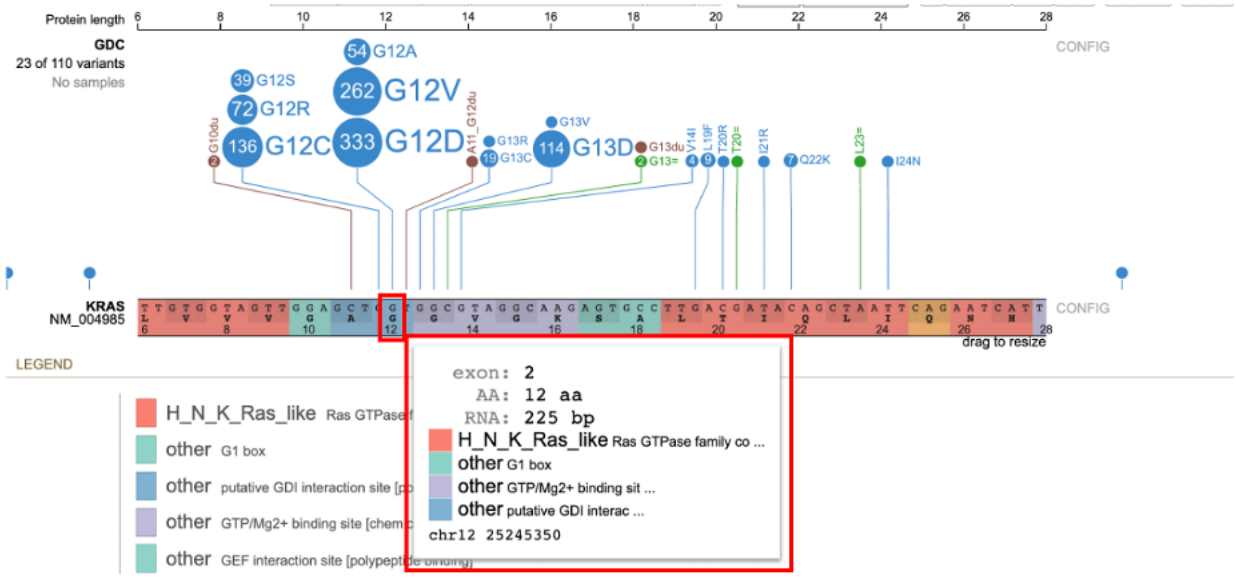
The Protein View, which displays the nucleotides, codons in the exon region, introns, and protein domains, is the primary area in which a user will visualize and interact with protein coding regions.



To zoom into the Protein View, users can highlight a region or use the zoom buttons (`In`, `Out x2`, `x10`, and `x50`) in the toolbar. For viewing a nucleotide of interest, click and drag in the top Protein length scale. The region appears highlighted in red with the calculated protein length in center.



Zooming in to the protein track displays the codons and the nucleotides. Hovering over the nucleotide position displays a tooltip with the exon, amino acids position, RNA position, and protein domain.

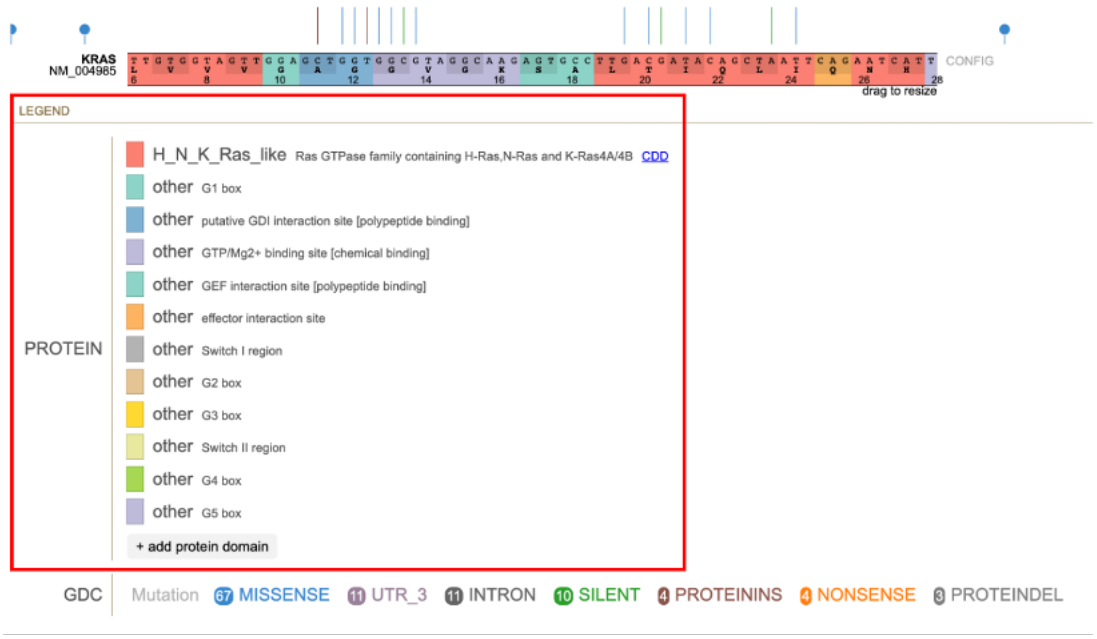


By clicking on the isoform number in the Lollipop chart, users can switch the display track between genomic, splicing RNA, exon only, protein, and aggregate of all isoforms.

Legend Panel

Protein Domains

The Protein View color codes regions by the protein domain present on the full-length protein region in the exon display.



The legend offers simple filtering for the variants shown in the lollipop. To the right of `PROTEIN`, users can click on the color to hide that particular protein domain. Clicking on the color again shows the protein domain.

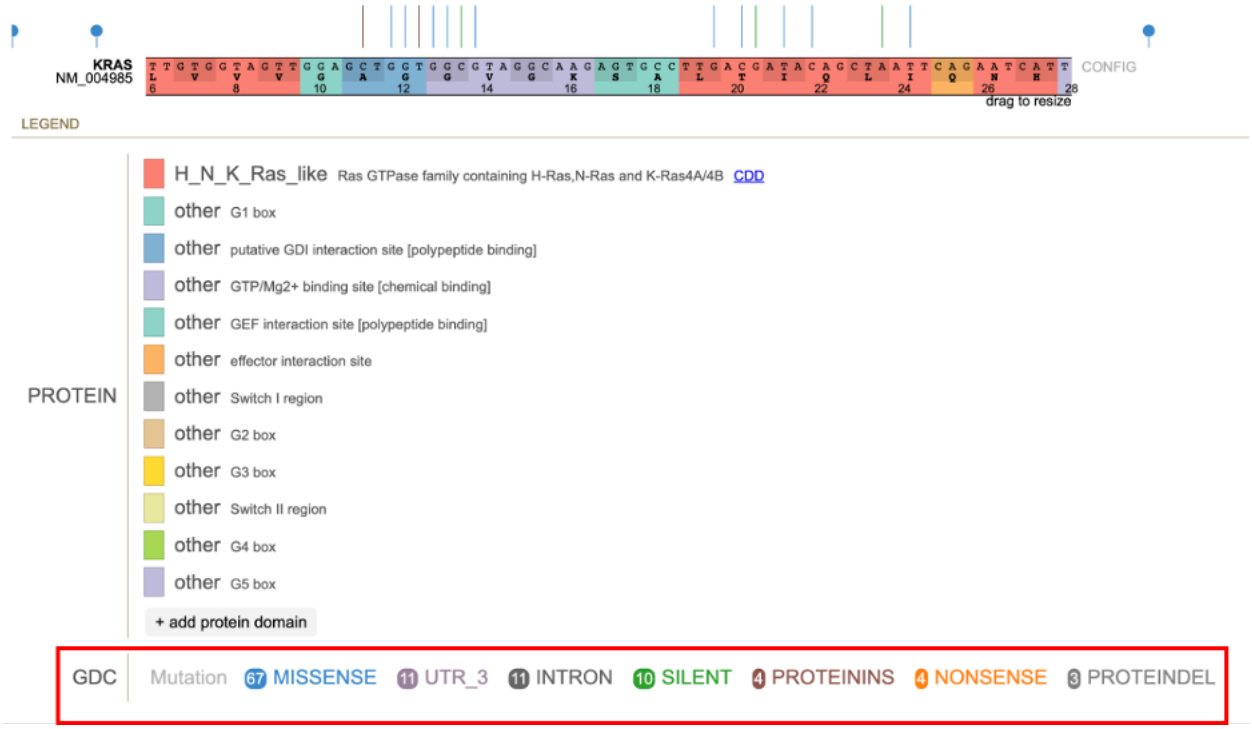
Custom protein domains are added by clicking on the `+ add protein domain` button at the bottom of the list. An input box appears requiring the following information:

1. Name, text with space, no semicolon: Name of the protein domain
2. Range, two integers joined by space: Codon position - start and stop
3. Color (e.g., red, #FF0000, rgb (255,0,0)): Color to assign to the protein domain

The protein domains also include links to databases of protein families such as the Conserved Domains Database (CDD), Simple Modular Architecture Research Tool (SMART), and Pfam.

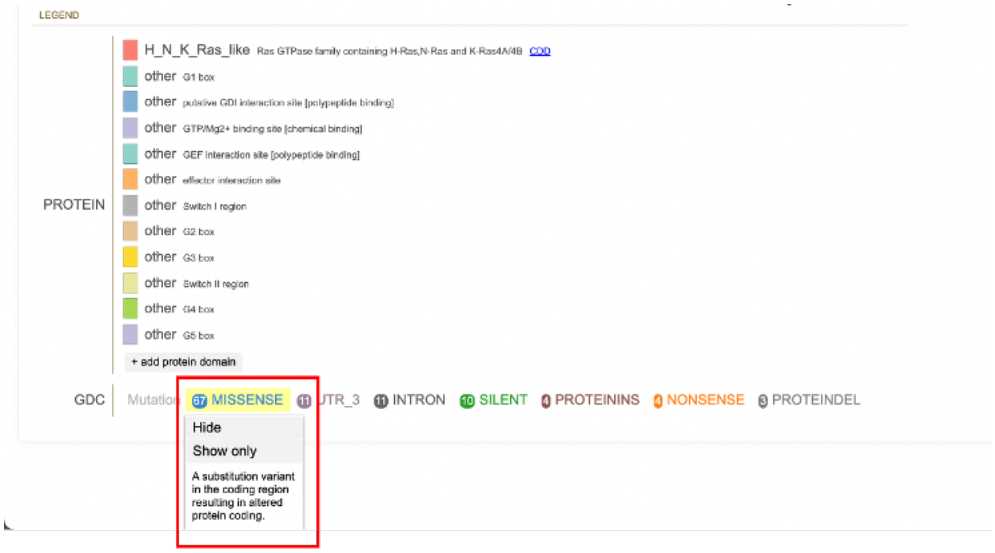
GDC Mutation Class

The GDC mutation class color coding for the lollipop discs appears below the legend for the protein domains.



Clicking on a mutation class opens a pop-up menu with show/hide functionalities:

- **Hide:** Remove all of the lollipop discs for the particular mutation class
- **Show only:** Only show the lollipop discs for the particular mutation class
- **Show all:** Display the lollipop discs for all mutation classes



Additional Features

In the toolbar, the `More` button offers methods to download figures and data:

- **Export SVG:** Download the Lollipop and legend as an SVG file
- **Reference DNA Sequence:** Display the DNA sequence as plain text for easy copying and pasting
- **Highlight:** Highlight a region in the Lollipop by selecting it in the chart or entering it in a text box

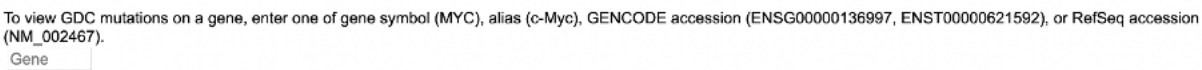
ProteinPaint Features

When selected, ProteinPaint will display the search-box as illustrated below. Once a user enters a gene symbol, alias, or GENCODE accession, a lollipop frame is displayed with the name of the chart in the header. The example below is of the gene AKT1. All gene symbols are based on the HGNC guidelines.

There are 3 main panels as outlined in the figure below:

1. Search box
2. Lollipop chart panel
3. Legend panel

Search Box



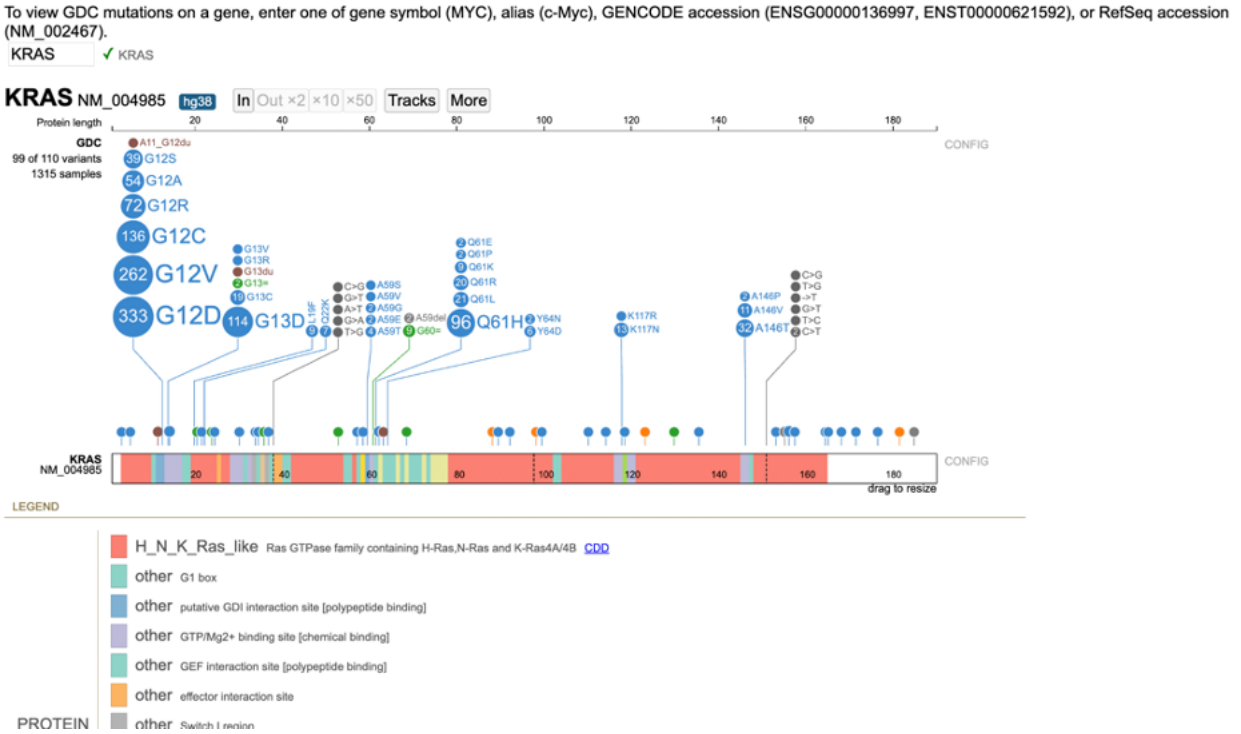
The example below uses the KRAS gene. The name of the gene (e.g., 'KRAS'), GENCODE accession no. (e.g., ENST00000311936, ENSP00000308495) or RefSeq accession (e.g., NM_004985) can be used as the search item. In case a wrong gene is entered, the

search box will display an error. For gene searches only, typing a few letters reveals a menu of possible matches. Choose from either a menu option or hit enter.

Lollipop Chart Panel

Protein View

After searching for KRAS, the Protein View for the default isoform appears in a new frame. The Protein View displays the nucleotides, codons in the exon region, introns, and protein domains as shown below.



The legend offers simple filtering for the variants showing in the lollipop. Clicking the color for a protein domain on the right of PROTEIN for example, hides that protein domain. Clicking on the color again shows the protein domain. Similar show/hide functions are available by clicking on the legend labels.

The default isoform for KRAS on hg38 genome build is NM_004985. Hovering over the isoform label will highlight it as shown below.

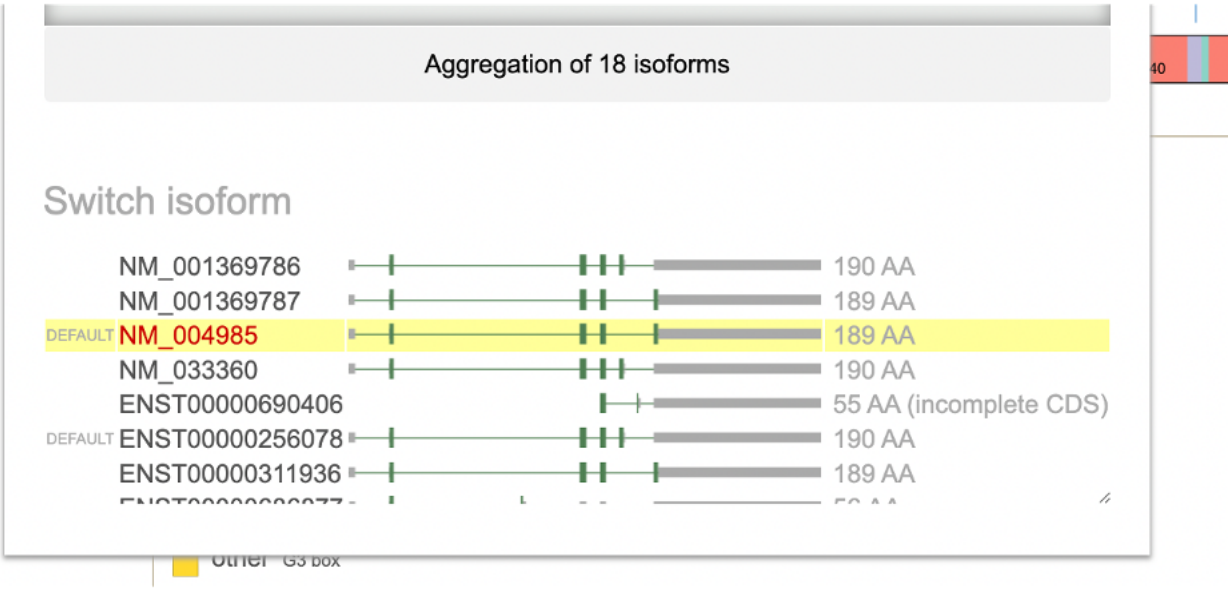
A user can select the isoform by clicking on the isoform number as shown in the figure above. Clicking this will open a display to view all the other isoforms as well as the option to switch the display track as shown below in the figure.

From **Switch Display**, a user can update to one of the following: 1. Genomic display 2. Splicing RNA 3. Exon only 4. Protein track 5. Aggregate of all isoforms

The Protein track is the primary area in which a user will visualize and interact with protein coding regions.

Protein Track

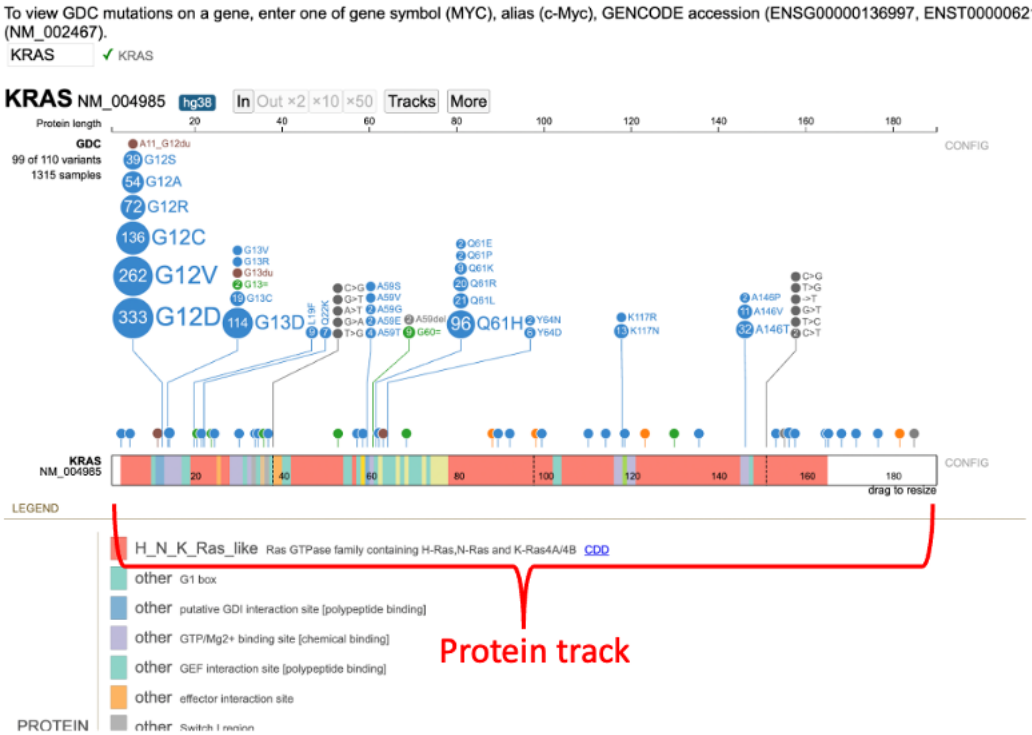
Under **Switch Isoform**, the available RefSeq and Ensembl isoform builds are listed. A condensed display and the protein length is shown for each isoform. The current selection appears in red text. The default KRAS isoform for example, is NM_004985 with 189 amino acids. To change the isoform, click on the appropriate line highlighted in yellow.



The pop-up window disappears and the lollipop track rerenders with the newly selected isoform.

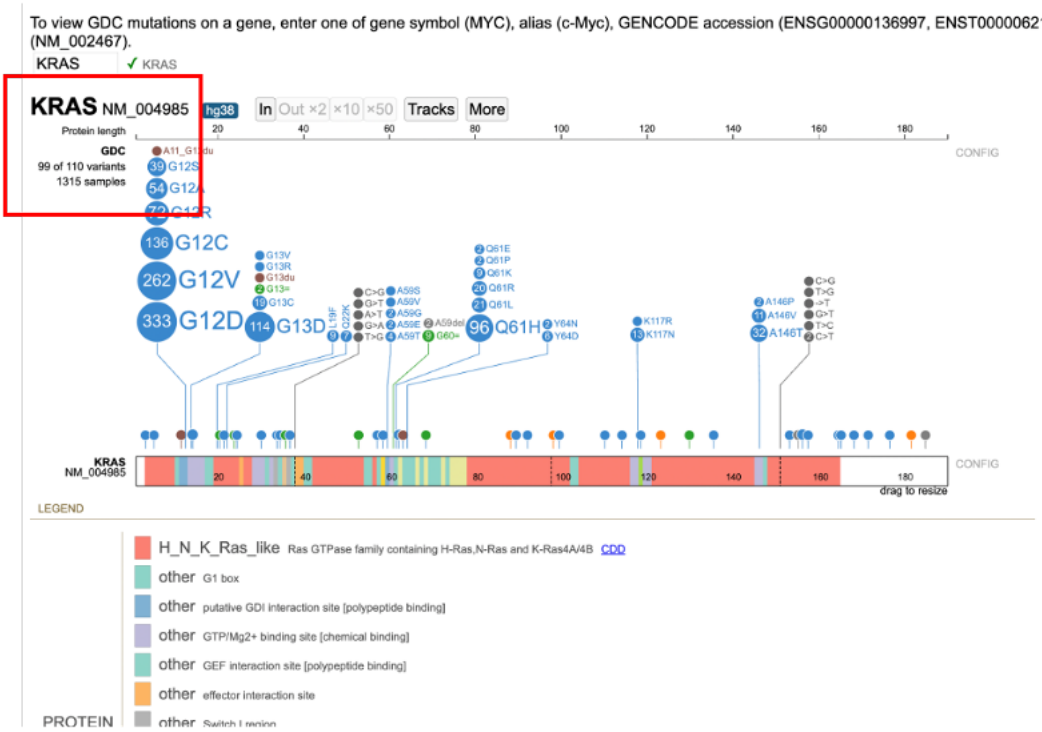
Lollipop Charts

The lollipop chart for the GDC variants appears above the Protein View. The circular disc for each variant is proportional to the number of occurrences. Variants in the same position are arranged in descending order of magnitude. There are eight types of variants found in the lollipop chart (see legend).



Exon variants report the amino acid change at the referenced codon. For example, G12D is a G > D substitution at the 12th codon of the protein.

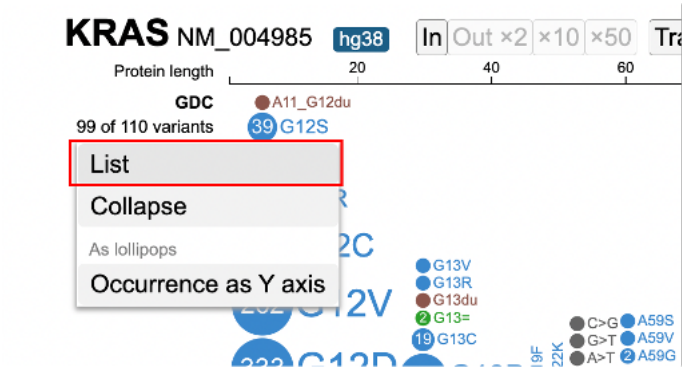
Clickable links for the number of cases (e.g. 1315 samples) and number of variants (e.g. 99 out of 110 variants) appear to the left of the lollipop. Clicking on these links reveals detailed annotations about the samples and variants, described in Viewing Variants and Case Samples.



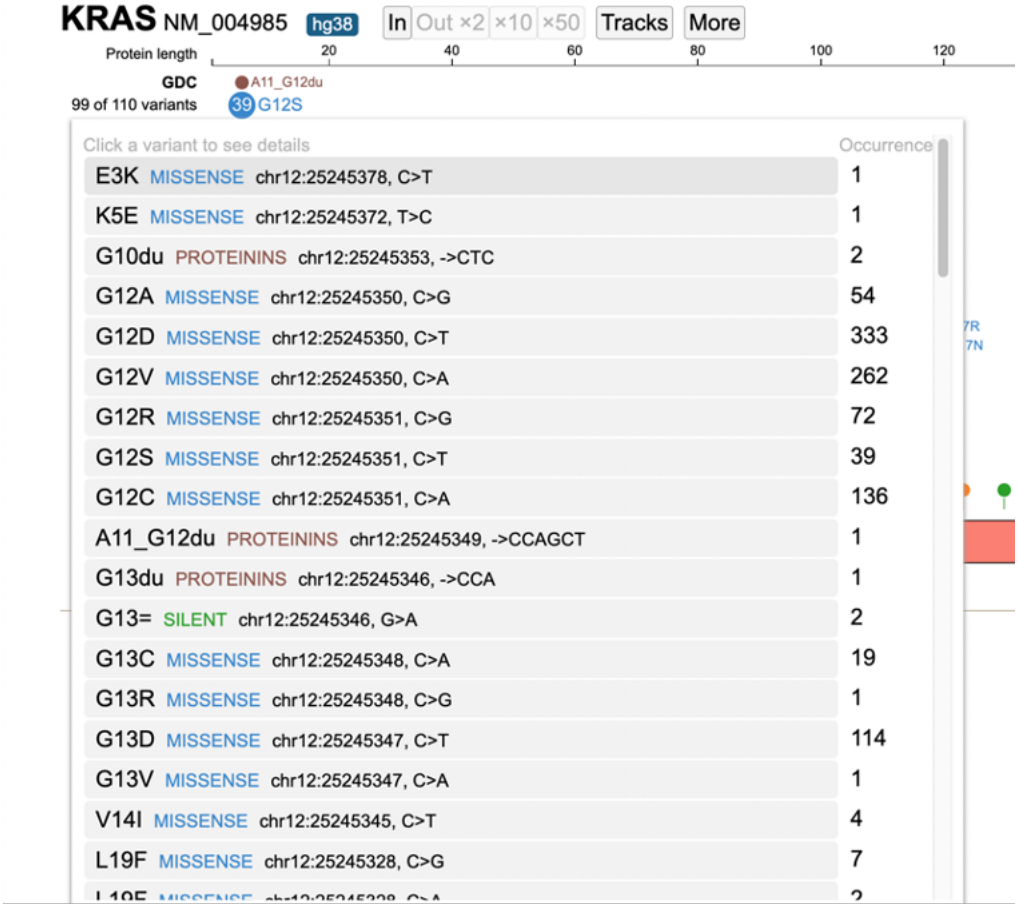
Viewing Variants and Case Samples

Variant Annotations and Chart Manipulation

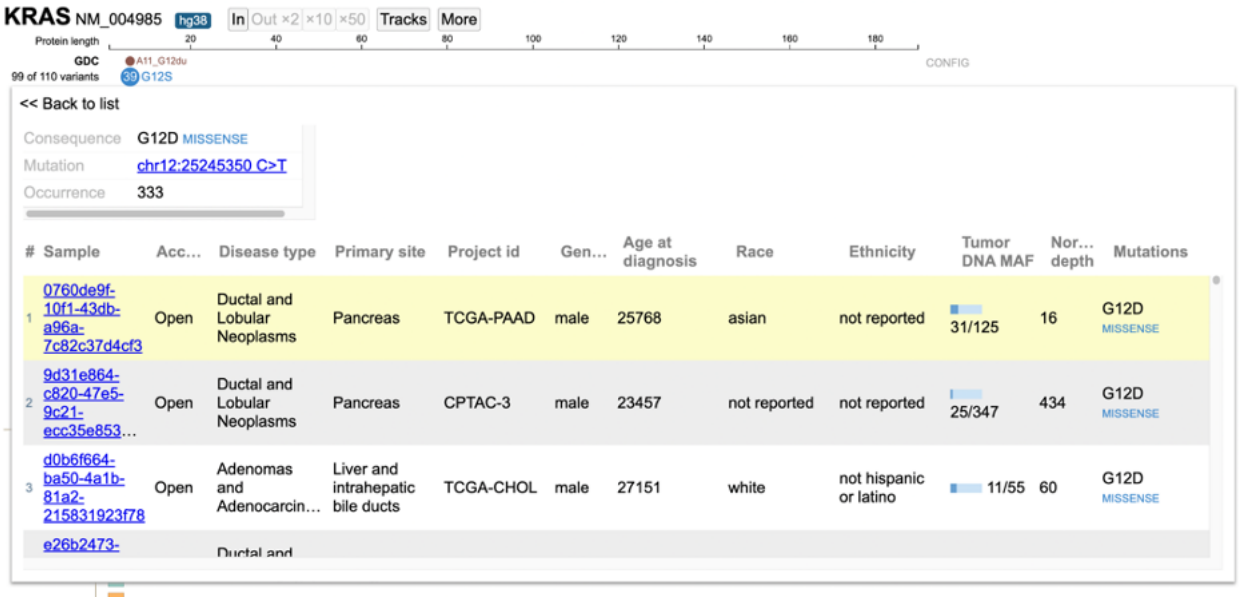
Click on the number of variants linked to the left of the lollipop for viewing annotations and manipulating the lollipop. For variant annotation, click on 'List'.



A pop-up window appears with the entire list of variants, as shown below.



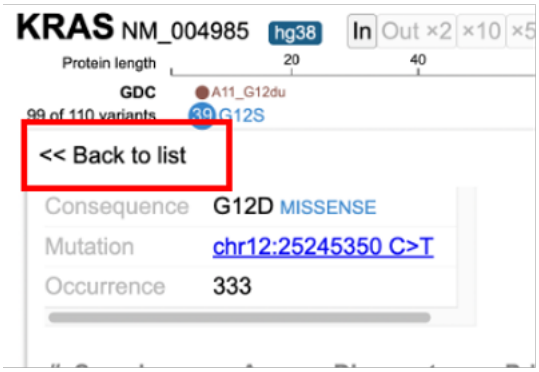
Click on the variant of interest and a new annotation table appears. From the table, view various associated features per sample such as: Disease type, Primary site, Project id, Gender, Race, Ethnicity, and Tumor DNA Mutant Allele Frequency(MAF). In the figure below, 333 occurrences are shown for the G12D variant, which represents a missense mutation at chromosome chr12:25245350 C>T.



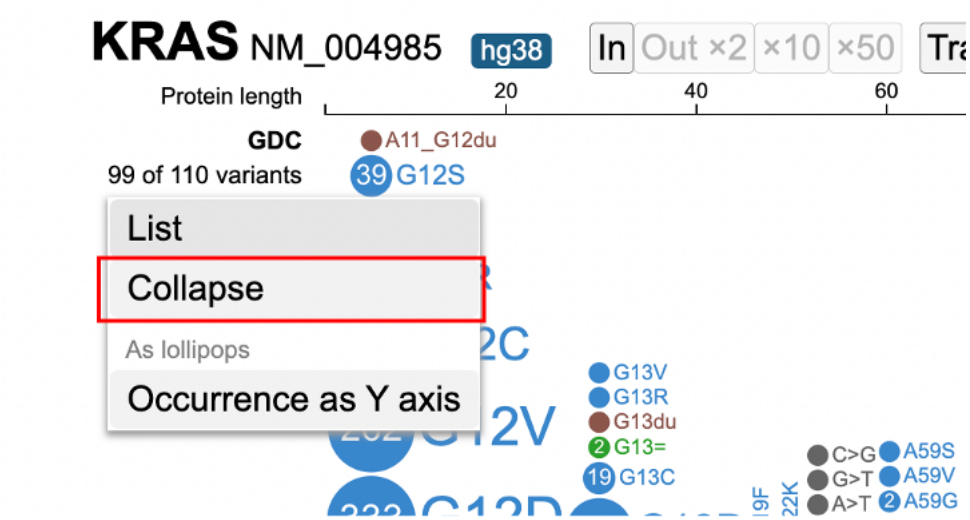
The first sample that is highlighted in yellow is a male with ductal and lobular neoplasms with a tumor DNA MAF of 31/125. This indicates 31 mutant alleles were found out of 125 total alleles.

The GDC dataset includes an 'Access' column to indicate whether the data is controlled or open. Users must obtain permission from dbGaP to view controlled data See Obtaining Access to Controlled Data. Click on the sample hyperlink and the GDC's case summary for the sample will appear in a new tab.

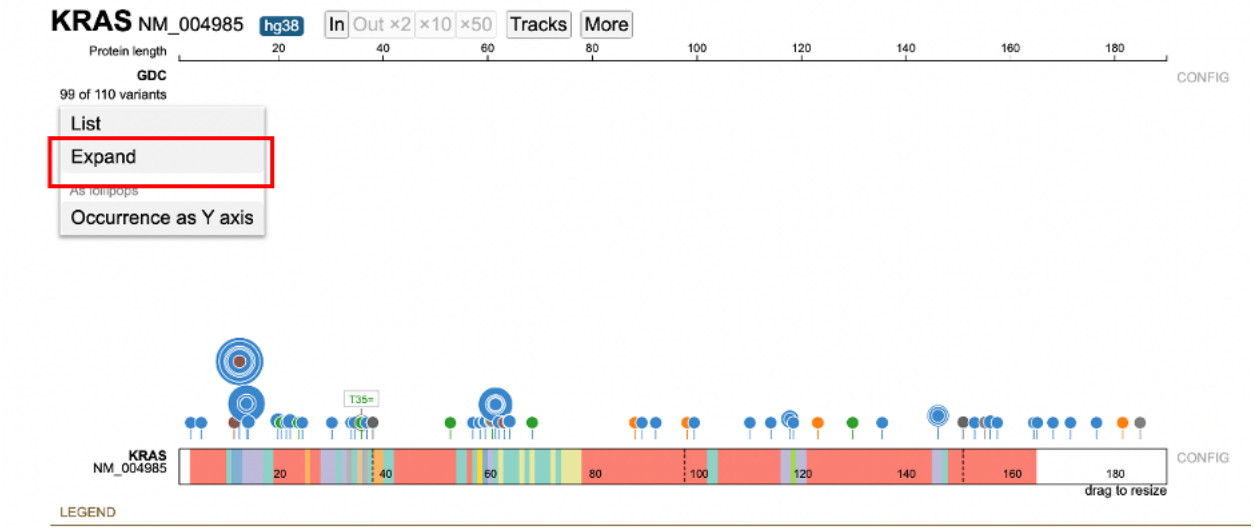
Click ‘Back to list’ and select another sample, as shown below.



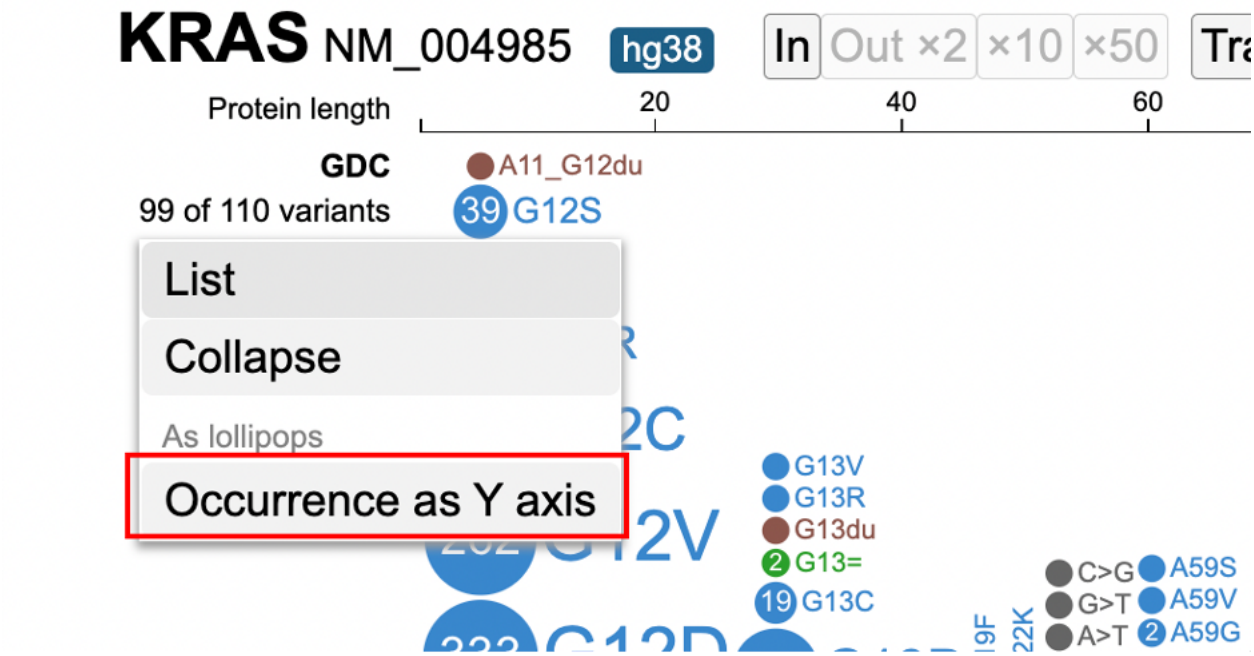
After clicking on the variant menu again, select the ‘Collapse’ option to collapse all skewers in the lollipop.



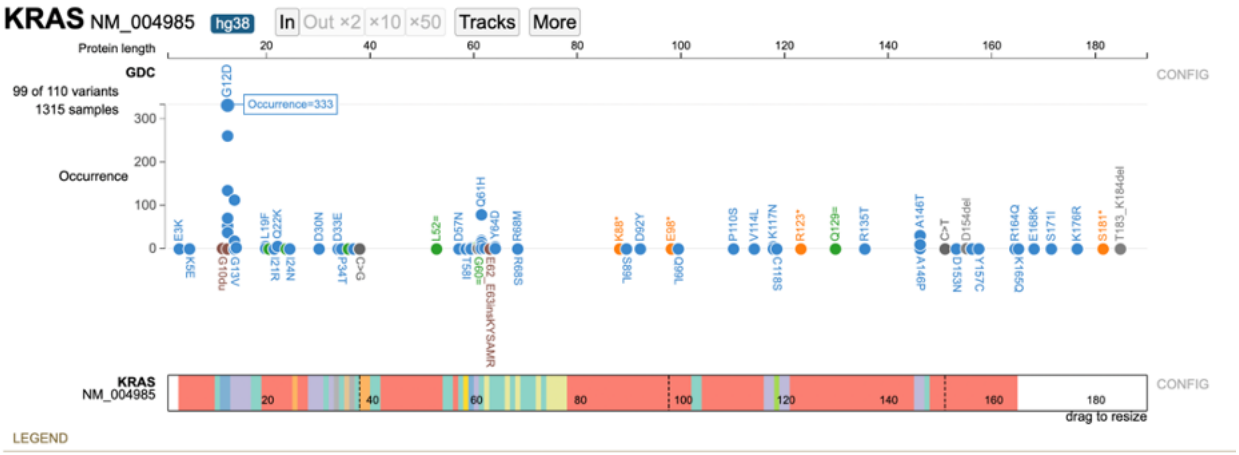
To expand any previously collapsed skewers, open the variant menu, and click on ‘Expand’.



The lollipop chart includes an option to arrange variants by the range of occurrences. Open the variant menu and click on ‘Occurrence as Y axis’.



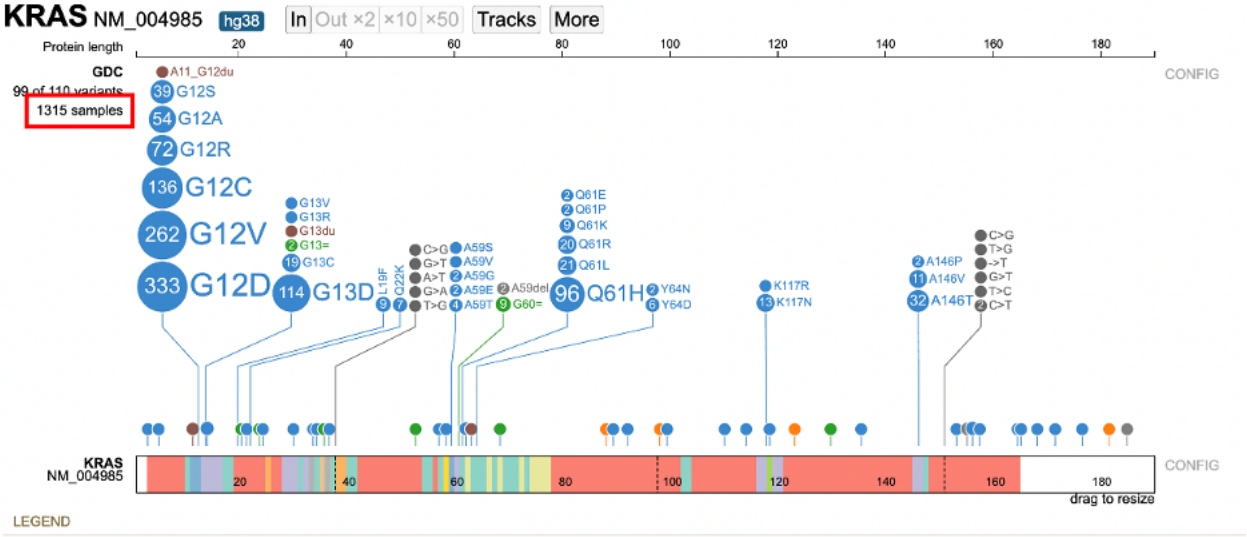
The lollipop re-renders with the variants sorted on the y-axis from lowest and highest occurrence. Hover over a variant to display the number of occurrences. In the example below, a user is hovering over G12D to display 333 occurrences of this variant.



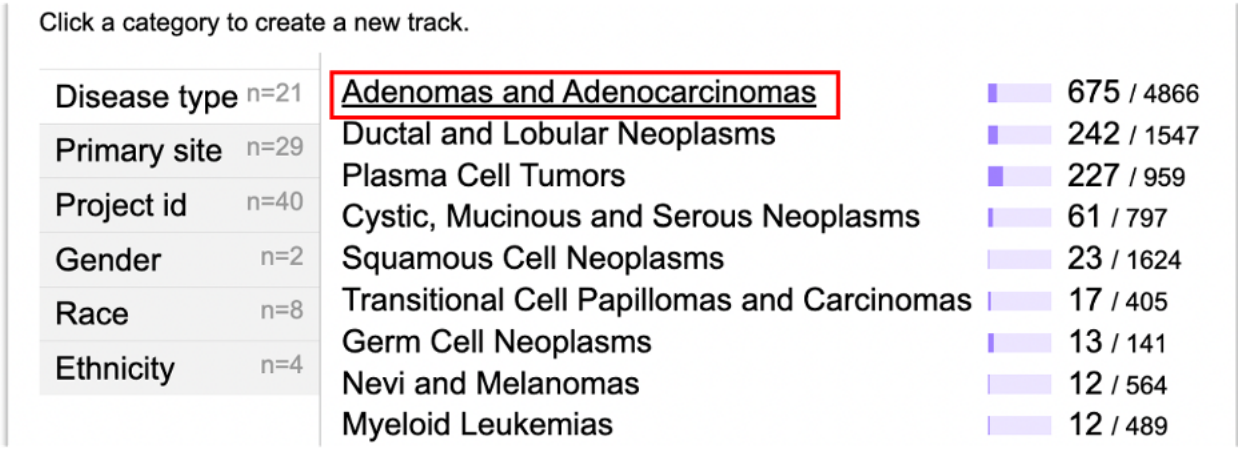
Clicking on the variant loads the sample table again as shown below.

Case Filtering

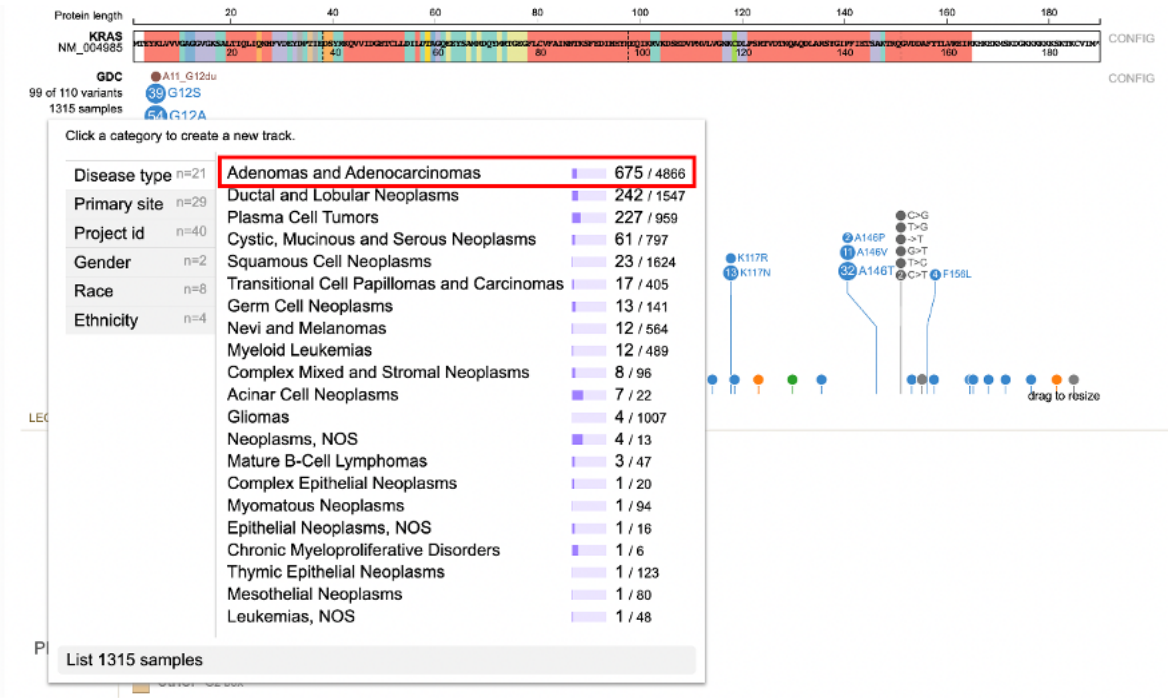
Clicking on the sample hyperlink on the left of the lollipop (e.g. 1315 samples) opens a menu to list all samples. Aggregate data for all samples by attribute appears in a series of tabs. The ability for advanced filtering and creating subtracks is available from this new display.



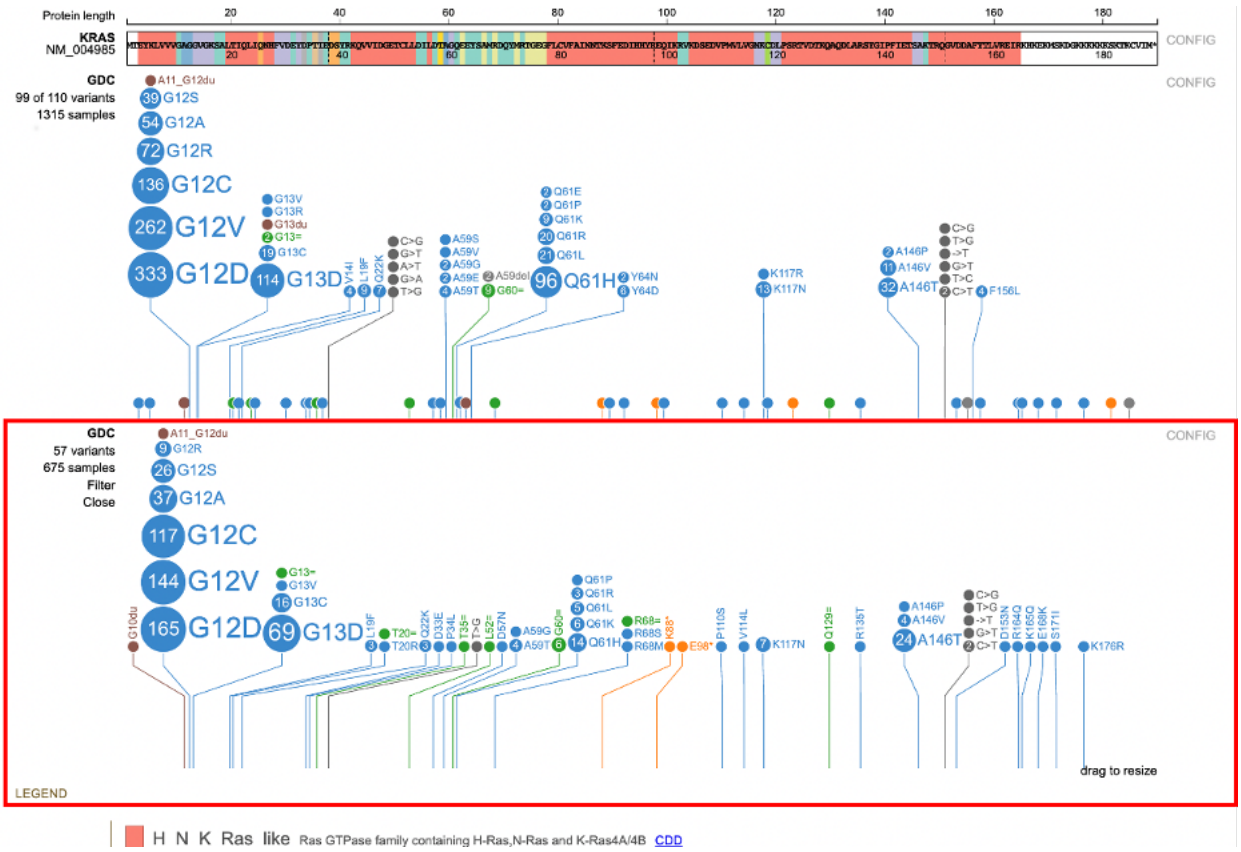
Click on 1315 samples to view annotations grouped by attributes such as: Disease type, Primary site, Project id, Gender, Race, Ethnicity, etc.. For each attribute, the number of values is represented by 'n' to the right of the group label. In the figure below, 21 values for Disease type are reported.



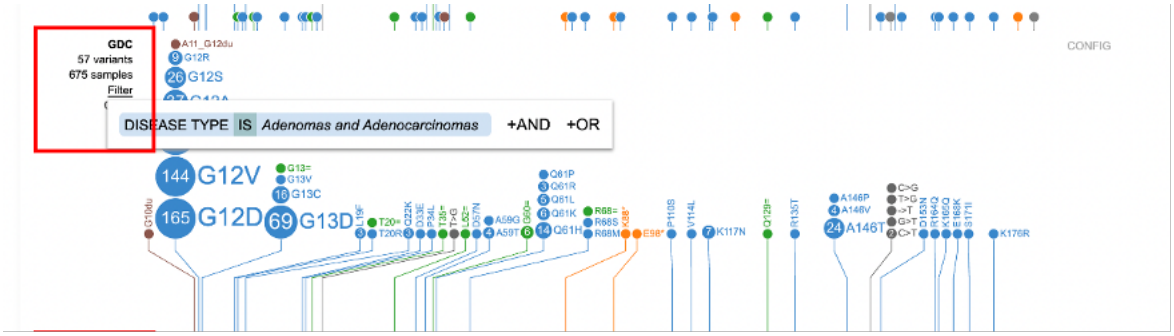
To start filtering, click on the value label or the value's sample fraction. Clicking on 'Adenomas and Adenocarcinomas' or '675/ 4866' for example, loads a new lollipop subtrack underneath the main GDC lollipop track.



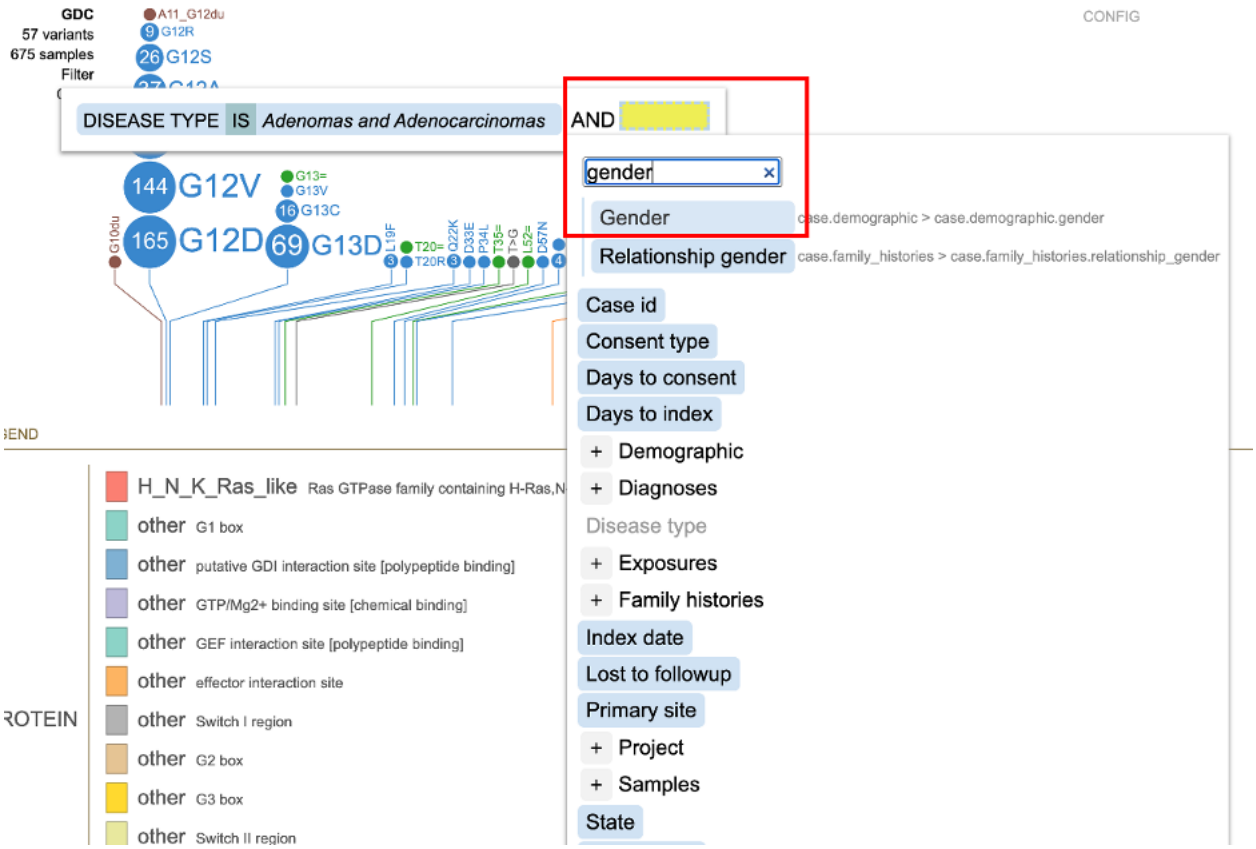
This new subtrack only shows the 675 Adenomas and Adenocarcinomas samples. This side-by-side view allows for a comparison between the mutations in the main track vs the subtrack.



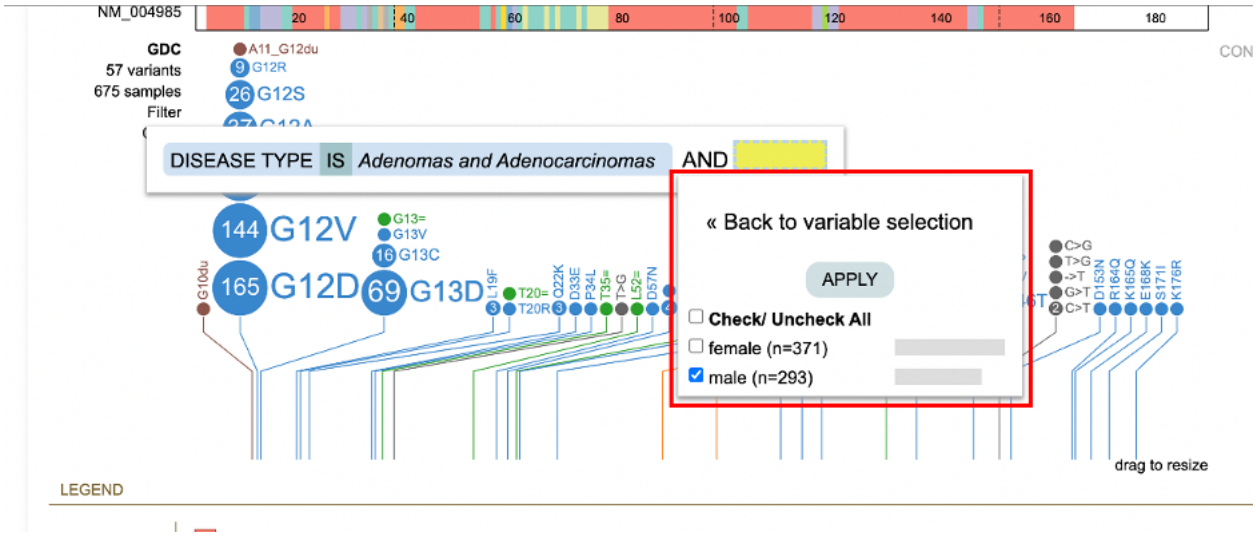
Each subtrack offers advanced filtering, shown below, for users to narrow down particular features.



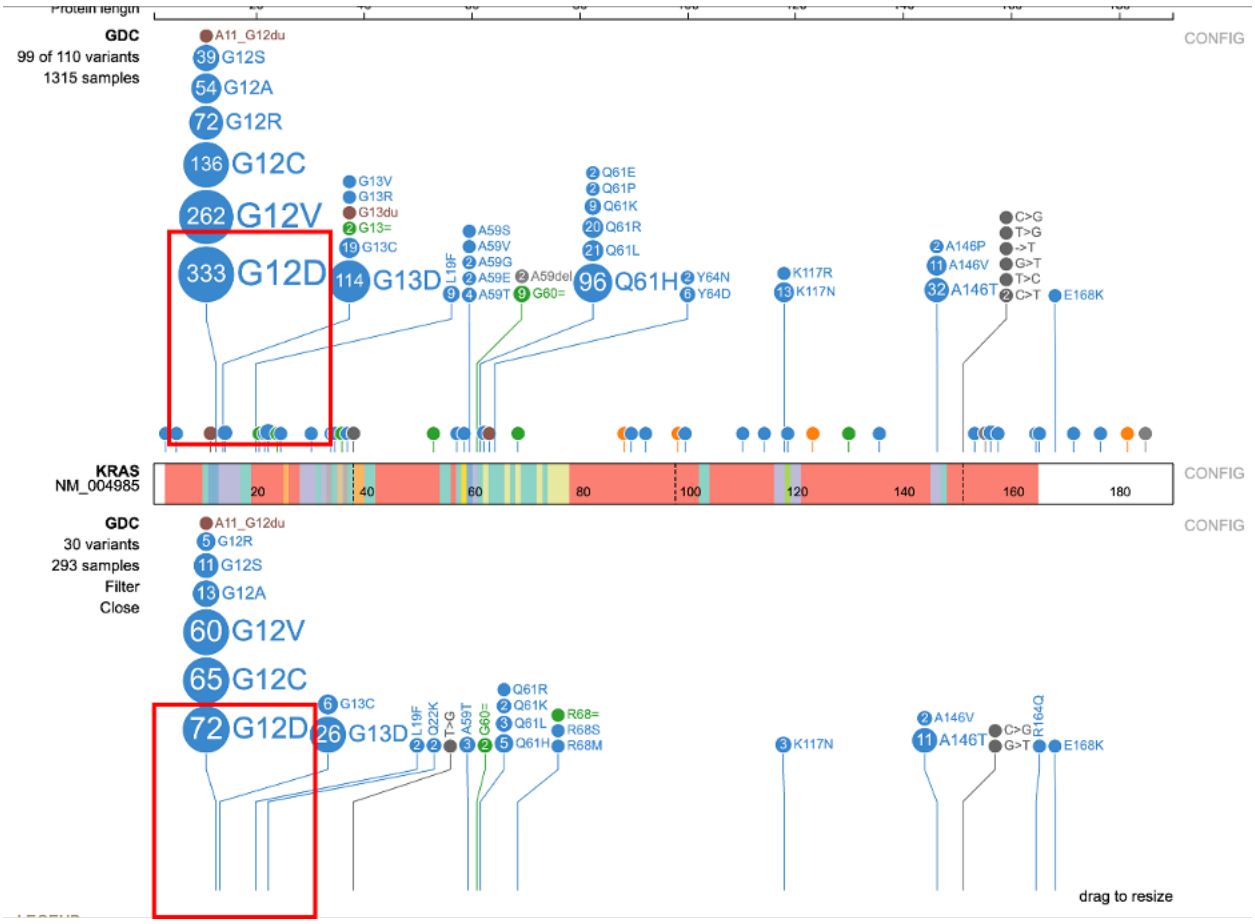
Clicking on 'Filter' displays a pop-up window with the feature the user selected previously from the sample annotation menu (e.g. Disease type: Adenomas and Adenocarcinomas). Clicking on either +AND or +OR displays a new pop-up with a search bar. Search for the desired term and click on the term's button. In the image below a user selected 'gender' by clicking the '+AND'.



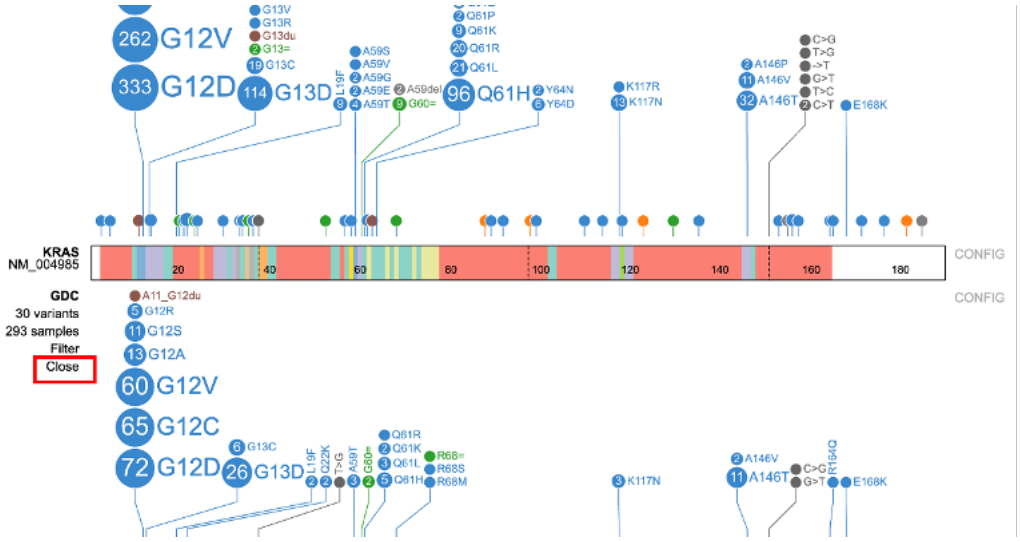
By clicking on 'Gender', all available values appear with checkboxes (i.e. male and female) as shown below. In this example, male with 293 data points is selected.



Click 'Apply' and the subtrack re-renders to reflect the updated filter. In the example below, the subtrack reduces from 675 samples to the 293 male samples with adenomas and adenocarcinomas. The figure shows the difference in mutations in the two tracks. Out of the original 333 samples, 72 of 293 males report the G12D mutation.

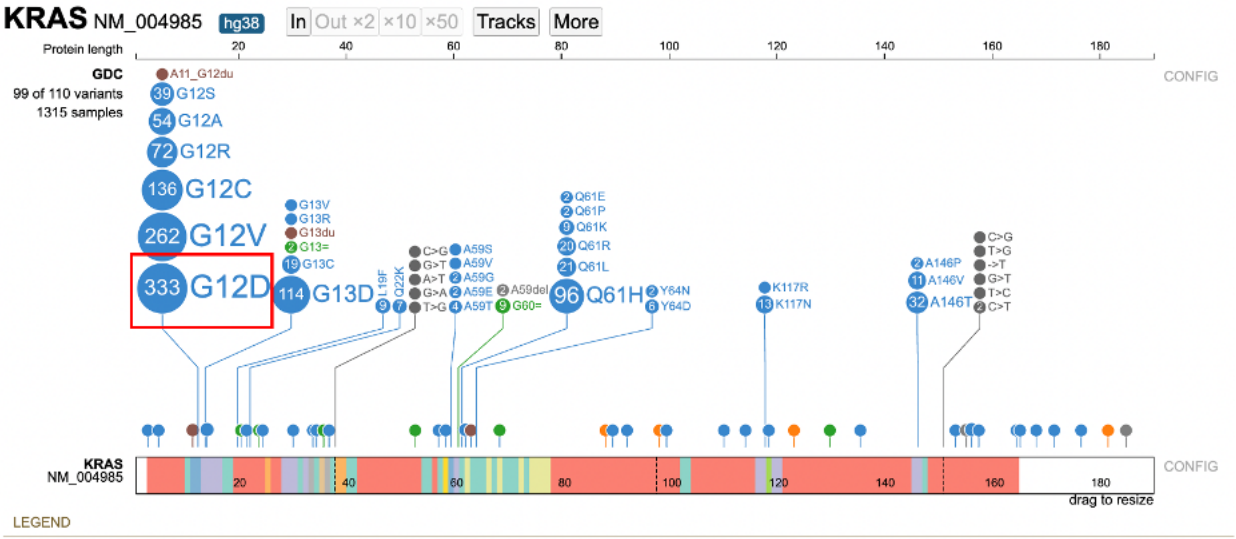


Click on the ‘Close’ option to remove the subtrack from the page.



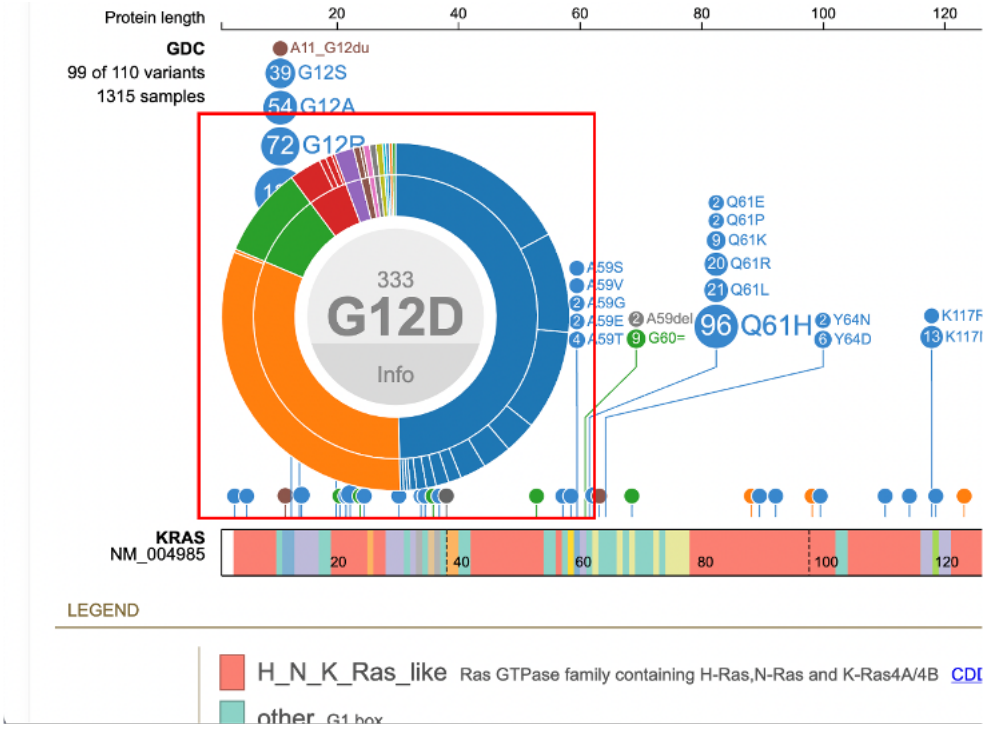
Viewing in the Lollipop Display

In the lollipop chart, users can drag the protein track down by clicking the name of the gene on the left of the protein track and pulling it below the lollipop chart.



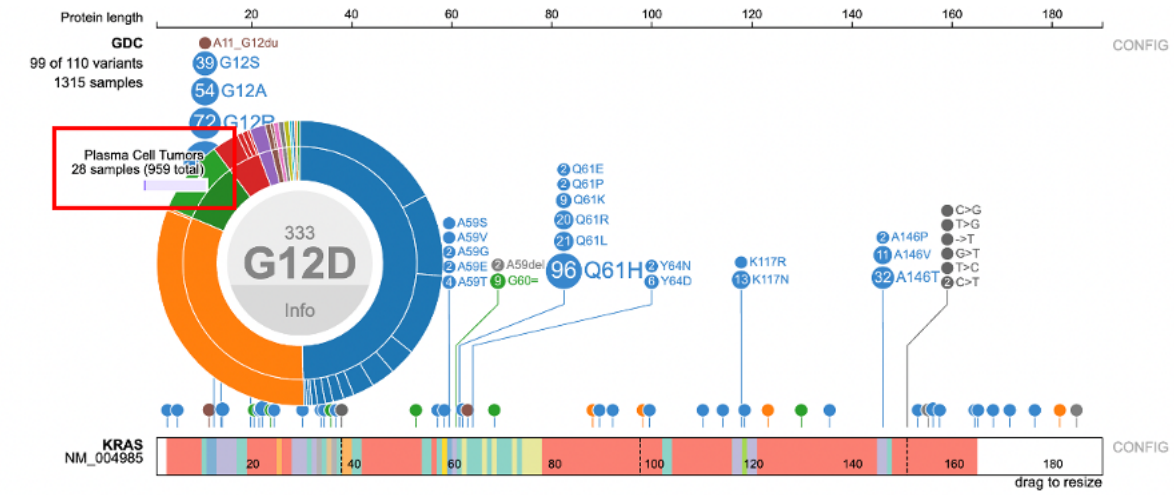
Detailed variant annotation is viewable by clicking on the variant disc next to the label. For G12D highlighted in a red outline in the image above, click on the '333' disc. A sunburst chart will appear, shown below.

The center displays the occurrence of the variant (333) above the variant label. The ring

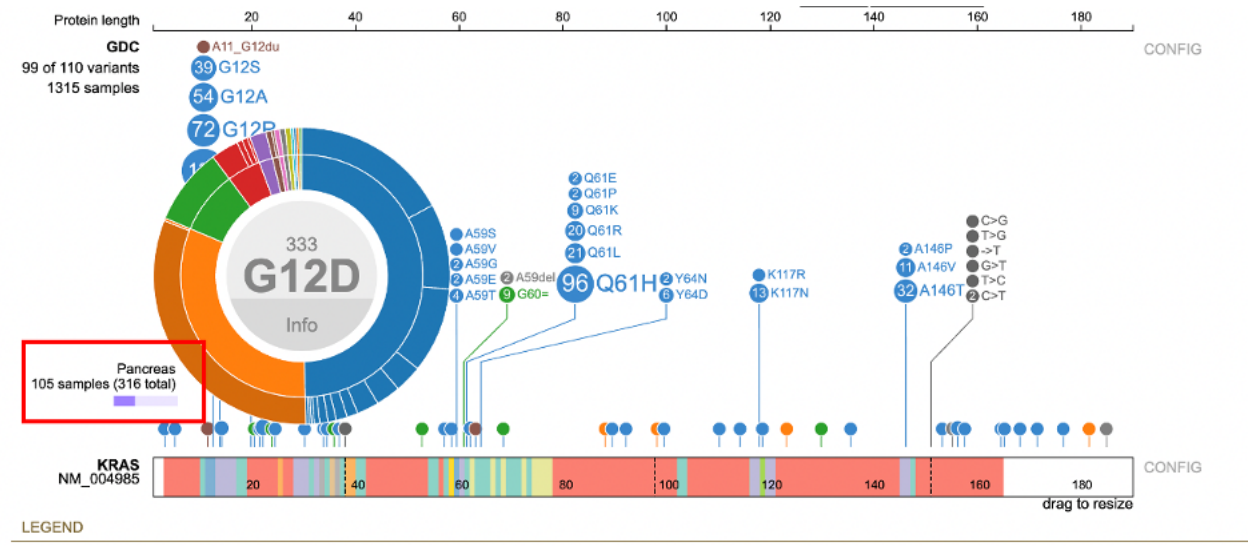


hierarchy is arranged by disease types then broken down by primary sites. Hovering over the inner ring displays the disease type, number of samples, and cohort size. In this example, the inner green ring displays 'Plasma Cell Tumors' with 28 samples out of a total 949 samples.

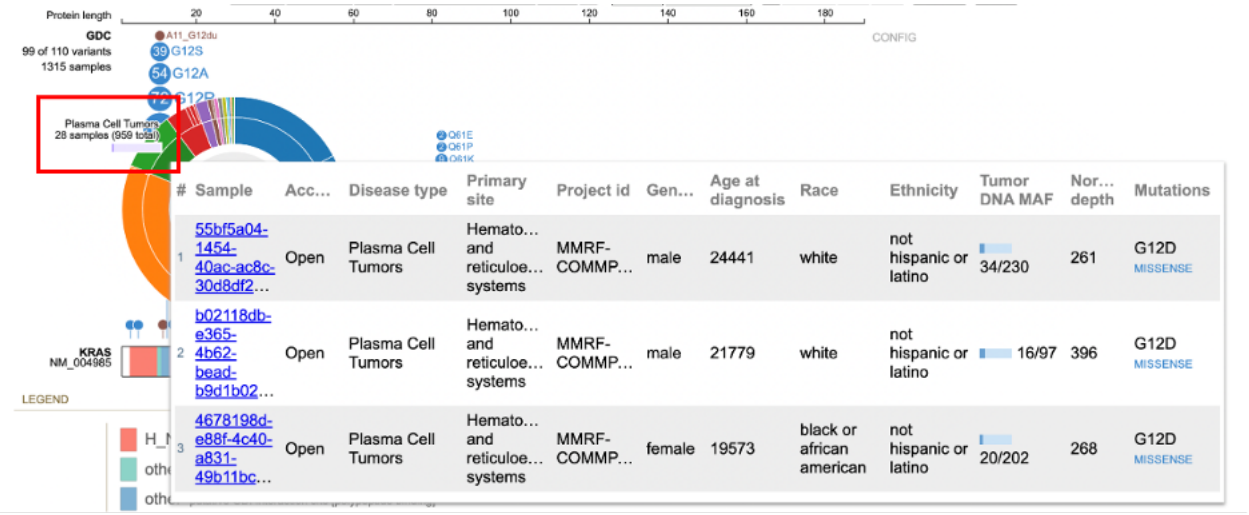
The outer ring represents the primary sites. Hovering over the primary site displays the number of samples relative to the disease type. In the figure below, for Ductal and lobular



neoplasms, there are 105 samples with the primary site as pancreas out of 316 total samples.



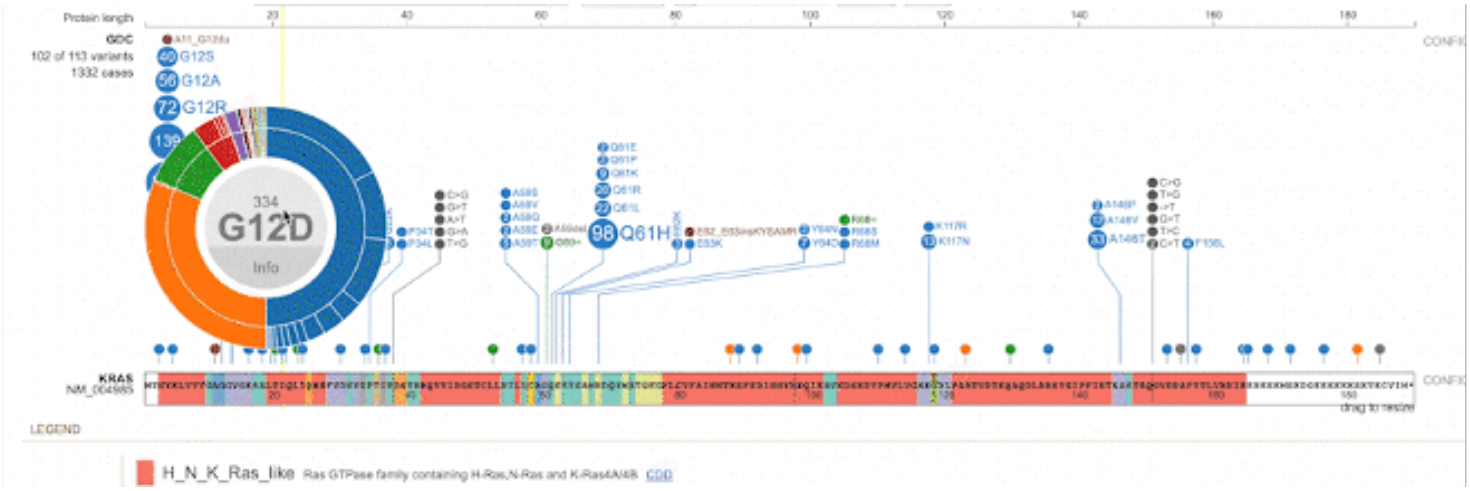
Clicking on a node displays a sample table for the disease type or primary site. In the figure below, the user selected 'Plasma Cell Tumors'. The sample annotation table appears for all Plasma Cell Tumors.



An aggregate sample table is available by clicking the 'Info' button in the center of the sunburst. This displays all the samples associated with that variant. In the screen recording below the aggregate sample table appears for KRAS - G12D.

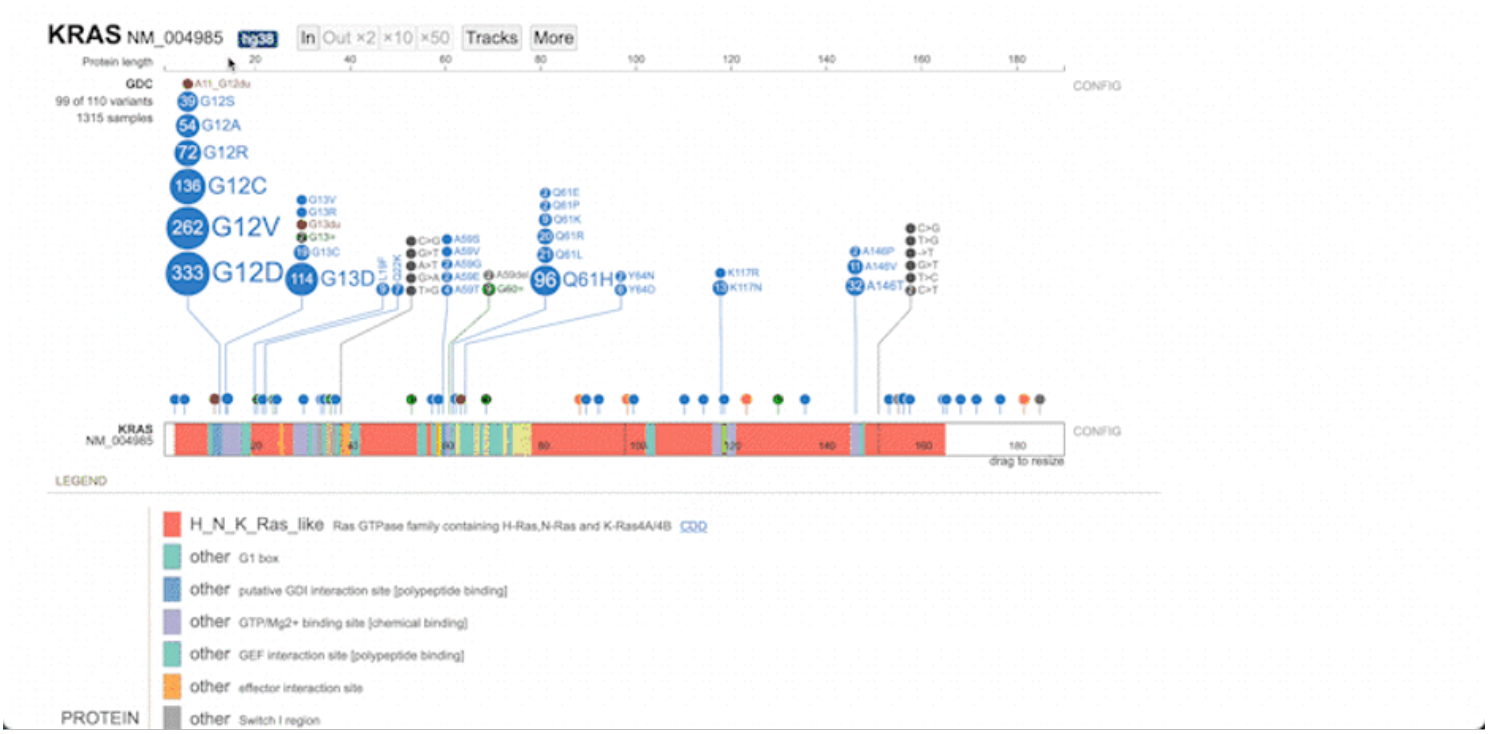
Clicking on the sample name hyperlink opens a new tab to the sample's GDC Case Summary page.

Clicking on the variant label in the center removes the sunburst chart.



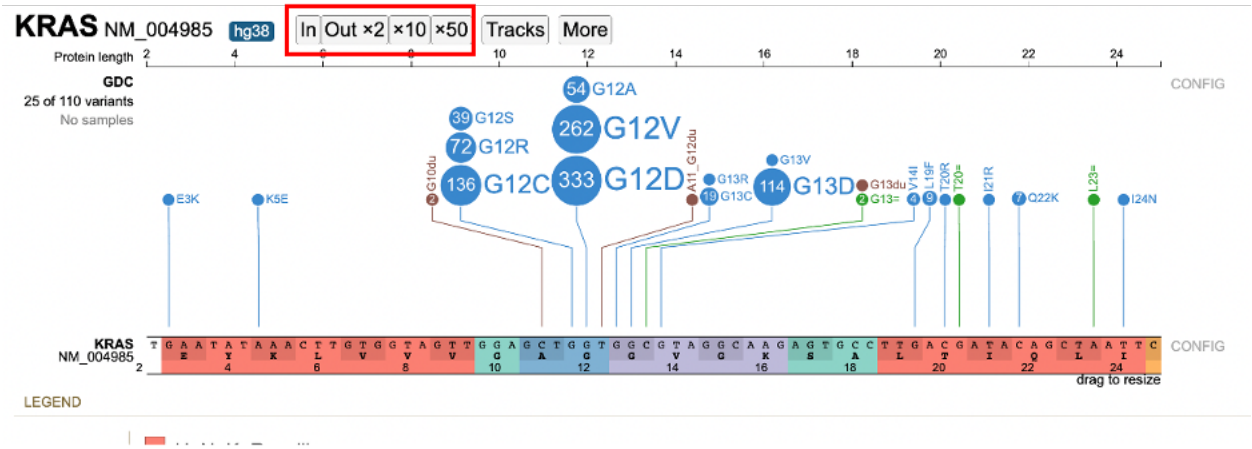
Working With the Protein Track

There are two zoom methods: highlighting a region and zoom buttons in the toolbar. For viewing a nucleotide of interest, click and drag the mouse in the top, x-axis, Protein length scale. The region appears highlighted in red with the calculated protein length in center.

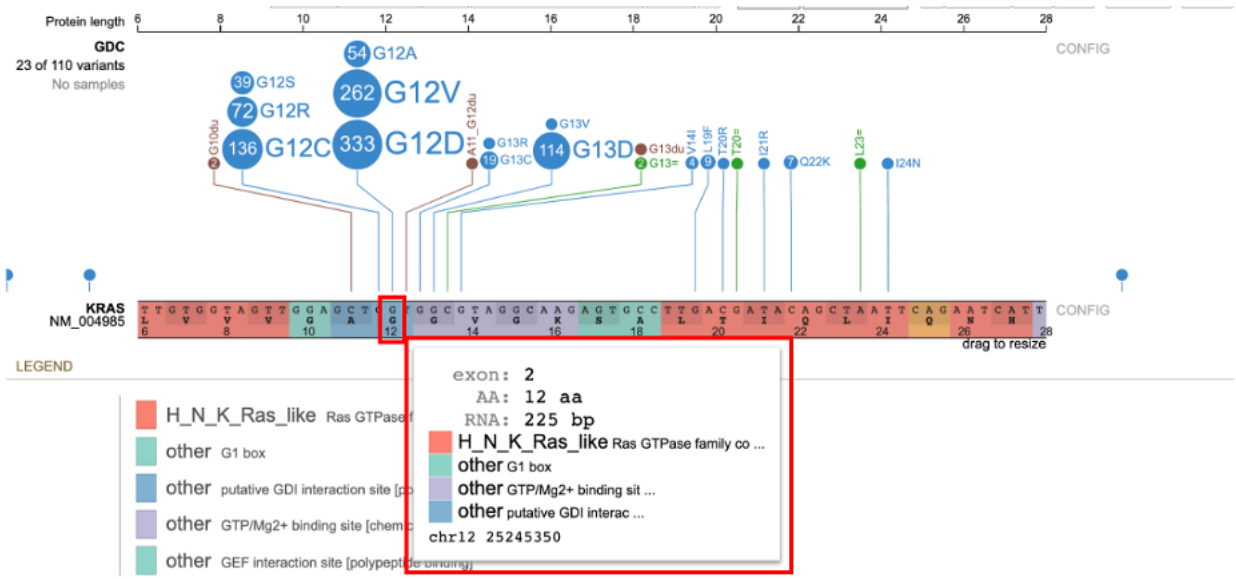


Once the mouse is released, the lollipop re-renders as the selected region.

The zoom buttons in the toolbar is the second option to zoom in and out based on the center position of the lollipop. For zooming out, users can choose to zoom out 2x, 10x or 50x times.

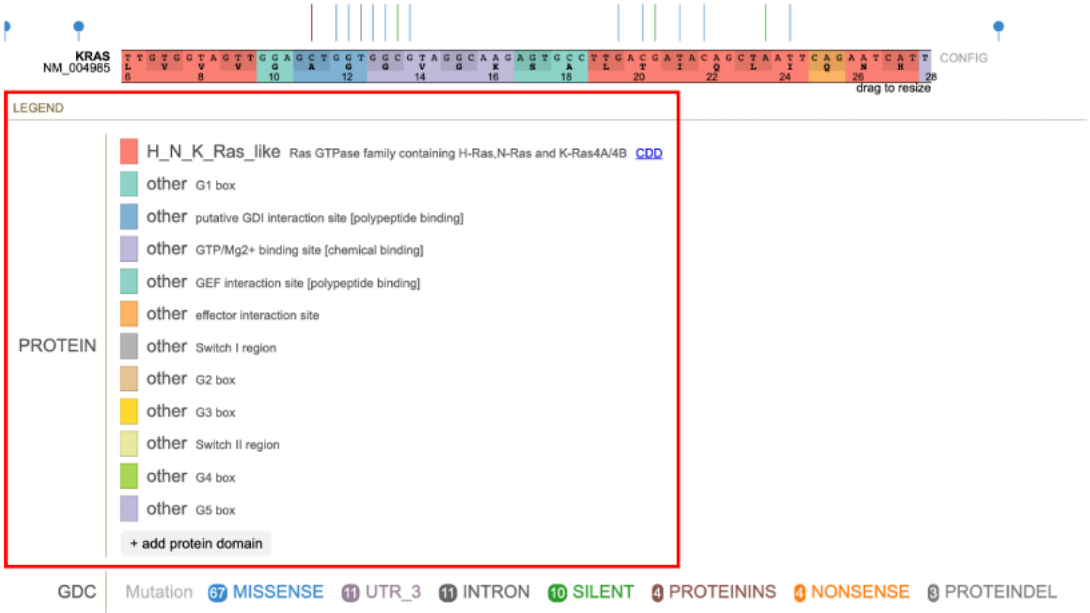


Zooming in causes the protein track to display the codons and the nucleotides as shown below. Hovering over the nucleotide position displays a tooltip with the exon, amino acids position, RNA position, and protein domain. As shown in the image below, at codon 12, the second exon of the transcript, RNA position 225 bp, the reference allele is a 'G'. There is a substitution at 'G' to A, V and D in the KRAS gene for isoform NM_004985 for which the cases are as shown below.



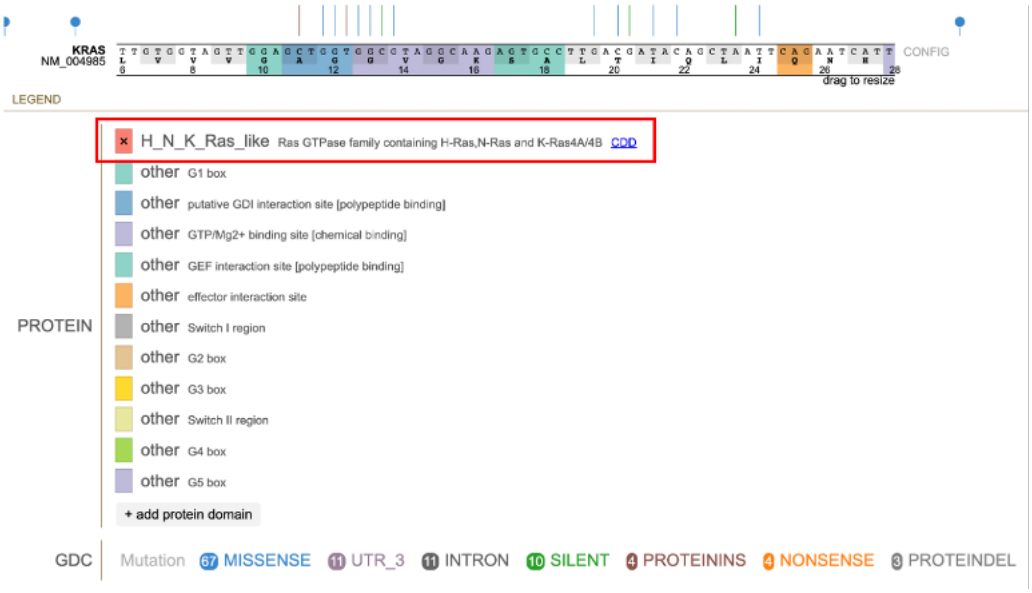
Legend Panel

The protein track color codes regions by the protein domain present on the full-length protein region in the exon display. For KRAS, the protein domains are shown in the red box in the image below.



Protein Domain Legend

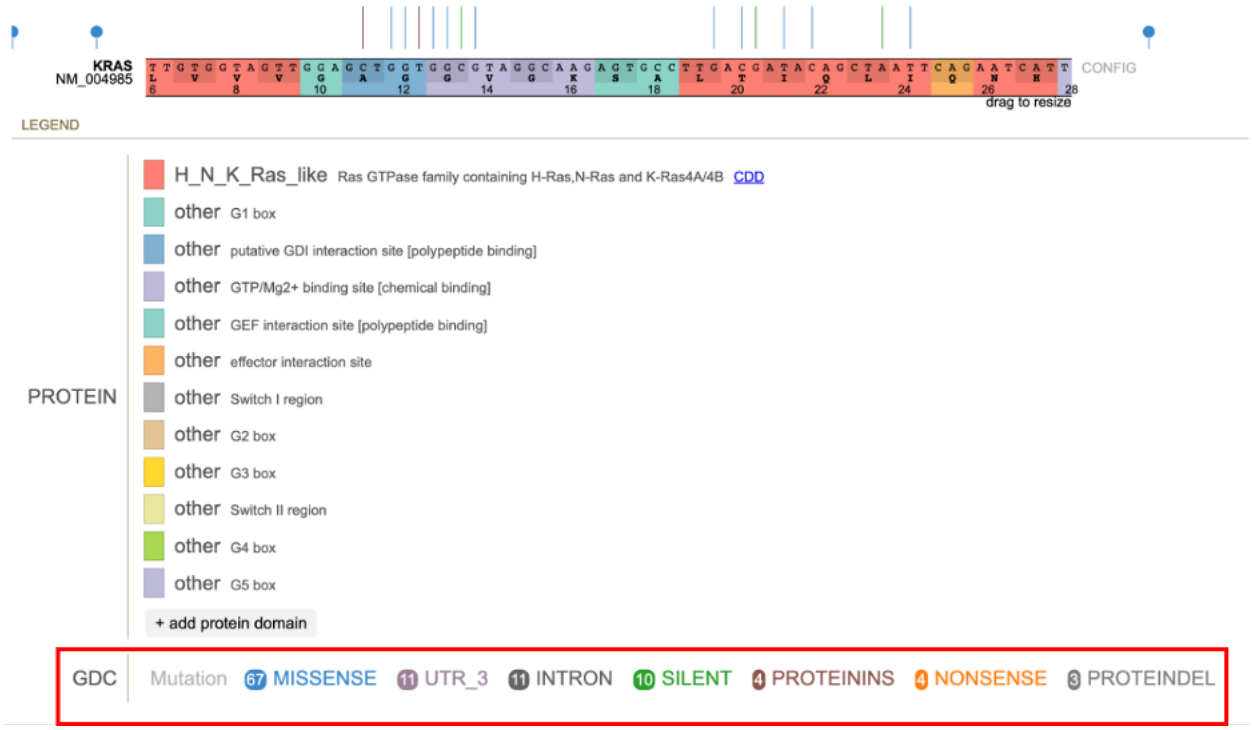
Clicking on the colored box next to the protein domain label removes the color from the protein track, as depicted below.



Custom protein domains are added by clicking on the '+add protein domain' button at the bottom of the list. An input box appears requiring the following information: 1. Name, text with space, no semicolon: This is the name of the protein domain 2. Range, two integers joined by space: This is the codon position - start and stop 3. Color (e.g., red, #FF0000, rgb (255,0,0)): This is the color to assign to the protein domain.

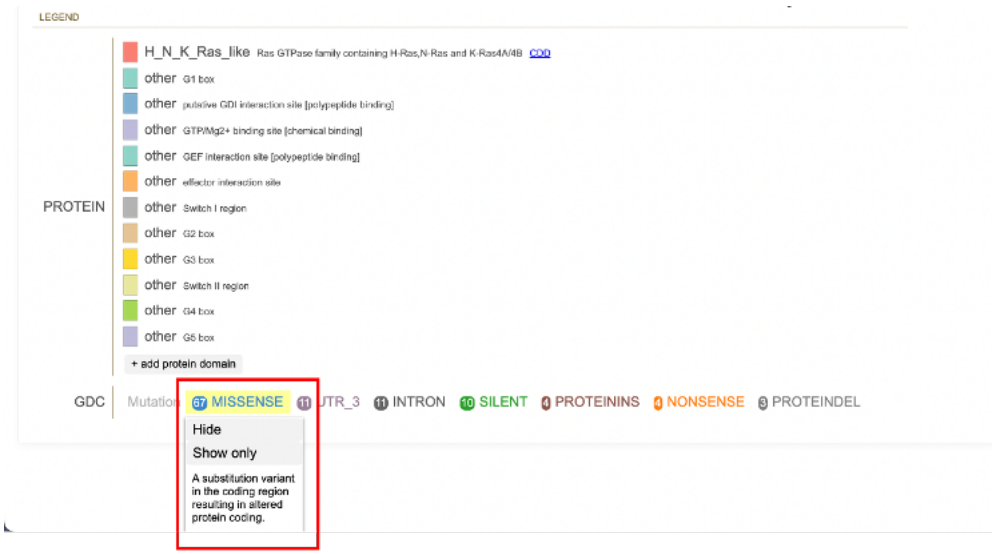
GDC Mutations

The lollipop discs are color coded per GDC mutation classes. The legend for the mutations appears below the protein domains with more advanced show/hide functions.



The classification for the type of variant is color coded as follows:

Clicking on a mutation prompts a pop-up menu to appear with the description of the mutation. Options to ‘hide’ or ‘show only’ are specific to the mutation. The option ‘show all’ includes all previously hidden mutations. Selecting ‘MISSENSE’ shown in the figure below by the yellow highlight displays the initial menu with the ‘hide’ and ‘show only’ buttons.



Clicking ‘Hide’ removes all of the mutation discs from the lollipop. The mutation is reordered to the end of the list and the font is striked through and grayed out. The discs reappear when the mutation label is clicked again.

More Options

ProteinPaint offers methods to download figures and data. Click the ‘More’ button in the toolbar to display various options as shown below.

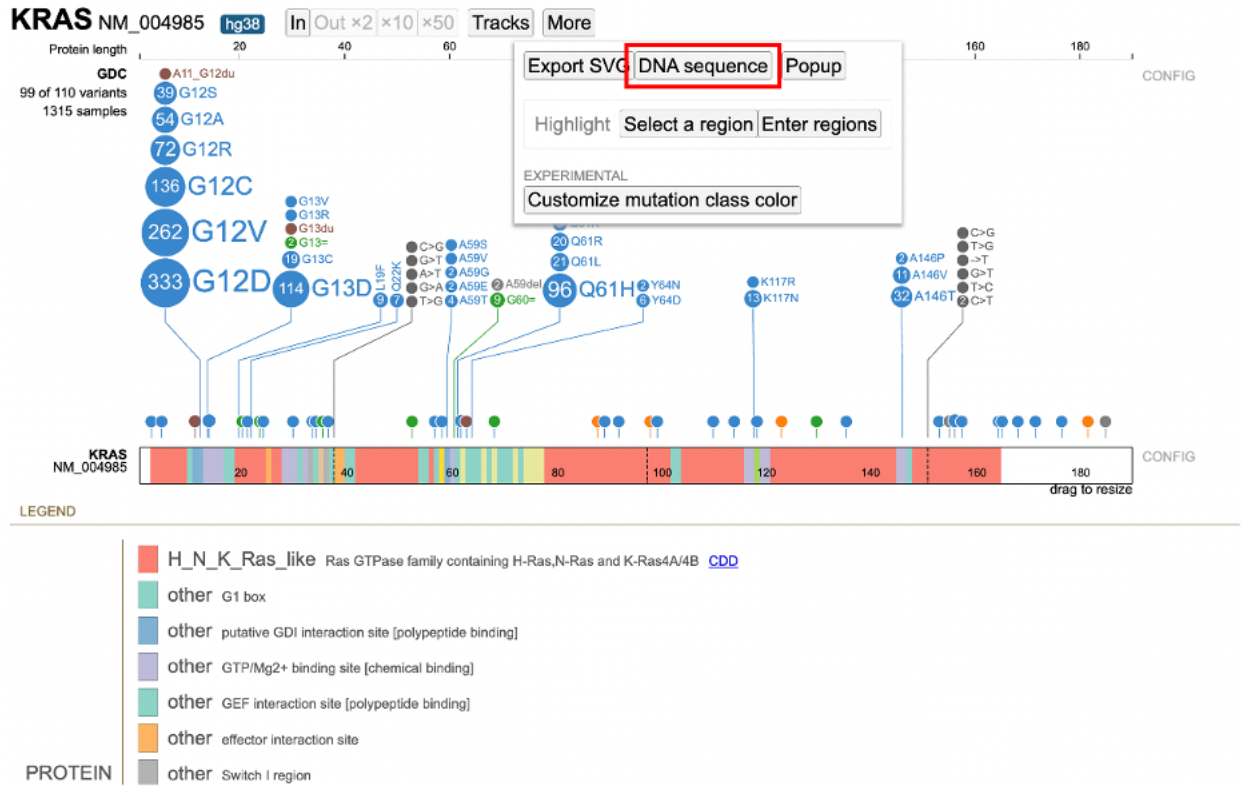
Exporting the Figure

Click “Export SVG” to download the lollipop and legend as an SVG file.

The exported figure will contain following contents, reflecting a user’s customization: * Displayed datasets, including custom data * Expand/fold states of all mutations * Sequences in the protein if at zoom-in level * Show/hide state of exon boundaries * Sunburst charts * Protein domains without the hidden ones * All mutations without the hidden classes or origins * Legend for protein domain, mutation class and origin

Copying the DNA Sequence

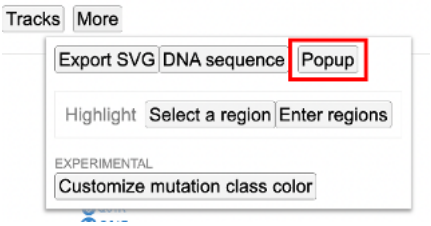
The ‘More’ button also includes a ‘DNA sequence’ button.



Clicking on 'DNA sequence' displays the DNA sequence as plain text for easy copying and pasting.

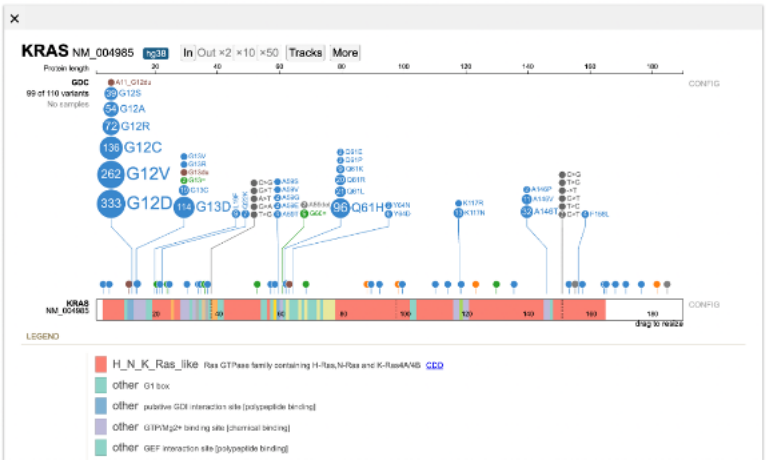
Popup Option

The pop up option under the More button allows for popping open another window with the same lollipop display selected by the user. Below is an example.



To view GDC mutations on a gene, enter one of gene symbol (MYC), alias (c-Myc), GENCODE accession (ENSG00000136997, ENST00000621592), or RefSeq accession (NM_002467).

KRAS ✓ KRAS



Chapter 14

Sequence Reads

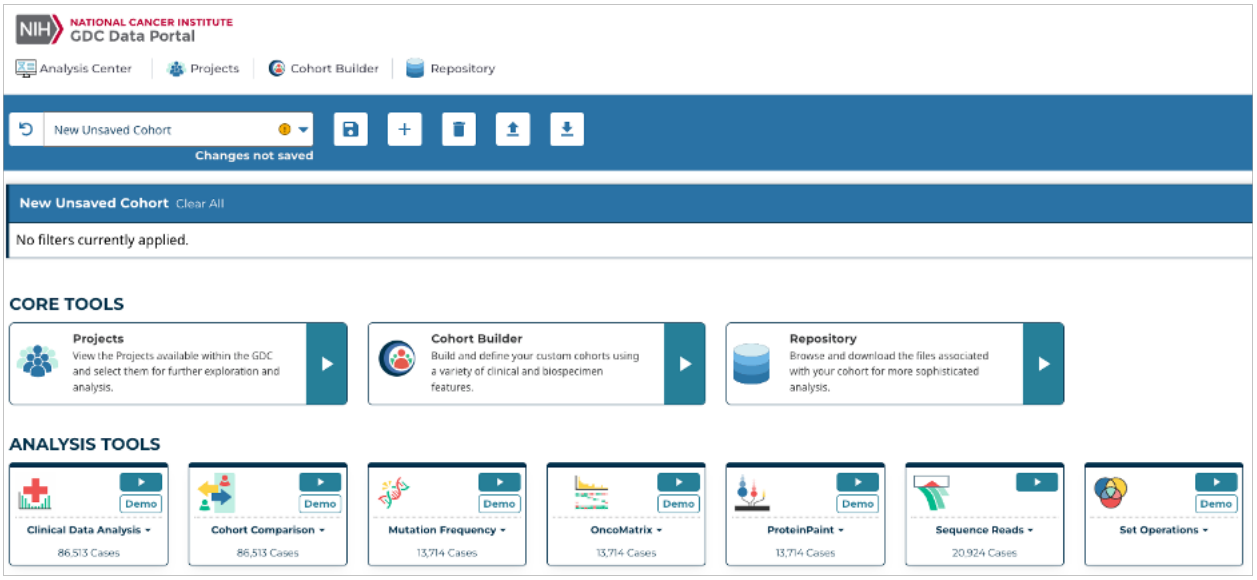
ProteinPaint Sequence Reads Tool

Introduction to ProteinPaint Sequence Reads

Sequence Reads is a web-based tool that uses the ProteinPaint BAM track and NCI Genomic Data Commons (GDC) BAM Slicing API to allow users to visualize read alignments from a BAM file. Given a variant (i.e. Chromosome number, Position, Reference Allele and Alternative Allele), the Sequence Reads tool can classify reads supporting the reference and alternative allele into separate groups.

Quick Reference Guide

At the Analysis Center, click on the 'Sequence Reads' card to launch the app.



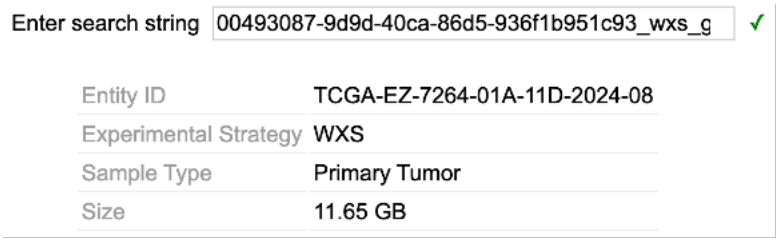
This feature requires access to controlled data, which is maintained by the Database of Genotypes and Phenotypes (dbGaP) See Obtaining Access to Controlled Data. In order to use this tool, users must be logged in with valid credentials. Otherwise, users will be prompted to login.

Selecting BAM Files and Variants

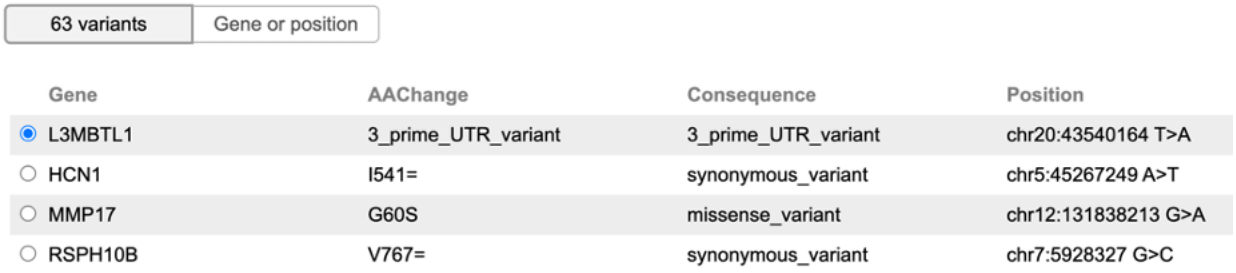
Once logged in, the Sequence Reads tool will display a search bar, as well as a link to browse the first 1,000 available BAM files for the active cohort. Users can choose to select a BAM from the available list, or search for a specific BAM file by entering four types of inputs: file name, file UUID, case ID, or case UUID.

The tool will verify the query string and return all matching GDC BAM files in a table, from which the user can select one or multiple to use with the tool.

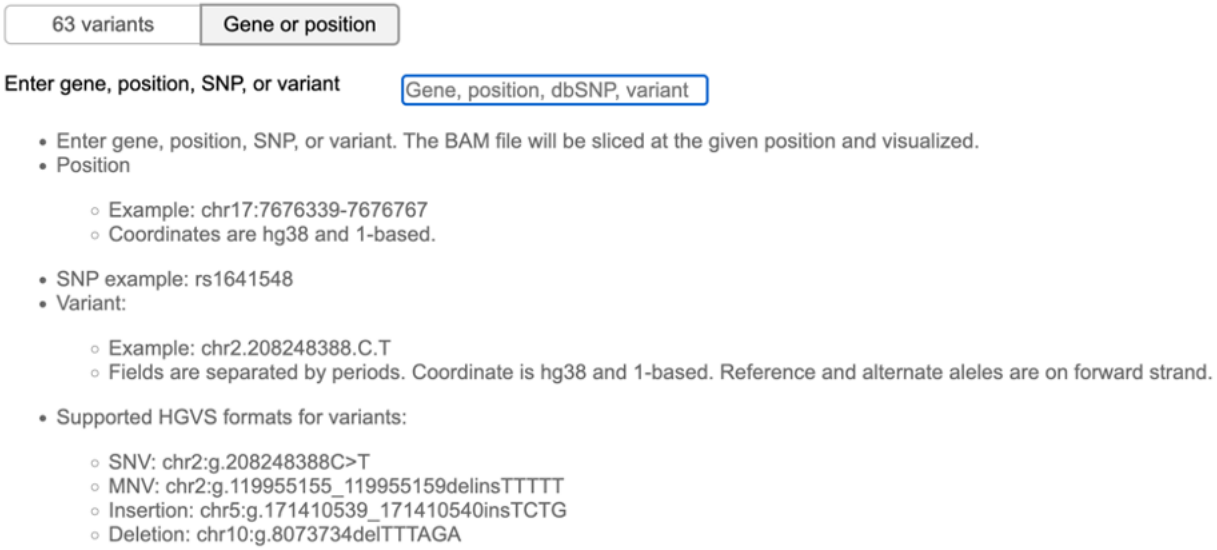
If an exact match is entered (i.e. a file name or file UUID), the Sequence Reads tool will find that BAM and present brief information about the file.



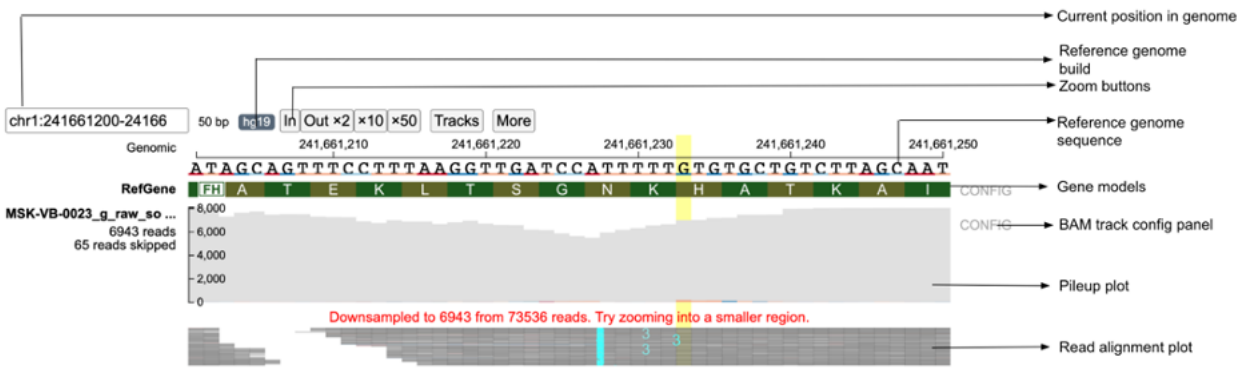
In the subsequent section, the mutation table displays somatic mutations catalogued by the GDC for this case, if available. Users can select a mutation to visualize read alignments on this variant.



Alternatively, the `Gene or position` button at the top of the mutation table allows users to enter a custom genomic region for BAM visualization.



Once at least one BAM file is selected and a gene, position, SNP, or variant is entered, the Sequence Reads tool will display the BAM visualization.



Toolbar

- **Current Position in Genome:** Displays the coordinates of the region currently displayed
- **Reference Genome Build:** Refers to the genome build that was used for mapping the reads; the GDC uses Reference Genome Build 38 (hg38)
- **Zoom Buttons:** Zooms in (`In`) or out (`Out x2`, `x10`, and `x50`) of the current view

Reference Genome Sequence

The Reference Genome Sequence displays the reference genome build against which the reads have been aligned.

Gene Models

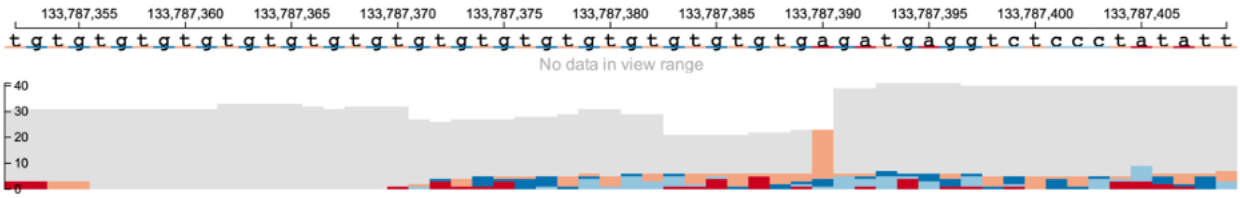
The Genome Models row displays the gene model structure from the view range. When zoomed into a coding exon, the letters correspond to the 1-letter amino acid code for each amino acid and are placed under its corresponding 3-letter nucleotide codon under the reference genome sequence. The arrows describe the orientation of the strand of the gene model being displayed (right arrow for forward strand and left arrow for reverse strand).

Graphical representations of the reads are displayed as they are aligned on the chromosome. Sequence can be read when zoomed in.

ProteinPaint BAM Track

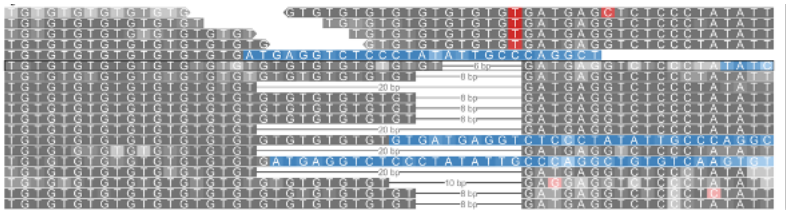
Pileup Plot

The Pileup Plot shows the total read depth at each nucleotide position of the region being displayed.

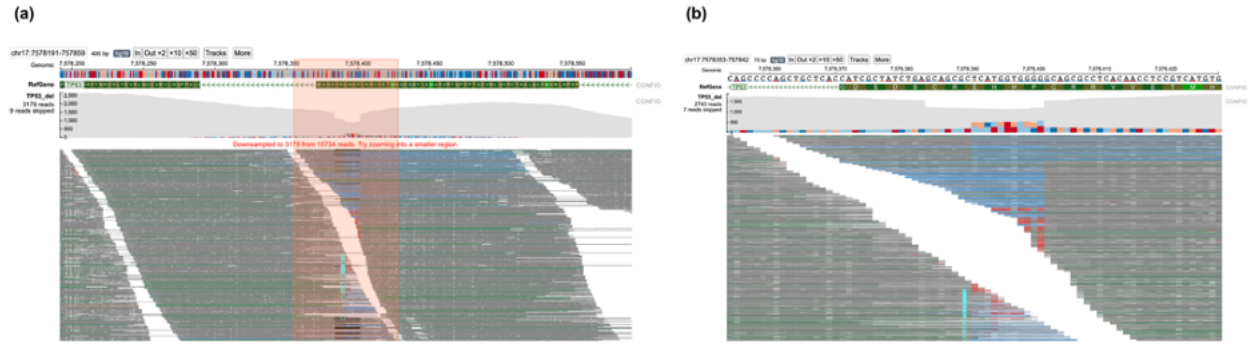


Read Alignment Plot

This visual contains the main read alignment plot of the reads from the BAM file.



When completely zoomed out, base-pair quality of each nucleotide in each read is not displayed. Users can zoom into the plot via the toolbar or by dragging on the genomic ruler (a) to zoom into the selected region (b).



Clicking on a read in the plot launches a window that displays the alignment between the read and reference sequence, as well as the chromosome, coordinates, read length, template length, CIGAR score, flag, and name. If the read is paired, the position of the other segment will be displayed below. This pop-up also contains two buttons, `Copy read sequence` and `Show gene model`, which copy the nucleotide sequence of the read to the computer clipboard and display the gene model, respectively.

Mutation Rendering

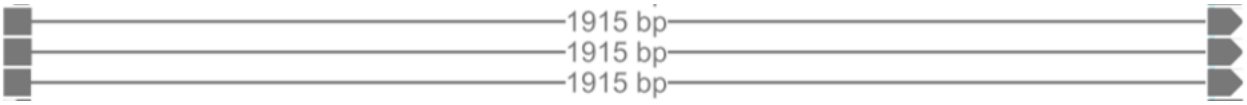
Mutations are rendered as follows:

- **Insertion:** The alphabet representing the nucleotides is displayed between the two reference nucleotides in cyan color, with the shade scaled by base quality
  - If more than one nucleotide is inserted, a number is printed between the two reference nucleotides indicating the number of inserted nucleotides
  - Clicking the read with multiple insertions will display the complete inserted nucleotide sequence

- **Deletion:** A black line represents the span of deleted bases

- **Substitution or Mismatch:** The substituted nucleotide is highlighted in red background, with the shade of red scaled by base quality

- **Splicing:** The different fragments of a read separated due to splicing are joined by a gray line



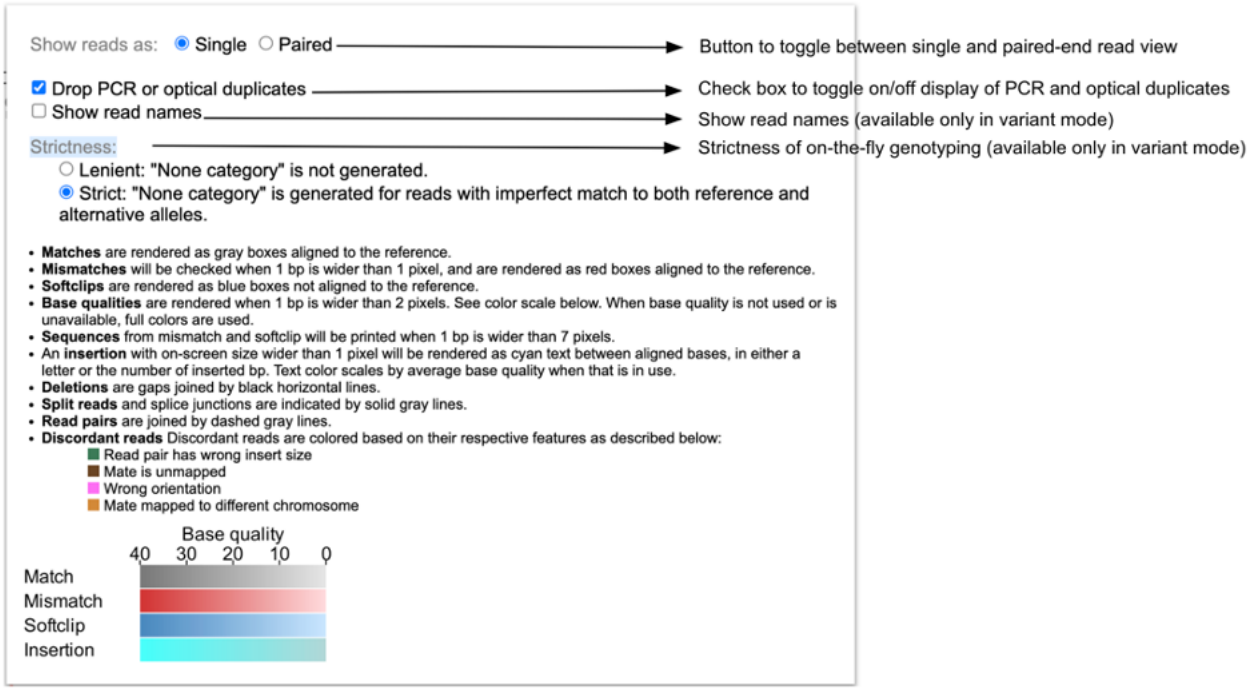
Color Coding of Reads

Color codes in the background of the read describe the quality of the read alignment and its mate (in case of paired-end sequencing). These colors are assigned both on the basis of the CIGAR sequence (if it contains a softclip) and the flag value of both the read and its mate.

- **Gray:** Suggests that both the read (at least part of it) and its mate are properly aligned and the insert size is within expected range
- **Blue:** Indicates that part of the read is soft clipped
- **Brown:** Indicates that the mate of the read is unmapped
- **Green:** Indicates that the template has the wrong insert size
- **Pink:** Indicates that the orientation of the read and its mate is not correct
- **Orange:** Indicates the read and its mate are mapped in different chromosomes

BAM Track Configuration Panel

The BAM Track Configuration Panel, which can be accessed by clicking the `CONFIG` option next to the Pileup plot, provides buttons for toggling between single-end and paired-end mode.

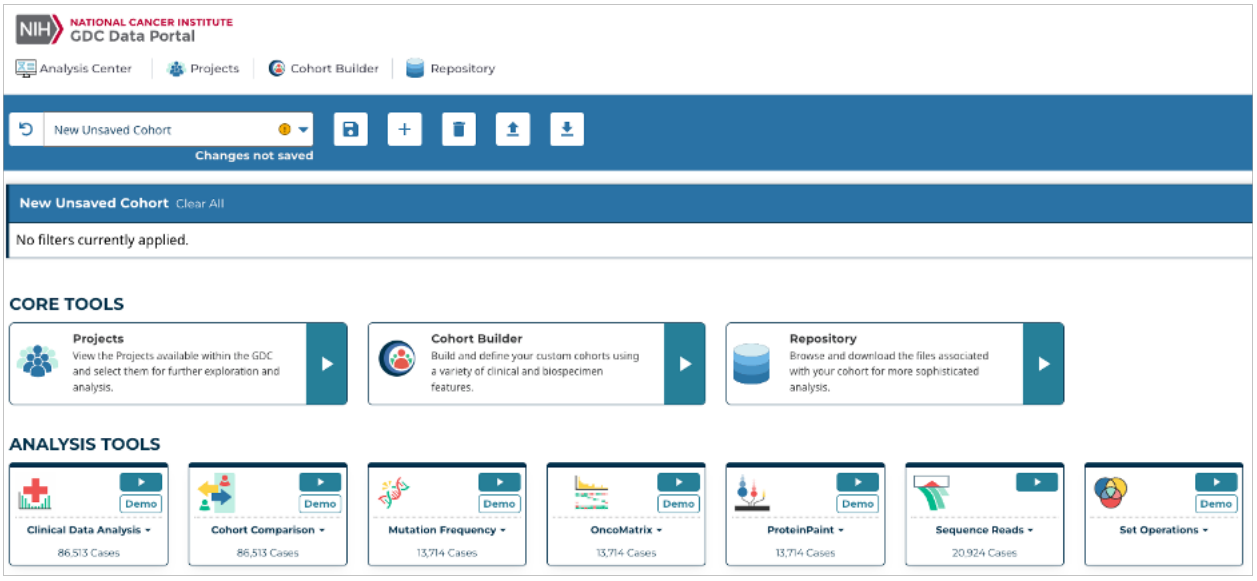


In single-end display each read is displayed individually without displaying any connections with its respective mate. In case of the paired-end display the two paired reads are joined by a gray dotted-line if the coordinates of the two reads do not overlap. When the coordinates of the two read-pairs overlap, the overlapped region is highlighted by a blue line.

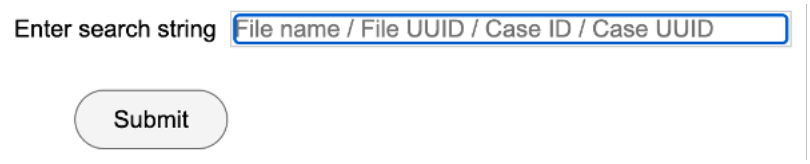
The BAM Track Configuration Panel also provides a check box to show/hide PCR and optical duplicated reads.

Launch the Sequence Reads Tool

At the Analysis Center, click on the “Sequence Reads” card to launch the app.



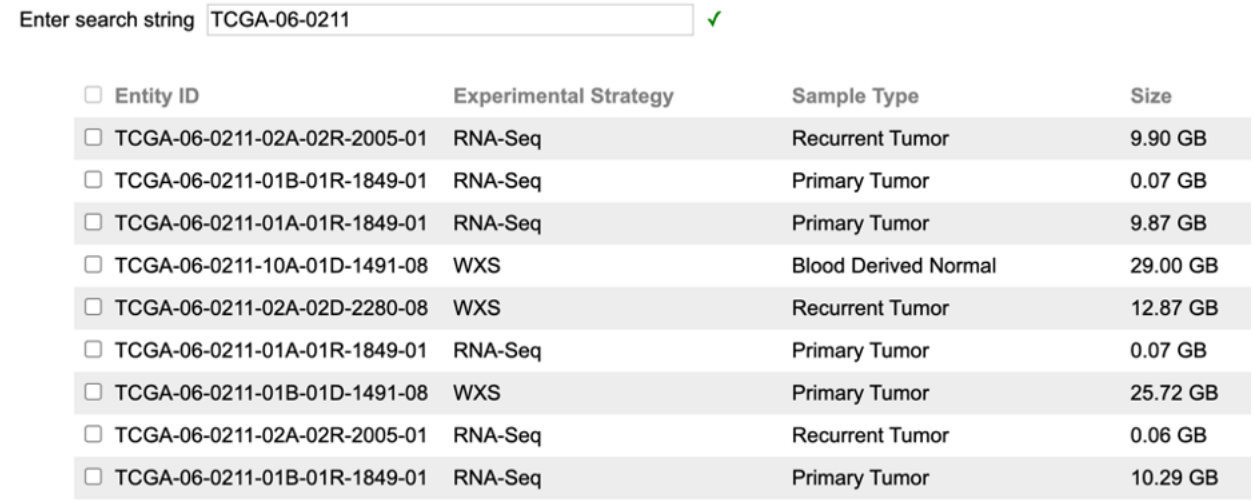
A user needs to be logged in to use this feature. If not, the user will be prompted to log in. Once the user logs in, a search bar and submit button will appear as below.



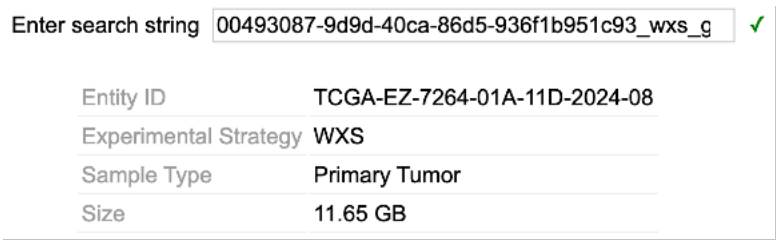
Find and Display BAM Files in GDC

To find a BAM file in GDC, a user can enter four types of inputs including file name, file UUID, case ID, or case UUID. The tool will verify the query string and return matching BAM files.

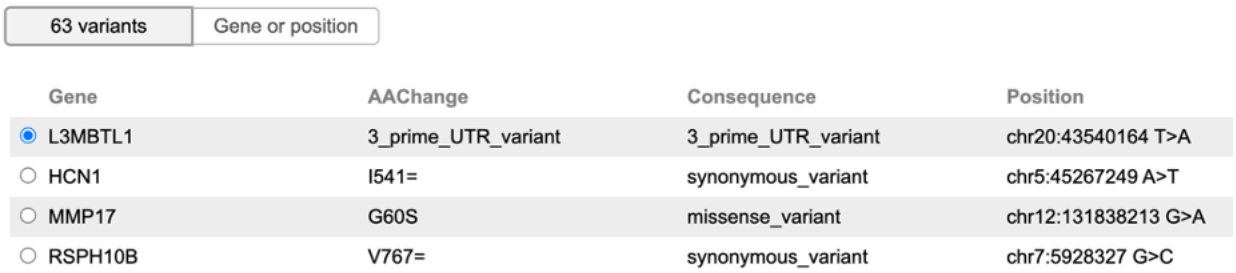
As an example, using case ID "TCGA-06-0211" will return 9 BAM files available from this case displayed in a table. One or multiple BAM files from this table must be selected to proceed.



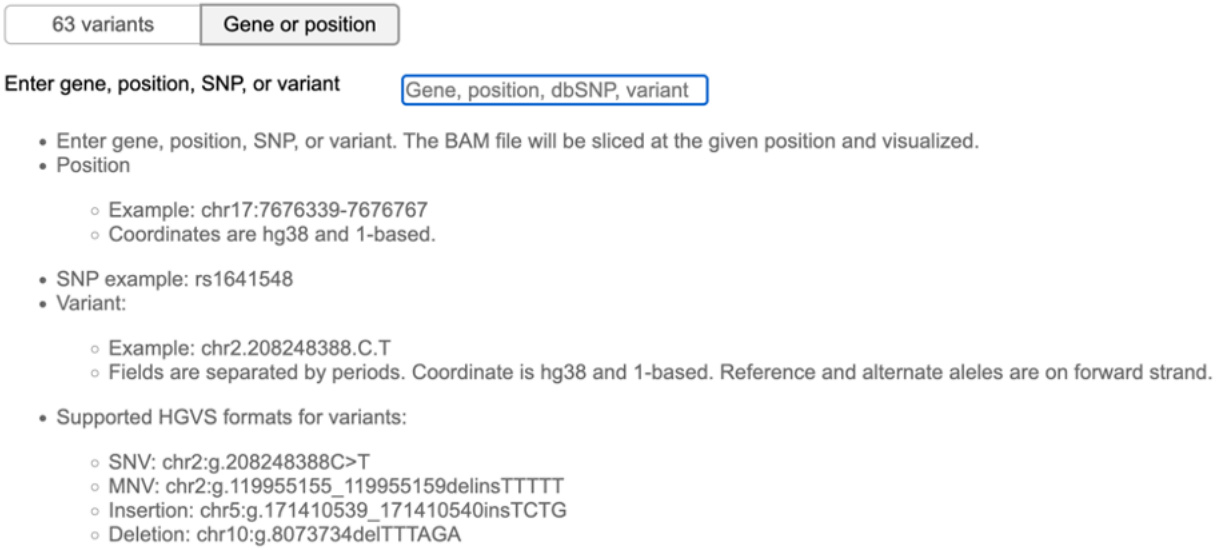
When a file name or UUID is provided, it will display brief information about the file. A user does not need to select anything here as the file is automatically selected.



The subsequent section displays somatic mutations catalogued by GDC for this case, if available. A user can select a mutation to visualize read alignment on this variant.



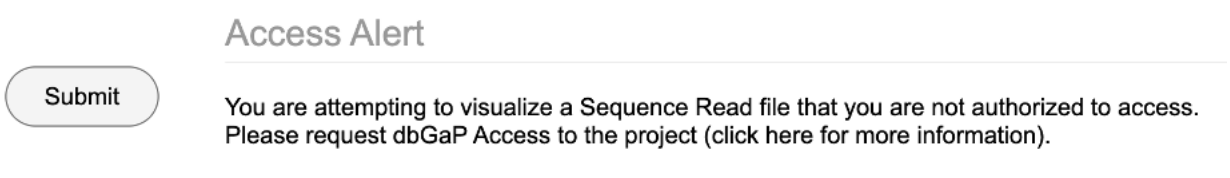
Alternatively, a user can enter a custom genomic region for BAM visualization. At the toggle button on top of the mutation table, click the "Gene or position" option to show the gene search box.



Follow the instructions to enter gene, position, SNP, or variant. Press ENTER to validate the input.

Lastly, press the "Submit" button to view read alignment from the selected BAM file over the selected mutation or genomic region. The server will verify the user's access to the requested BAM file and query the GDC API to slice the BAM file at the selected region. This may take from 10 seconds to a minute.

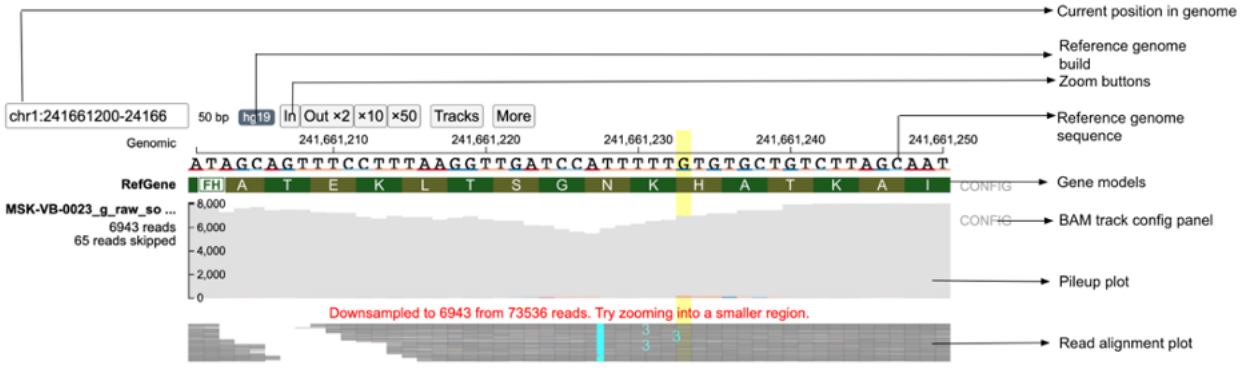
An error message will appear if the user does not have access to the requested BAM file. Please follow the instructions to obtain access.



Once the BAM visualization is successfully displayed, the search interface is hidden, and a button named "Back to input form" is shown. Clicking the button will bring the user back to the search interface so a user can change the BAM file or mutation.

Click the "Download GDC BAM Slice" button to download the BAM slice file used in this visualization.

Using ProteinPaint Genome Browser



Various fields labeled in the above figure are described below:

Current Position in Genome

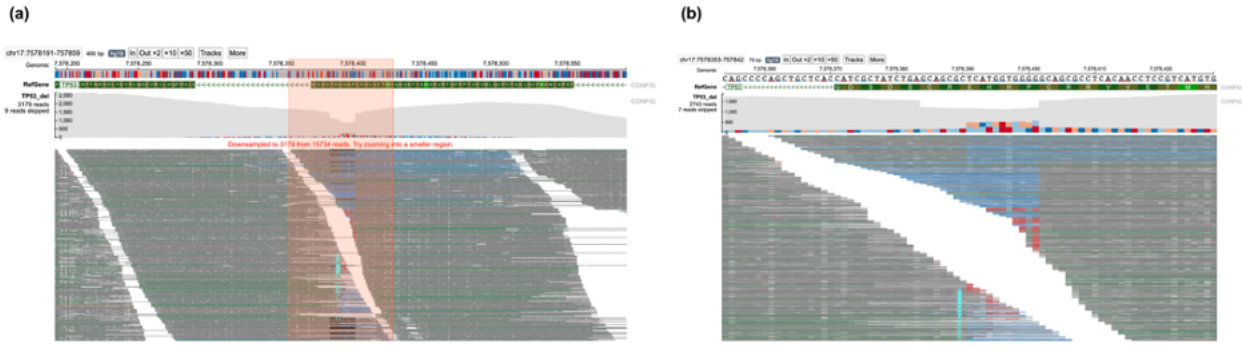
The Current Position in Genome text box displays the coordinates of the region currently displayed on the screen. It initially shows the coordinates specified in the URL. On pan/zoom by the user, this region displays the updated coordinates of the view region.

Reference Genome Build

The Reference Genome Build button refers to the genome build specified by the user that was used for mapping the reads. The GDC uses Reference Genome Build 38 (hg38).

Zoom Buttons

A user can zoom in/out of the current view by clicking the "In" (zoom in) or "Out x2" (zoom out) buttons. By clicking on the x10 and x50 button, a user can zoom out 10 and 50-fold respectively. Alternatively, a user may choose to zoom into a smaller region by dragging on the genomic ruler (a) to zoom into the selected region (b) as shown below.



Reference Genome Sequence

The Reference Genome Sequence displays the reference genome build against which the reads have been aligned.

Gene Models

This Genome Models row displays the gene model structure from the view range. When zoomed into a coding exon, the letters correspond to the 1-letter amino acid code for each amino acid and are placed under its corresponding 3-letter nucleotide codon under the reference genome sequence. The arrows describe the orientation of the strand of the gene model being displayed (right arrow for forward strand and left arrow for reverse strand).

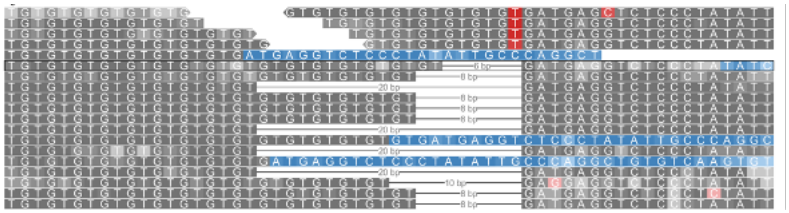
ProteinPaint BAM Track Features

Pileup Plot

The Pileup Plot shows the total read depth at each nucleotide position of the region being displayed. Color codes of bars representing various possibilities:

Gray - Reference allele nucleotides Blue - Soft clipped nucleotides Mismatches: * nucleotide “A” - Red (color code: #ca0020) * nucleotide “T” - Orange (color code: #f4a582) * nucleotide “C” - Light blue (color code: #92c5de) * nucleotide “G” - Dark blue (color code: #0571b0)

Read Alignment Plot



The Read Alignment Plot contains the main read alignment plot of the reads from the BAM file.

Rendering of Various Mutations

Insertion

In case of a single nucleotide insertion, the alphabet representing the nucleotide (A/T/C/G) is displayed between the two reference nucleotides in cyan color.

Darkness of the inserted nucleotide is determined by the base quality, as an example below of an inserted T with low quality.

If more than one nucleotide is inserted, a number is printed between the two reference nucleotides indicating the number of inserted nucleotides. The text color is full cyan and does not account for the quality of inserted bases. Showing below is a read with two insertions, first with 2 bases, and second with T.

On clicking this read, the read information panelthe read information panel is displayed where the complete inserted nucleotide sequence is shown in cyan color.

Deletion

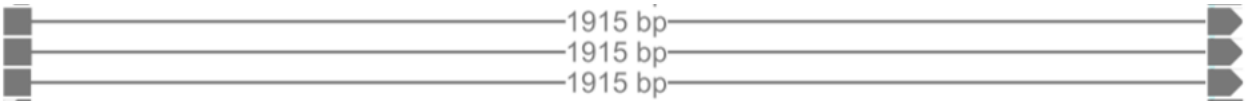
A black line represents the span of deleted bases.

Substitution (or Mismatch)

In case of substitutions (or mismatches), the substituted nucleotide ("A") is highlighted in red background, with the shade of red scaled by base quality.

Splicing

In case of splicing, the different fragments of a read separated due to splicing are joined by a gray line as shown below. In the example below, the reads contain spliced fragments that are separated by a 1915bp intron.



Zooming the Read Alignment Plot

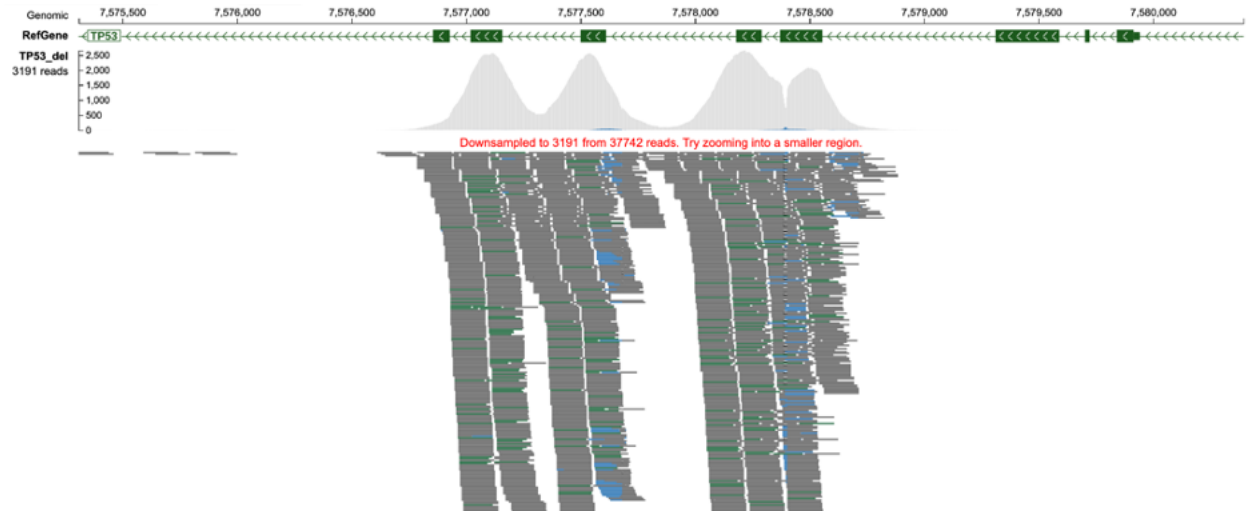
The rendering of the reads depends upon the zoom level (horizontal zoom) chosen by the user and the number of reads mapped at the display region (vertical zoom).

Horizontal Zoom

The BAM track has three levels of horizontal zoom:

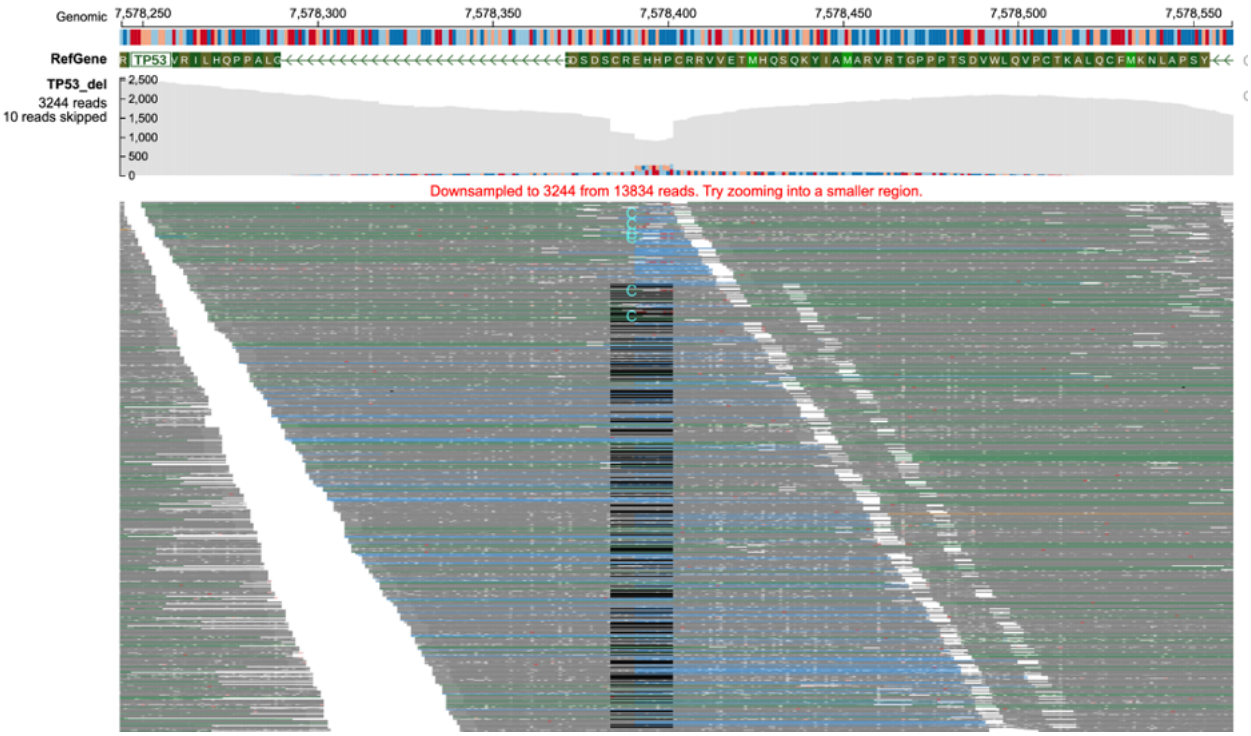
Overview Level

This is the completely zoomed out mode (shown below). At this resolution, base-pair quality of each nucleotide in each read is not displayed as each read occupies a very small area on the screen. Also the reference sequence at the top is not displayed. Only reads which contain big insertions/deletions/softclips or are discordant are represented by their respective colors Also the (see color codes of various reads).



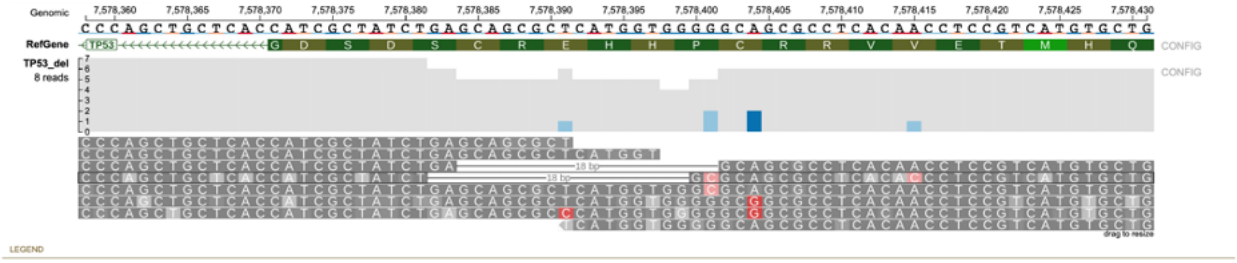
Base-pair Quality Level

At this level of zoom (shown below), in addition to color codes of reads, the phred base pair quality score of each read is also displayed. Poor base-pair quality of nucleotides is represented by lighter shades of the respective color and darker shades represent high base-pair quality. For example, dark gray color represents a higher quality nucleotide in a properly mapped read than light gray which represents poor base-pair quality.



Base-pair Resolution Level

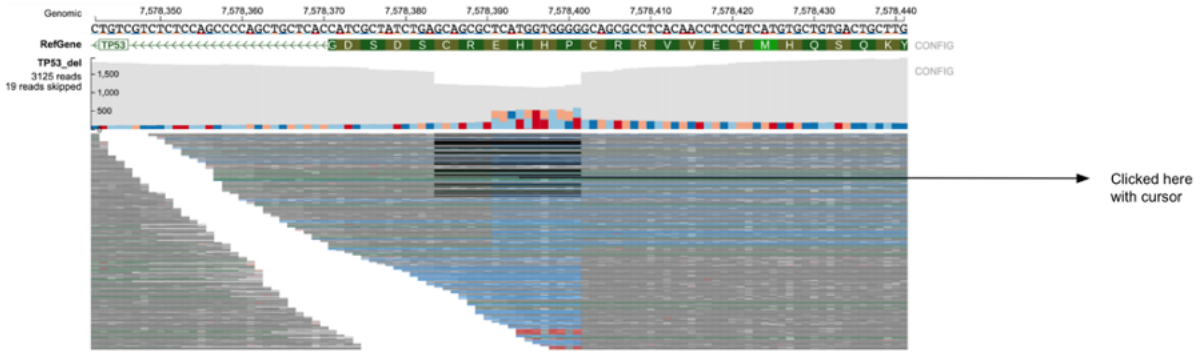
At this resolution, all information including the read sequence of each read is displayed along with reference genome nucleotides at the top. For simplicity (as discussed later under vertical zoom), only a few reads are shown in the figure below.



Vertical Zoom: Examining Subset of Reads

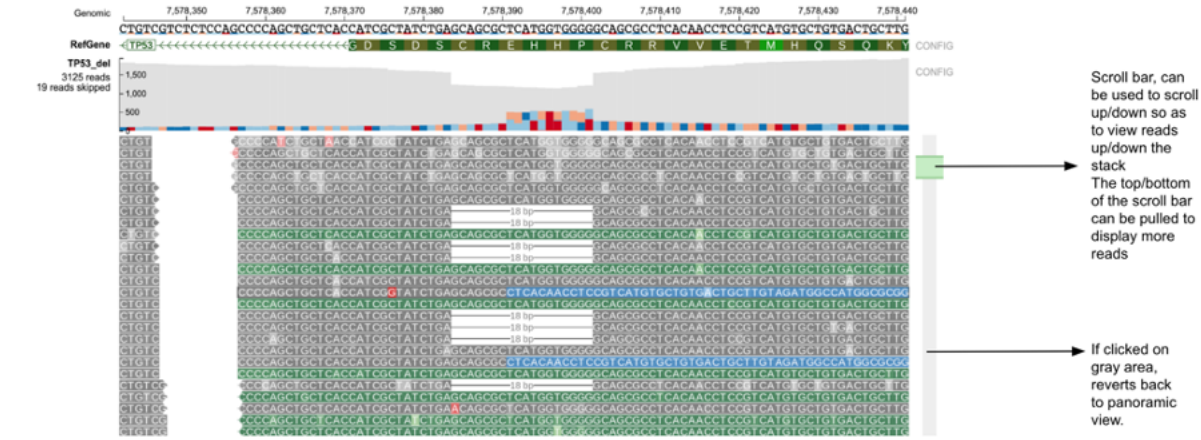
ppBAM can display up to 7000 reads, and will downsample if the number of reads in a region is over 7000. This is especially helpful for displaying high-depth sequencing data. However, displaying nucleotides from each read for such a large number of reads is not feasible. Therefore, the pixel width of each read is reduced to accommodate all reads in the region (Panoramic view, figure below). When the user clicks on a read, that part of the alignment stack is enlarged to show the nucleotides within each read (Nucleotide view, figure below) stacked near the cursor click. Reads at the top and bottom of the stack can be viewed by scrolling up/down with the scroll-bar. The top/bottom of the green scroll-bar can be adjusted to display more reads on the screen by reducing the individual width of each read. On clicking the gray area of the scroll bar region, the panoramic view is displayed again.

(a) Panoramic view



After clicking on a read

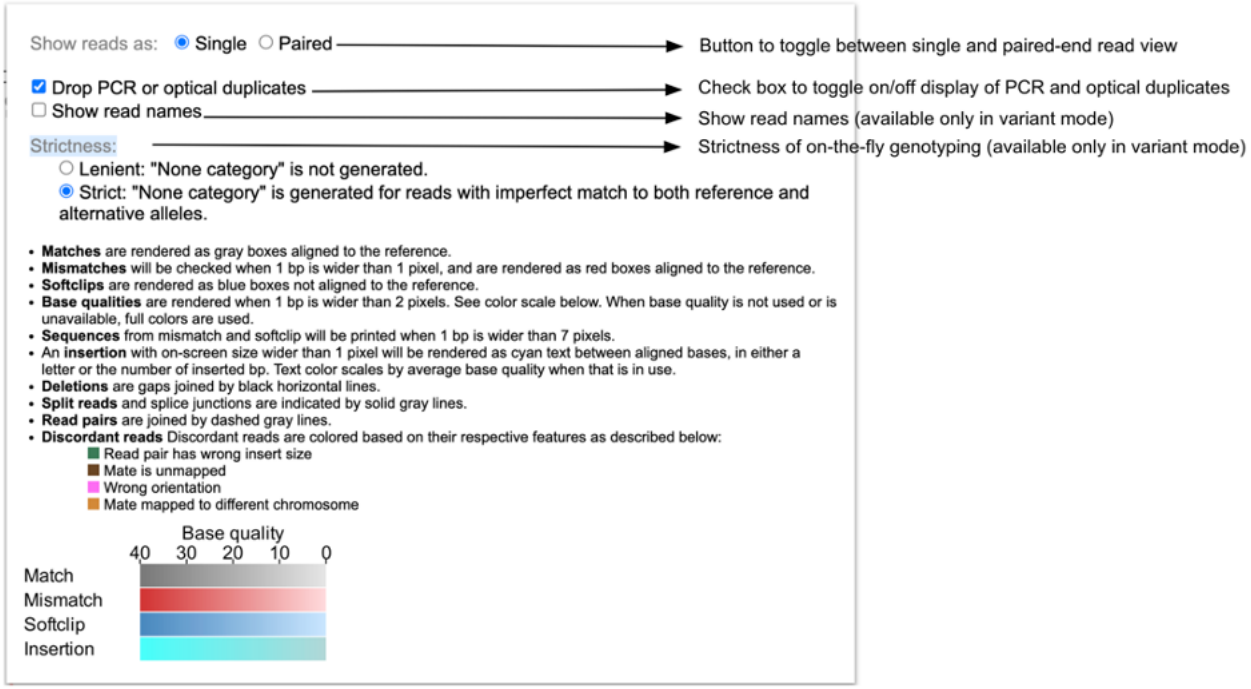
(b) Nucleotide view



BAM Track Configuration Panel

The BAM Track Configuration Panel can be accessed by clicking the "CONFIG" option next to the pileup plot. The BAM Track Configuration Panel (shown below) provides buttons for toggling between single-end and paired-end mode. It also provides a check box to show/hide PCR and optical duplicated reads.

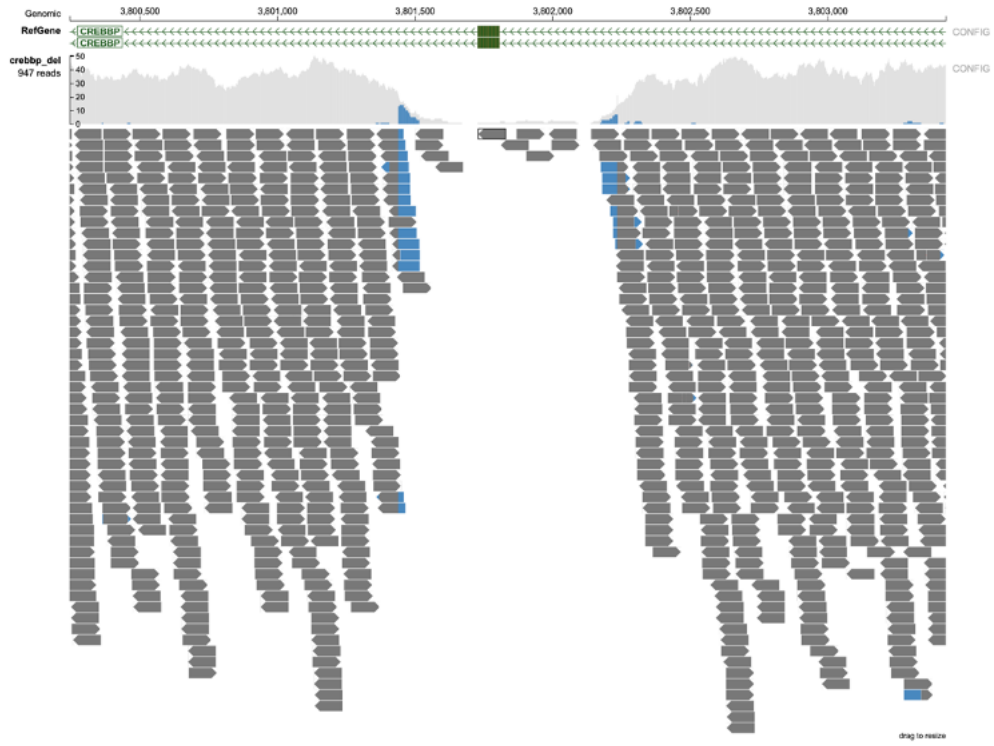
BAM track configuration panel figure



Single and Paired-end Read

The configuration panel (above) provides a toggle to change view between single-end (default) and paired-end view (shown below: see Link for example). In single-end display each read is displayed individually without displaying any connections with its respective mate. In case of the paired-end display the two paired reads are joined by a gray dotted-line if the coordinates of the two reads do not overlap. When the coordinates of the two read-pairs overlap, the overlapped region is highlighted by a blue line.

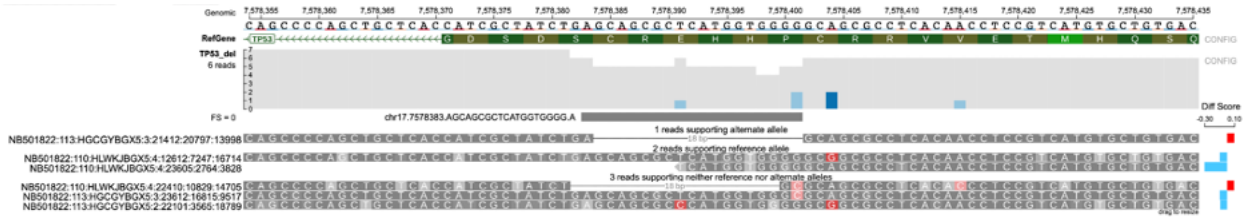
The following shows reads in single-end mode.



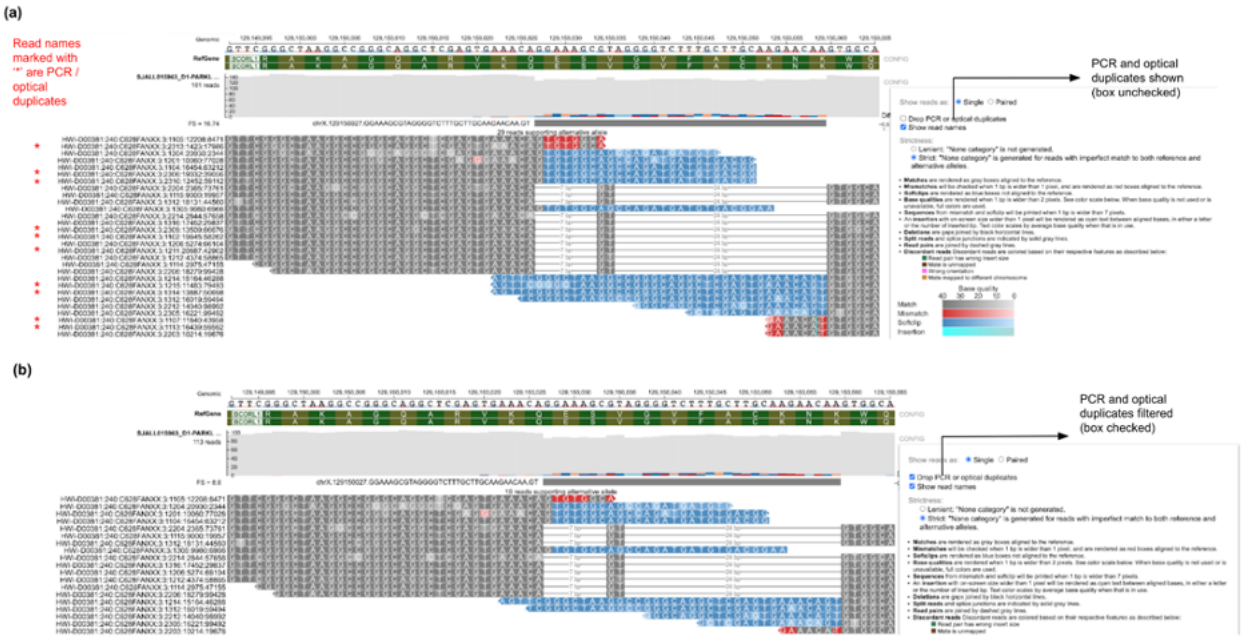
The same track above shows in paired mode.

Show/Hide Read Names

Read names are available only when the variant field is specified. There is a checkbox that displays read names on the left side of the main BAM track as shown below. The read names are only displayed when the main BAM track has base-pair level resolution and is in nucleotide view (vertical zoom in case of high-depth sequencing data).



Displaying PCR and Optical Duplicated Reads



The checkbox in the configuration panel can be toggled to switch on/off the display of PCR and optical duplicates. In the above figure, a total of 29 reads are shown when PCR/optical duplicates are displayed (Figure a) whereas a total of 19 reads are displayed supporting the alternative allele when PCR/optical duplicates are not displayed (default, Figure b).

Strictness

Strictness of the on-the-fly genotyping analysis. This option is available when the BAM track is performing on-the-fly genotyping against a variant. The user can toggle between Lenient and Strict (default) mode as shown in the ppBAM configuration panel figure.

Read Information Panel

For displaying the various features of individual reads, on clicking a particular read (in nucleotide view) a new panel opens displaying the information about the selected read (as shown below).

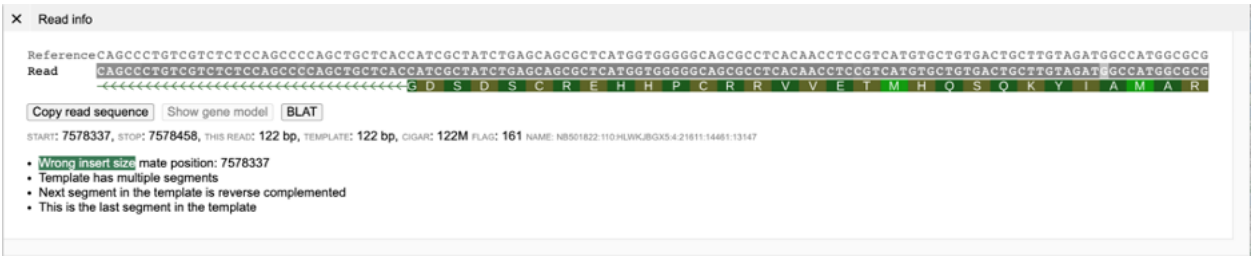
In this panel (as shown above), the top row shows the reference sequence that is aligned to the read. The second row shows the nucleotide sequence of the read. The colors of the nucleotides of the read are based on the CIGAR sequence of the read and follow the color codes as described in the section color coding of reads. In the third row, three clickable buttons are available which have the following functions as described below. The fourth row contains the start, stop, read length, template length, CIGAR sequence, flag and name of read.

Copy Read Sequence

The Copy Read Sequence feature copies the nucleotide sequence of the read being displayed to the computer clipboard so that it can be pasted outside of ppBAM.

Show Gene Models

On clicking the Show Gene Models button, the gene model (as shown below) (as described for the ProteinPaint Genome Browser figure) is displayed.



BLAT

On clicking the BLAT button, the read sequence is aligned against the given reference genome build using BLAT (as shown below).

Each of the columns obtained from BLAT alignment are explained below:

- **QScore** -Score of the BLAT alignment. Generally higher scores mean better alignment.
- **QStart** -Start position of alignment with respect to read i.e. from which nucleotide position the alignment started in the read.
- **QStop** -Stop position of alignment with respect to read i.e. from which nucleotide position the alignment stopped in the read.
- **QAlignLen** -Number of nucleotides in query sequence aligned to the reference genome.
- **RChr** -Chromosome of reference region aligned.
- **RStart** -Start position of alignment in reference genome.
- **RStop** -Stop position of alignment in reference genome.
- **RAlignLen** -Alignment length in the reference genome.

Read Details

The fourth row contains details about the read present in the BAM file

- **START** -Contains the start position of the read.
- **STOP** -Contains the stop position of the read.
- **This Read** -Contains length of the read.
- **TEMPLATE** -Contains length of the template of which the current read is part of.

- **CIGAR** -Contains CIGAR sequence of the read.
- **FLAG** -Contains the flag number (from BAM file) of the read.
- **NAME** -Contains the name of the read.

Color Coding of Reads

(a) Paired-end view

(b) Base-pair resolution mode showing nucleotides of each of the reads

In the figure above, structural variant deletion in the CREBBP gene is shown. Reads near the vicinity of the deletion have various colors (gray, green, brown and blue) based on their features as explained below. In the paired-end view (a) an overview of the deletion is shown. In Base-pair resolution mode (b) showing nucleotides of each of the reads there are softclipped reads starting near position chr16: 3,801,439.

Color codes in the background (as shown above) of the read describe the quality of the alignment of the read and its mate (in case of paired-end sequencing). These colors are assigned both on the basis of the CIGAR sequence (if it contains a softclip) and the flag value of both the read and its mate.

Gray

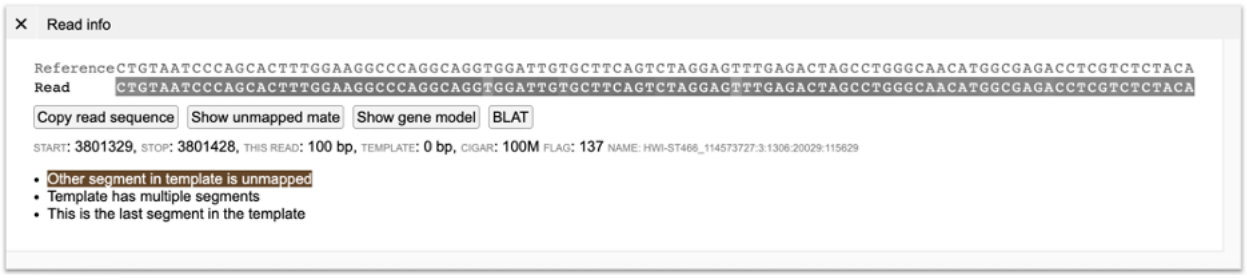
Presence of gray background nucleotides in a read suggests that both the read (at least part of it) and its mate are properly aligned and the insert size is within expected range (as shown below).

Blue

Presence of blue-background nucleotides in a read indicates that part of the read is soft clipped (as shown below). The last 94 nucleotides in the read below are softclipped based on CIGAR sequence (57M94S).

Brown

A brown colored background (in the main read alignment plot) indicates that the mate of the read is unmapped. Such reads have a flag value that contains the 0x8 bit. On clicking a read with unmapped mate in the read information panel, the current read sequence is displayed along with a button “Show unmapped mate”.



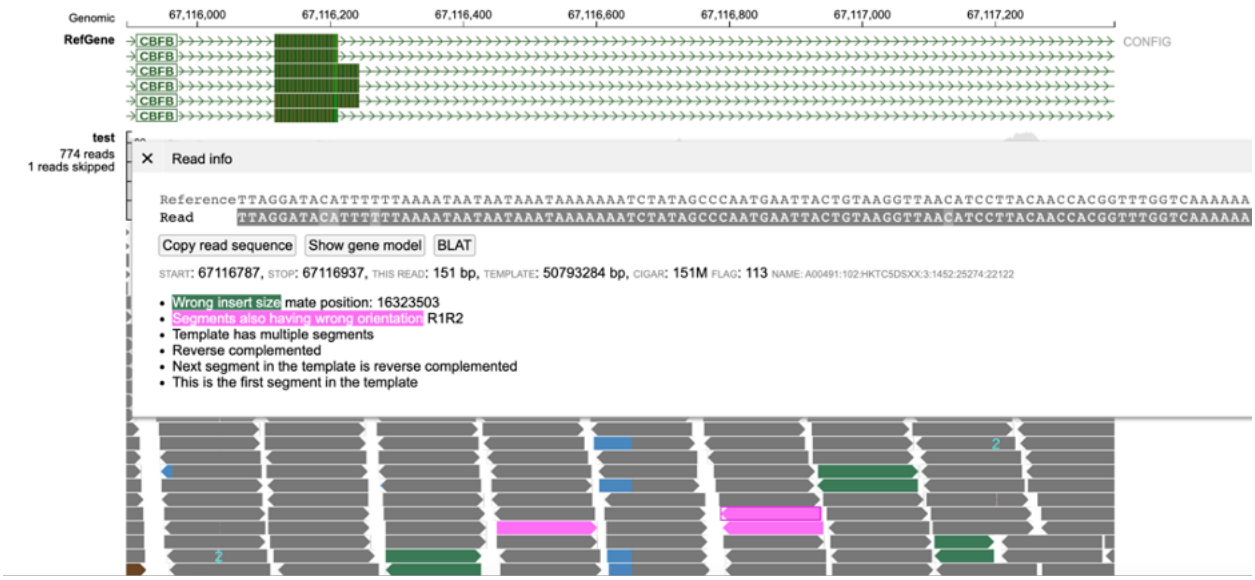
On clicking the button “Show unmapped mate”, the sequence of the unmapped mate is also displayed.

Green

A green background (shown below) indicates that the template has the wrong insert size. As shown in the Read Info figure below, the reads labeled green have a higher insert size than normal (gray) reads because of the structural deletion. In paired-end view, generally such read-pairs have a much longer gray-dashed line than properly aligned (Gray) read-pairs.

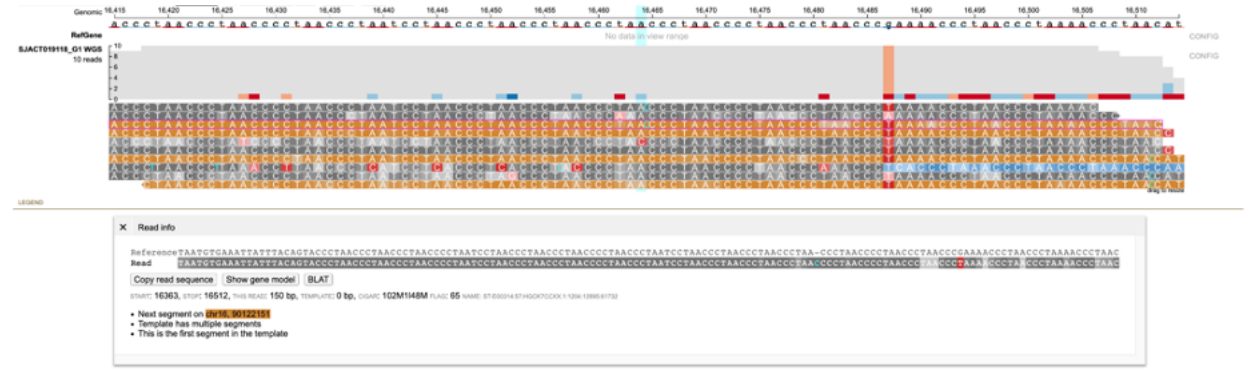
Pink

A pink color background indicates that the orientation of the read and its mate is not correct (See Link to wrong orientation example). Several orientations are taken into consideration. The figure below displays an example of an inversion caused due to CBFB-MYH11 gene fusion found in Acute Myeloid Leukemia (AML) patients. Here the read and its mate are oriented in the reverse direction (R1R2). * F1F2 - When both read and its mate are pointing in the forward direction (-> ->). * R1R2 - When both read and its mate are pointing in the reverse direction (<- <-). * F1R2 - When both read and its mate are pointing in forward and reverse direction but are pointing in opposite directions (<- ->).



Orange

Orange background color indicates the read and its mate are mapped in different chromosomes (as shown below). The displayed read is mapped in chr7:16363-16512 whereas its mate is mapped in chr16.



Variant Mode

Variant Mode provides an intuitive view of a variant specified by the user inside ppBAM. On specifying the chromosome, position, reference and alternative allele; the reads covering the variant region are displayed and classified into groups supporting the reference allele, alternative allele, none (neither reference nor alternative allele) and ambiguous groups. This mode is invoked when the "variant" field is specified containing the chromosome, position, reference and alternative allele of the variant.

Alternative, Reference, None and Ambiguous Read Classification Groups

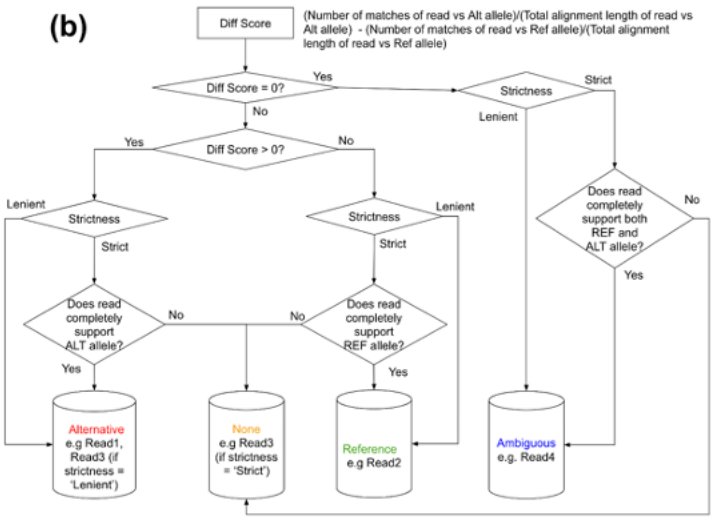
For a given variant (SNV or indel), reads mapping to the variant region are classified into Reference, Alternative, None (neither reference nor alternative allele) and Ambiguous (unclassified reads) groups by using the Smith-Waterman alignment (as shown in figure above). The difference (Diff score) between the ratio of sequence similarities (Number of matches in read alignment / Length of alignment) of the read with alternative sequence and with that of the reference sequence is used to classify the read as alternative (Diff score > 0) or reference supporting read (Diff score < 0). Reads with alleles which are neither reference nor alternative are classified into none group (when strictness level = 'Strict') and those that have equal similarity (Diff score $= 0$) to both reference and alternative allele are classified into the ambiguous group. This Diff score barplot on the right displays the Diff score (red if Diff score > 0 and blue if Diff score < 0) for each read. This barplot is especially helpful in analyzing reads classified into the none group by indicating the alternative/reference allele with which it has maximum sequence similarity.

(a)

Example variant: G > GTCA (insertion)
Example alternative sequence: ATCGTCATGC
Example reference sequence: ATCGTGC

Example reads	Alignment on alternative sequence	Alignment on reference sequence	Diff score	Group assignment
Read1 CGTCATG	ATCGTCATGC ||||||| CGTCATG	ATCG---TGC || || CGTCATG	Alternative: 7/7=1 Reference: 4/7=0.57 Diff score: 1-0.57=0.43	Alternative allele
Read2 TCGTG	ATCGTCATGC ||| || TCG---TG	ATCGTGC ||||| TCGTG	Alternative: 5/8=0.625 Reference: 5/5 =1 Diff score: 0.625-1=-0.375	Reference allele
Read3 TCGCCAT	ATCGTCATGC |||*||| TCGCCAT	ATCG---TGC ||| | TCGCCAT	Alternative: 6/7=0.857 Reference: 4/7=0.57 Diff Score: 0.857-0.57=0.287	Alternative allele
Read4 ATCGT	ATCGTCATGC ||||| ATCGT	ATCGTGC ||||| ATCGT	Alternative: 5/5=1 Reference: 5/5=1 Diff Score: 1-1=0	Ambiguous group

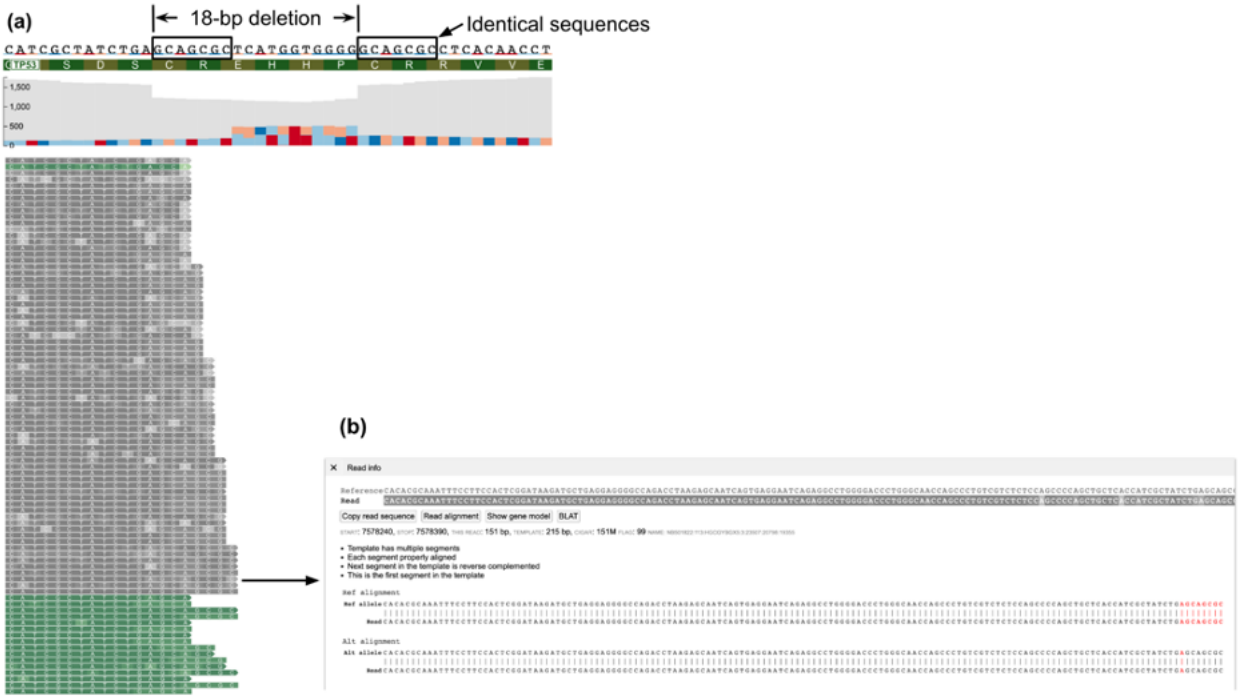
(b)



The above figure describes the methodology for classification of reads into Reference, Alternative, None (neither reference nor alternative allele) and Ambiguous groups. Classification of four reads are described for a variant (G/GTCA). (a) Sequence alignment of various reads to alternative and reference alleles (red colored nucleotides represent alternative and reference allele nucleotides): Read1 and Read2 completely support the alternative and reference allele respectively. Read3 has higher sequence similarity to the alternative allele but has a mismatch. Read4 has equal similarity to both reference and alternative groups. (b) Flow chart for classification of reads into four groups: Read1 and Read2 completely support alternative and reference allele respectively and are classified into these groups irrespective of strictness. Read3 contains a mismatch and is classified into none group when strictness = 'Strict' and into alternative group when strictness = 'Lenient'. Read4 has equal similarity (Diff score = 0) towards both alternative as well as reference sequences and is classified into the ambiguous group.

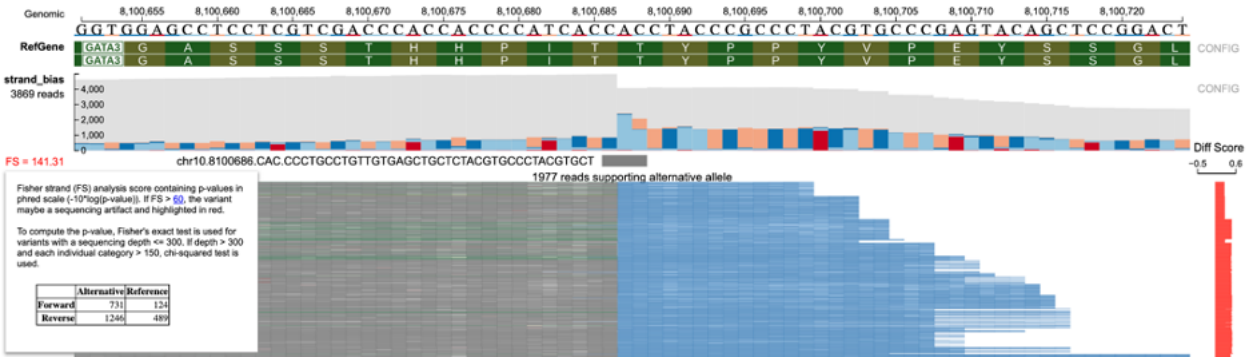
Ambiguous Reads

In certain indels such as in the TP53 example, certain reads in the variant region can have equal similarity towards both the reference and alternative allele. A large number of ambiguous reads are on the left-side of the indel (Figure a below) because the deletion starts with the sequence GCAGCGC which is also found in the right flanking sequence resulting in equal similarity to both reference and alternative alleles for any read ending within this part of the indel region (as shown in figure below). On viewing the read alignment for the ambiguous read (Figure b below) through the read information panel, it is observed that the read has equal similarity to both reference and alternative alleles. Nucleotides highlighted in red indicate those which are part of reference/alternative allele.



Fisher-strand Analysis to Check for Strand Bias in Variants

Fisher-strand (FS) analysis on ratio of forward/reverse strand reads in the alternative and reference groups can help in detecting possible strand bias that may be present in the variant of interest. The FS score is the Phred-scaled p-value from the Fisher test of the contingency table consisting of forward/reverse strand reads from both the alternative and reference alleles (as shown in figure below). To increase performance for high-depth sequencing examples, when the sequencing depth is greater than 300 the chi-square test is used (if equal or lower than this number, the fisher's exact test is used). If FS score is greater than 60, the FS score is highlighted in red (as shown below) indicating that there may be a possible strand bias in the variant.



In the figure above, an example of a complex indel is shown containing fisher strand bias. The FS score is highlighted in red indicating this particular variant may contain strand bias.

Strictness in On-the-fly Genotyping

A user can also optionally change the strictness of the algorithm to Lenient/Strict (default) from the ppBAM configuration panel. For strictness level = 'Lenient', reads are classified based on higher sequence similarity to reference/alternative allele. In case of strictness level = 'Strict', the exact sequence of the reference/alternative allele in the read is compared against the allele sequence given by the user. Reads that do not match either allele are classified into the none group.

The lenient strictness level can be helpful, when the user wants a lenient estimate of the number of reads supporting the particular indel of interest or when the user is confident that only one alternative allele exists. This can also be helpful when there are reads with low base-pair quality calls near the variant region. In contrast when the strictness level is set to 'Strict', a more conservative

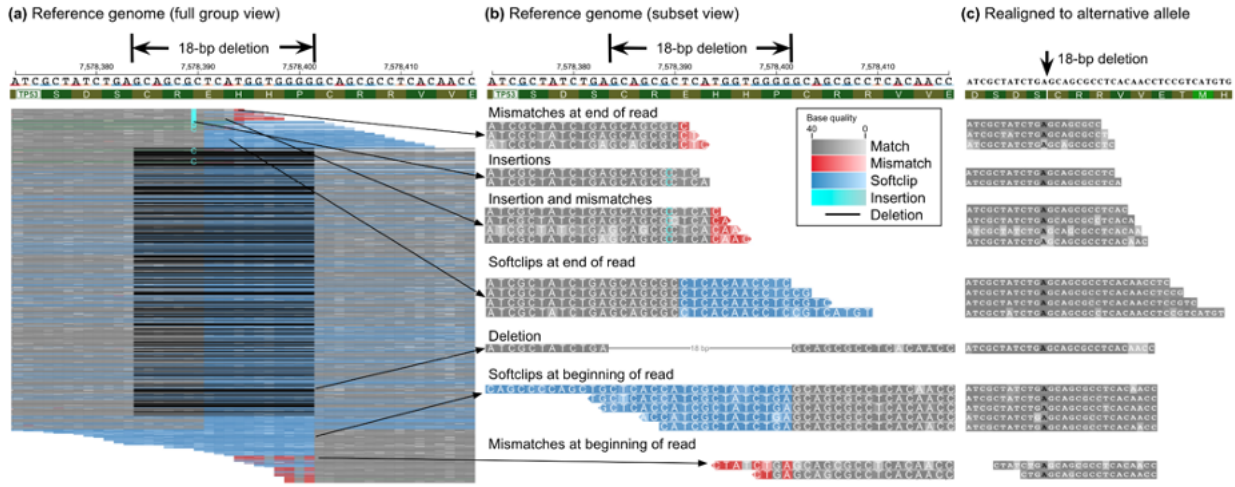
estimate of the read support is provided for each allele and may indicate the presence of a wrong variant call (if present) or may indicate presence of multiple alternative alleles.

In case of the TP53 deletion example, select reads with wrong base pair calls are shown. For strictness level = 'Lenient', there are two reads that support the alternative allele. However, read NB501822:110:HLWKJBGX5:4:22410:10829:14705 has a wrong base pair call at position 7578401. When the strictness level is changed to 'Strict', this read is classified into the none group. Similarly, reads NB501822:113:HGCGYBGX5:3:23612:16815:9517 (wrong base-pair call at 7578401) and NB501822:113:HGCGYBGX5:2:22101:3565:18789 (wrong base-pair call at 7578391) are classified in the reference allele group when strictness level = 'Lenient' but are classified into the none group when strictness level is set to 'Strict'.

The 'Lenient' strictness level is generally only helpful in cases where only one alternative allele is present as it assumes only the given reference and alternative allele are the only possible cases. For multi-allelic variants or when a region has a large number of reads with low Phred base-pair quality nucleotides, the 'Strict' (default) level should be used.

Realignment using Clustal Omega

In the original alignment shown in the main BAM track view, all the reads are aligned against the reference genome. Therefore, in the alternative allele group reads may be mapped differently although they have the same sequence in the variant region. For example, in the reads supporting the alternative allele in the TP53 example , the reads either have mismatches, deletions, soft clips or a combination of all three. Figure (a) shows the complete alternative allele group, whereas in Figure (b) selected set of reads from alternative allele group are displayed displaying various kinds of mapping inconsistencies.



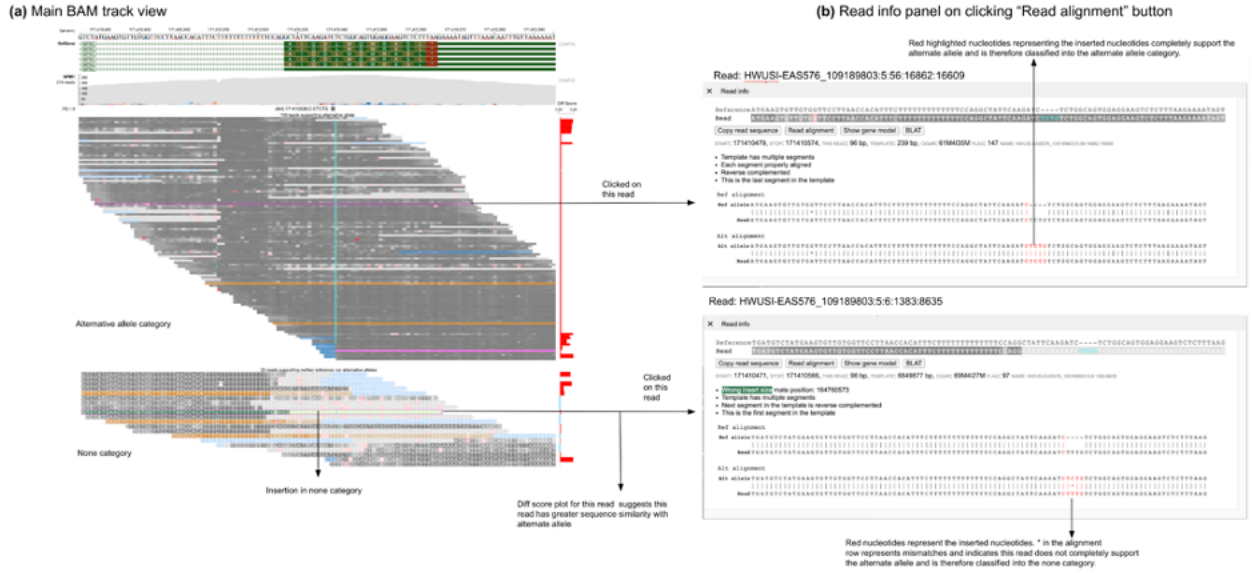
In Figure (c), the reads from (b) are realigned to the alternative allele using Clustal Omega (ClustalO) by clicking on the link showing the number of reads aligned to the alternative allele. This provides an intuitive view confirming the accuracy of the classification of reads to the designated allele. See subset of different reads with same sequence near variant region mapped differently.

Display of Read Alignment with Respect to Reference and Alternative Allele

In case of reads that are classified into the none group (when strictness level = 'Strict') it can be difficult to understand the classification into that group. For example, in case of insertions with the wrong nucleotide (with respect to the predicted alternative allele) the sequence of the inserted nucleotides is not shown in the main BAM track and can only be viewed through the read information panel. As an example, a 4bp insertion in NPM1 exon is discussed below. In Figure (a) (shown below) most reads with 4bp insertion have been classified into the alternative allele. However, there are some reads (as highlighted in Figure a) with 4bp insertions that are classified into the none group. The diff score plot suggests that these reads have higher sequence similarity to the alternative allele (and are classified into the alternative group when strictness = 'Lenient') and they seem to support the alternative allele. However, when we click on this read (Fig. b) and click on the "Read Alignment" button (which is available only when the variant field is specified in the URL) the Smith-Waterman alignment of the read with the reference and alternative allele is displayed (Figure b). The indel nucleotides are highlighted in red. In case of the read in the none group (HWUSI-EAS576_109189803:5:56:16862:16609), a mismatch is observed in the indel region between the read and the alternative allele (highlighted by '*' in the alignment row) which explains the classification into the none group. In contrast, the read shown

from the alternative allele group (HWUSI-EAS576_109189803:5:6:1383:8635) has a complete match with the alternative allele and is therefore classified into the alternative allele group.

Display of read alignment of the read with respect to both the reference and alternative allele helps provide an intuitive view for describing classification of a read into its respective group.

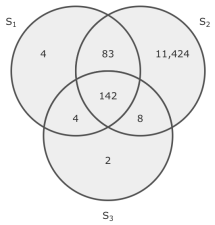


Chapter 15

Set Operations

Set Operations

Up to three cohorts, gene sets, and mutation sets can be compared and exported based on complex overlapping subsets. The features of this page include:



Alias	Entity Type	Name	# Items
S1	Cohort	Chemotherapy treatment	233
S2	Cohort	Radiation treatment	11,657
S3	Cohort	Surgery treatment	156

Select	Set Operation	# Items	Download
☐	(S1 ∩ S2 ∩ S3)	142	
☐	(S1 ∩ S2) - (S3)	83	
☐	(S2 ∩ S3) - (S1)	8	
☐	(S1 ∩ S3) - (S2)	4	
☐	(S1) - (S2 ∪ S3)	4	
☐	(S2) - (S1 ∪ S3)	11,424	
☐	(S3) - (S1 ∪ S2)	2	
Union of selected sets:		0	

- **Venn Diagram:** Visually displays the overlapping items included within the three cohorts or sets. Subsets based on overlap can be selected by clicking one or many sections of the Venn diagram. As sections of the Venn Diagram become highlighted in blue, their corresponding row in the overlap table becomes selected.
- **Summary Table:** Displays the alias, entity type, and name for each set included in this analysis
- **Overlap Table:** Displays the number of overlapping items with set operations rather than a visual diagram. Subsets can be selected by checking boxes in the "Select" column, which will highlight the corresponding section of the Venn Diagram. As rows are selected, the "Union of selected sets" row is populated. Each row has an option to create a new set or cohort from the subset, or export the subset as a TSV.

Chapter 16

For Developers

GDC Portal 2.0 Application Developer Guide

Introduction

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

The GDC Portal is designed to support the development of applications that allow for analysis, visualization, and refinement of cohorts. The GDC Portal is built on top of the GDC API, which provides access to the GDC data. The GDC Portal provides a framework for developing applications that can be used to analyze and visualize data from the GDC. The GDC Portal is built on top of the React framework and uses the Redux library for state management. The GDC Portal uses NextJS to provide server-side rendering of React components. Mantine.dev is the base component library used and styling is done with TailwindCSS.

The GDC Portal contains an Analysis Center where applications are displayed for users to use with their cohorts. The GDC Portal also provides a framework for developing applications that can be used to analyze and visualize data from the GDC.

Overview of an Application

Applications are High Order Components (HOC) that are rendered in the Analysis Center. The portal major functions like Project, Downloads, and Protein Paint are all applications. Each application handles a specific task and can be used to refine and analyze cohorts. Applications have access to all the current cohort information and can use that information to query the GDC API for additional information.

Local and Cohort filters are available to applications. Local filters are filters that are specific to the application and are used to refine the data that is displayed in the application. Local filters are those available from the GDC API and are typically not the most common. For example in the Mutation Frequency application, the local filters are the gene and mutation type filters. In the figure below the local filters are highlighted in yellow. These filters are used to refine the input cohort allowing users to drill down to specific genes and mutation types of interest in the cohort.

Local vs Cohort Filters

A Portal application's input can be anything including a single cohort or multiple cohorts. The application then can either add filters to refine the cohort by adding filters, create additional cohorts, or display the data in a visualization. Applications typically have:

- Local filters that are used to refine the data displayed in the application
- Cohort filters that are used to refine the cohort
- UI components that are used to display the data in the application
- State that is used to store the data displayed in the application
- Actions that are used to update the state of the application

Applications can also create new cohorts. These cohorts can be used by other Portal applications.

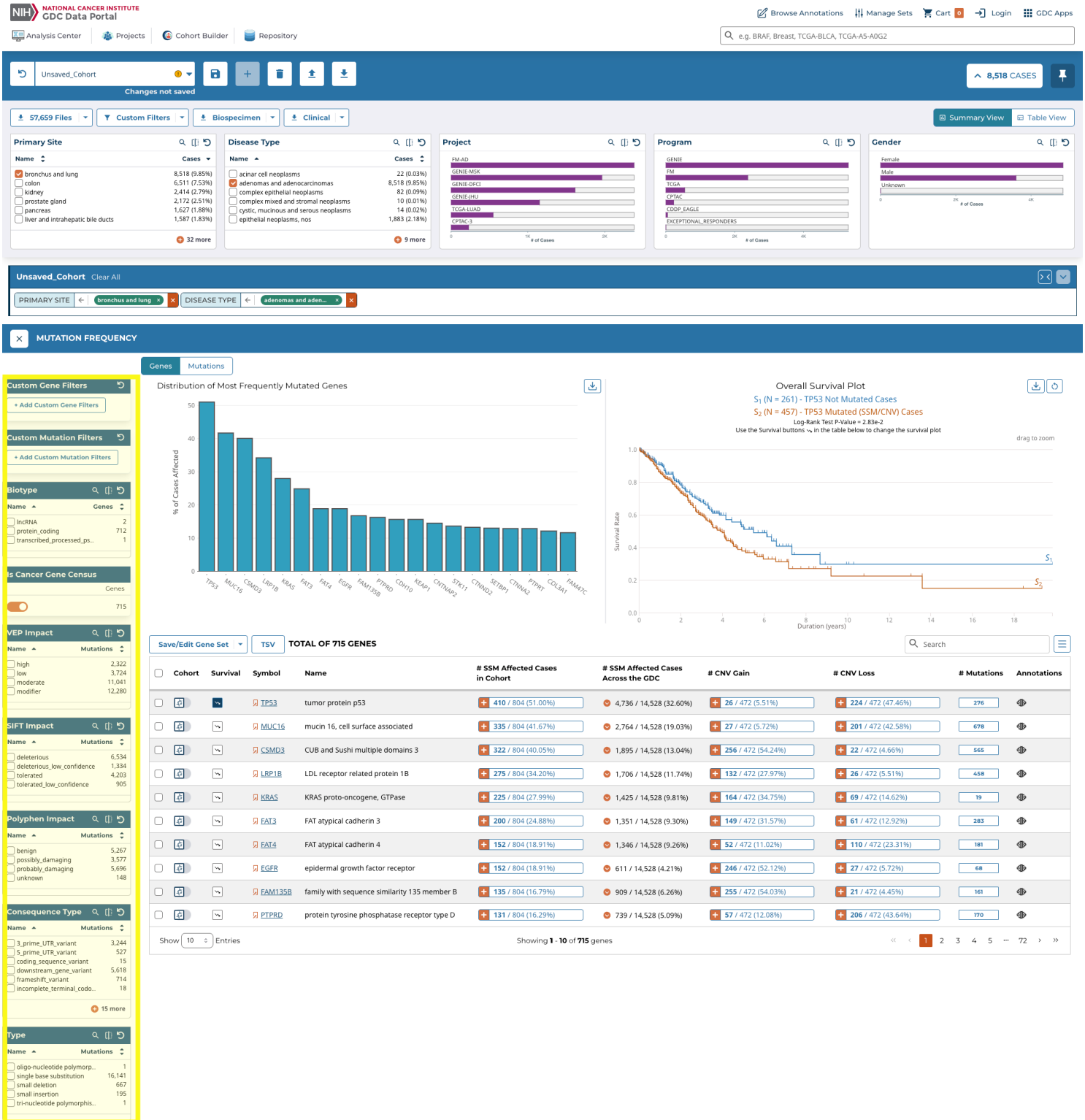


Figure 16.1: Mutation Frequency

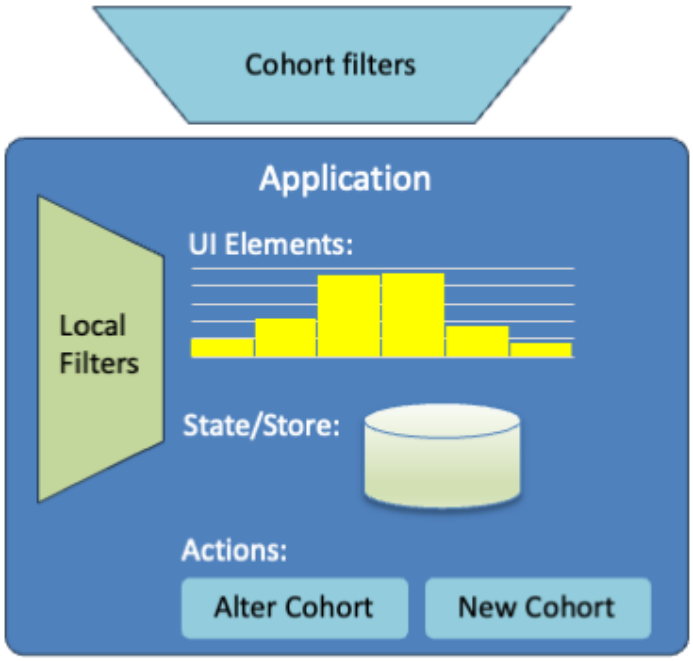


Figure 16.2: Structure of an Application

Cohorts and Filters

From an application perspective, a cohort is an Object containing the following information:

```
interface Cohort {
  id: string;        // unique id for cohort
  name: string;      // name of cohort
  filters: FilterSet; // active filters for cohort
  caseSet: CaseSetDataAndStatus; // case ids for frozen cohorts
  modified?: boolean; // flag which is set to true is modified and unsaved
  modified_datetime: string; // last time cohort was modified
  saved?: boolean; // flag indicating if cohort has been saved.
  counts: CountsDataAndStatus; //case, file, etc. counts of a cohort
}
```

Likely the most important part of the cohort is the `filters` field. The `filters` field contains the active filters for the cohort. The `filters` field is a `FilterSet` object. The `FilterSet` object contains the active filters for the cohort. When calling either the GDC API or GDC GraphQL API the `FilterSet`` is converted to the appropriate format for the API. The`FilterSet‘ object is of the form:

```
interface FilterSet {
  op: "and" | "or"; // operator for combining filters
  root: Record<string,Operation >; // map of filter name to filter operation
}
```

Operation are GDC API filters described in the GDC API Guide. These are: * Equals * NotEquals * LessThan * LessThanOrEquals * GreaterThan * GreaterThanOrEquals * Exists * Missing * Includes * Excludes * ExcludeIfAny * Intersection * Union

The `root` field is a map of filter names (as defined in the GDC API) to filter operation. The filter operation can be either a single operation or a `FilterSet` object. The `op` field will eventually support either `and` or `or`, however at this time only `and` is supported. The `and`operator is used to combine filters using the `and` operator. The `or` operator is used to combine filters using the `or` operator. The `FilterSet` object is converted to the appropriate format for the GDC API when the cohort is saved. When using the GDC REST API, the FilterSet can be converted into the appropriate format using the `filterSetToOperation`

function. When using the GDC GraphQL API, the FilterSet can be using the `convertFilterSetToGraphQL` function. The API guide will provide information on what format the filters should be in for the API. Also as the code is in TypeScript, the IDE will provide information on the format as well.

Getting Cohort Information

The current active cohort can be accessed via the selector `selectCurrentCohort`. This selector returns the current cohort, which is the cohort that is currently being displayed in the Cohort Management Bar. Accessing the current cohort is done via the selector:

```
import {useCoreSelector,  selectCurrentCohort } from '@gff/core';

const currentCohort = useSelector(selectCurrentCohort);
```

By using the selector, the component/application will be updated when the cohort changes. There are also selectors for getting a particular field from the cohort. For example, to get the cohort name, the selector `selectCurrentCohortName` can be used. The selectors are:

- `selectCurrentCohort`
- `selectCurrentCohortName`
- `selectCurrentCohortId`
- `selectCurrentCohortFilters`
- `selectCurrentCohortModified`
- `selectCurrentCohortModifiedDatetime`
- `selectCurrentCohortSaved`
- `selectCurrentCohortCounts`

The current active filters can be accessed via the selector `selectCurrentCohortFilters`. This selector returns the current filters, which are the filters that are currently being displayed in the Cohort Management Bar. Accessing the current filters is done via the selector:

```
import {useCoreSelector,  selectCurrentFilters } from '@gff/core';

const currentFilters = useSelector(selectCurrentCohortFilters);
```

By using the selector, the application will be updated when the filters change. The filters are returned as a `FilterSet` object described above.

All the cohorts can be selected using the selector `selectAllCohorts`. This selector returns all the cohorts in the store. Accessing all the cohorts is done via the selector:

```
import {useCoreSelector,  selectAllCohorts } from '@gff/core';

const allCohorts = useSelector(selectAllCohorts);
```

Using the Portal Application API

The GDC Portal provides a number of hooks for querying the GDC API. These hooks are located in the `@gff/core` package. The hooks are designed to work in a manner similar to the RTL Query hooks. The hooks take arguments and return an object. The object contains the data and the status of the query. The status of the query is stored in the `isSuccess` variable. The @gff/core package also provides a set of selectors that return values stored in the core redux store: `CoreStore`.

There are a number of hooks and selectors that are available for querying the GDC API, a subset of which are shown below:

Case Information

The GDC Portal provides several hooks for querying case information. These hooks are located in the `@gff/core` package. Cases can be queried using several different methods. The 'useAllCases' hook returns all the cases in the GDC and can be filtered by the current cohort as shown below:

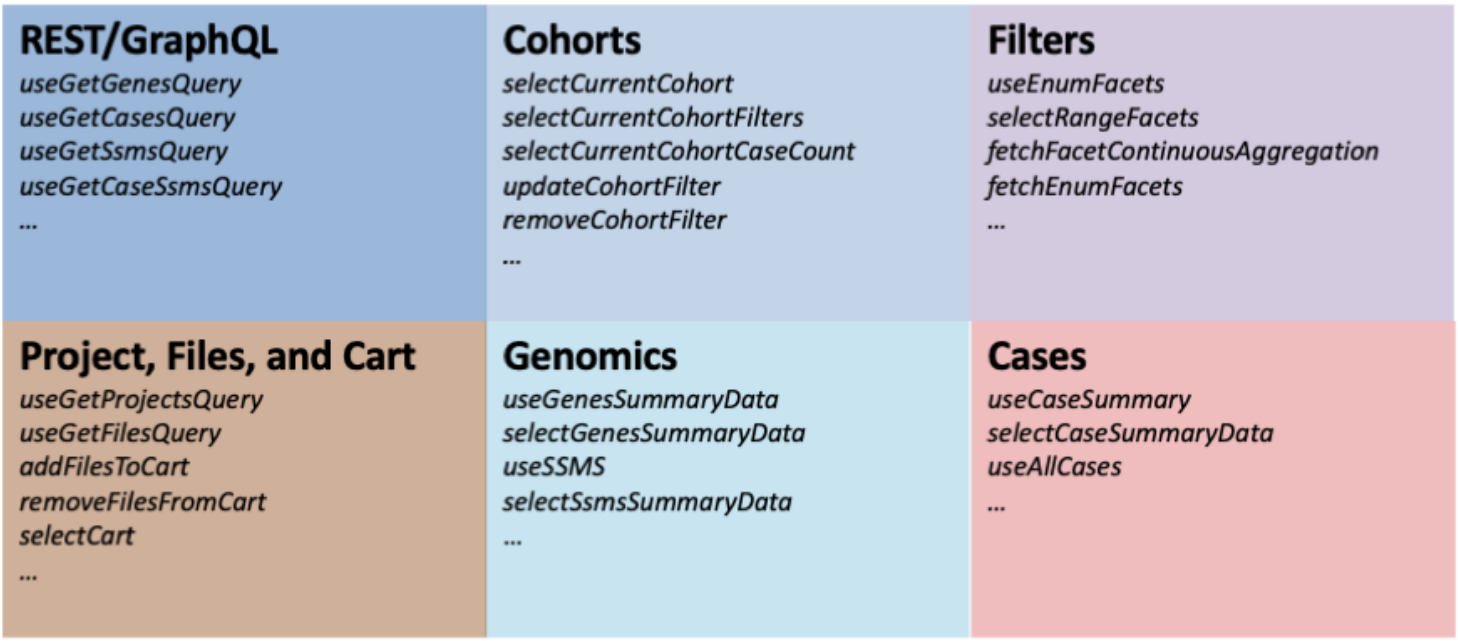


Figure 16.3: hooks and selectors

```
import { useCoreSelector, useAllCases } from '@gff/core';

...

const [pageSize, setPageSize] = useState(10);
const [offset, setOffset] = useState(0);
const [searchTerm, setSearchTerm] = useState<string>("");
const [sortBy, setSortBy] = useState<SortBy[]>([]);
const cohortFilters = useCoreSelector((state) =>
        selectCurrentCohortFilters(state),
);


const { data, isFetching, isSuccess, isError, pagination } = useAllCases({
  fields: [
    "case_id",
    "submitter_id",
    "primary_site",
    "disease_type",
    "project.project_id",
    "project.program.name",
    "demographic.gender",
    "demographic.race",
    "demographic.ethnicity",
    "demographic.days_to_death",
    "demographic.vital_status",
    "diagnoses.primary_diagnosis",
    "diagnoses.age_at_diagnosis",
    "summary.file_count",
    "summary.data_categories.data_category",
    "summary.data_categories.file_count",
    "summary.experimental_strategies.experimental_strategy",
    "summary.experimental_strategies.file_count",
```

```
    "files.file_id",
    "files.access",
    "files.acl",
    "files.file_name",
    "files.file_size",
    "files.state",
    "files.data_type",
  ],
  size: pageSize,
  filters: cohortFilters,
  from: offset * pageSize,
  sortBy: sortBy,
  searchTerm,
});
```

The `useAllCases` hook takes a number of arguments: * `fields` - the fields to return from the GDC API * `size` - the number of cases to return * `filters` - the filters to apply to the cases * `from` - the starting index of the cases to return * `sortBy` - the fields to sort the cases by * `searchTerm` - the search term to use to search the cases

This call is used in the Table view tab of the Cohort Management Bar.

Information for a single case can be queried using the `useCaseSummary` hook. This call is used in the caseView page: portal.gdc.cancer.gov/cases/5693302a-4548-4c0b-8725-0cb7c67bc4f8

```

  const { data, isFetching } = useCaseSummary({
  filters: {
    content: {
      field: "case_id",
      value: case_id,
    },
    op: "=",
  },
  fields: [
    "files.access",
    "files.acl",
    "files.data_type",
    "files.file_name",
    "files.file_size",
    "files.file_id",
    "files.data_format",
    "files.state",
    "files.created_datetime",
    "files.updated_datetime",
    "files.submitter_id",
    "files.data_category",
    "files.type",
    "files.md5sum",
    "case_id",
    "submitter_id",
    "project.name",
    "disease_type",
    "project.project_id",
    "primary_site",
    "project.program.name",
    "summary.file_count",
    "summary.data_categories.file_count",
    "summary.data_categories.data_category",
    "summary.experimental_strategies.experimental_strategy",
```

```
    "summary.experimental_strategies.file_count",
    "demographic.ethnicity",
    "demographic.demographic_id",
    "demographic.gender",
    "demographic.race",
    "demographic.submitter_id",
    "demographic.days_to_birth",
    "demographic.days_to_death",
    "demographic.vital_status",
    "diagnoses.submitter_id",
    "diagnoses.diagnosis_id",
    "diagnoses.classification_of_tumor",
    "diagnoses.age_at_diagnosis",
    "diagnoses.days_to_last_follow_up",
    "diagnoses.days_to_last_known_disease_status",
    "diagnoses.days_to_recurrence",
    "diagnoses.last_known_disease_status"]
  });
```

The `useCaseSummary` hook takes a number of arguments: * `fields` - the fields to return from the GDC API * `filters` - the filters to apply to the cases and where the caseId is passed in

File Information

Similar to the case information, the GDC Portal provides a number of hooks for querying file information. These hooks are located in the `@gff/core` package. To get a list of files associated with a cohort, the `useGetFilesQuery` hook can be used. This call is used in the repository application. The repository application is used to display the files associated with a cohort. The `useGetFilesQuery` hook takes a number of arguments:

```
import {
  useCoreDispatch,
  useCoreSelector,
  selectCurrentCohortFilters,
  buildCohortGqlOperator,
  joinFilters,
  useFilesSize,
} from "@gff/core";

...

const coreDispatch = useCoreDispatch();
const [sortBy, setSortBy] = useState<SortBy[]>([]); // states to handle table sorting and pagination
const [pageSize, setPageSize] = useState(20);
const [offset, setOffset] = useState(0);

const repositoryFilters = useAppSelector((state) => selectFilters(state)); // as this is a app get the repository filters from the app state (local filters)
const cohortFilters = useCoreSelector((state) =>    // get the cohort filters from the core state (global filters)
    selectCurrentCohortFilters(state),
);

const { data, isFetching, isError, isSuccess } = useGetFilesQuery({
  case_filters: buildCohortGqlOperator(cohortFilters),
  filters: buildCohortGqlOperator(repositoryFilters),
  expand: [
    "annotations", //annotations
    "cases.project", //project_id
```

```
    "cases",
  ],
  size: pageSize,
  from: offset * pageSize,
  sortBy: sortBy,
});
```

The `useGetFilesQuery` hook takes a number of arguments: * `case_filters` - the filters to apply to the cases * `filters` - the filters to apply to the files * `expand` - the fields to expand * `size` - the number of files to return * `from` - the starting index of the files to return * `sortBy` - the fields to sort the files by

Note this hook was designed to take global filters (e.x the current cohort as `case_filters`) and local filters (the repository filters).

Information for a single file can be queried using the `useFileSummary` hook. This call is used in the `File Summary View` page portal.gdc.cancer.gov/files/0b5a9e7e-8e2e-4b7a-9b7e-ff5d9c5b2b2b

```
const { data: { files } = {}, isFetching } = useGetFilesQuery({
  filters: {
    op: "=",
    content: {
      field: "file_id",
      value: setCurrentFile,
    },
  },
  expand: [
    "cases",
    "cases.annotations",
    "cases.project",
    "cases.samples",
    "cases.samples.portions",
    "cases.samples.portions.analytes",
    "cases.samples.portions.slides",
    "cases.samples.portions.analytes.aliquots",
    "associated_entities",
    "analysis",
    "analysis.input_files",
    "analysis.metadata.read_groups",
    "downstream_analyses",
    "downstream_analyses.output_files",
    "index_files",
  ],
});
```

The `useFileSummary` hook takes several arguments: * `filters` - the filters to apply to the cases and where the file uuid is passed in * `expand` - the fields to expand

Sets: Gene, SSMS, and Case

Sets are supported by the GDC API and are used to create an entity that represents a set of items as a `set_id`. Sets can be gene sets, SSM sets, or case sets. All GDC APIs support passing sets as a filter parameter. The GDC Portal provides a number of hooks for creating and querying set information.

a set can be created using one of the following hooks:

- `useCreateGeneSetFromValuesMutation`
- `useCreateSsmsSetFromValuesMutation`
- `useCreateCaseSetFromValuesMutation`
- `useCreateGeneSetFromFiltersMutation`
- `useCreateSsmsSetFromFiltersMutation`

- `useCreateCaseSetFromFiltersMutation`

These functions will create a set from either a list of values or a filter set. The create from Values hooks take a single parameter `values` which is an array of values, while the create from filters hooks take one required parameter `filters` which is a filter set or JSON object. Both calls return the created `set_id` if the set was successfully created.

As the above hooks are Redux Toolkit Query hooks, namely mutation hooks, they return a tuple of the form: `[mutationHook, response]` which is a function to call the mutation and the response from the mutation. The mutation hook can be used like:

```
const [createSet, response] = createSetHook();

 const handleCreateSet = async () => {
   const { data } = await createSet({
     variables: {
       values: ["TP53", "KRAS", "EGFR"],
     },
   });
   if (response.isSuccess) {
     dispatch(
       addSet({
         setType,
         setName: form.values.name.trim(),
         setId: response.data as string,
       }),
     );
   }
   ;
 }
```

Once a set is created it can be altered using the following hooks: * `useAppendToGeneSetMutation` * `useAppendToSsmSetMutation` * `useRemoveFromGeneSetMutation` * `useRemoveFromSsmSetMutation`

Sets can be managed using the following actions:

- `addSet`
- `removeSet`
- `updateSet`

The following selectors are available for getting set information:

- `selectAllSets`
- `selectSetById`
- `selectSetByName`
- `selectSetByType`

Finally, the following hooks are available for querying set size:

- `useGeneSetCountsQuery`
- `useSsmSetCountsQuery`
- `useCaseSetCountsQuery`

Creating a cohort

Depending on what your application does, you may want to create a new cohort. Although the GDC Portal SDK provides a number of functions for creating a new cohort. It is highly recommended that the application use the provided Button and SaveCohortModal components to create a new cohort. The Button and SaveCohortModal components are located in the `@gff/portal-proto` package.

To create a cohort using the SaveCohortModal component the following code can be used: In summary, the above code flow is:

1. The ProjectsCohortButton component renders a button with the label “Save New Cohort”.

2. When the button is clicked, it sets the state variable `showSaveCohort` to true, which triggers the rendering of the `SaveCohortModal` component.
3. The `SaveCohortModal` component passed:
   - an onClose function that sets the showSaveCohort state variable to false.
   - a filters prop, which is an object defining the filters for the cohort based on the selected projects.
4. The `SaveCohortModal` will use the passed filter to create, name and save the cohort when the save button is clicked.

Additional details on the `SaveCohortModal` component can be found in the Component Library section as well as buttons to create a saved cohort.

Altering a cohort

Altering a cohort is done by dispatching actions to add, remove, or clear filters. The following actions are available for altering the current cohort:

- `updateCohortFilter`
- `removeCohortFilter`
- `clearCohortFilters`

Note that all of these operations are applied to the current cohort. The current cohort is the cohort that is currently being displayed in the Cohort Management Bar. The current cohort can be accessed via the `selectCurrentCohort` selector. The current cohort's filters can be accessed via the `selectCurrentCohortFilters` selector.

Updating, removing, and clearing filters

to update the current selected cohort's filter, the `updateCohortFilter` action can be used. The `updateCohortFilter` action takes two arguments:

```
interface UpdateFilterParams {
  field: string;
  operation: Operation;
}
```

where `field` is the field to update and `operation` is the operation to apply to the field. For example to update the `cases.project.project_id` field to include the project `TCGA-ACC` the following code can be used:

```
import { useCoreDispatch, updateCohortFilter } from '@gff/core';

const coreDispatch = useCoreDispatch();

coreDispatch(updateCohortFilter({
  field: "cases.project.project_id",
  operation: {
    op: "in",
    content: {
      field: "cases.project.project_id",
      value: ["TCGA-ACC"],
    },
  },
}));
```

This will update the current cohort's filter to include the project `TCGA-ACC`. The `removeCohortFilter` action can be used to remove a filter from the current cohort. The `removeCohortFilter` action takes a single argument:

```

interface RemoveFilterParams {
  field: string;
}
```

where `field` is the field to remove. For example to remove the `cases.project.project_id` field from the current cohort's filter the following code can be used:

```
import { useCoreDispatch, removeCohortFilter } from '@gff/core';

const coreDispatch = useCoreDispatch();

coreDispatch(removeCohortFilter({
  field: "cases.project.project_id",
}));
```

This will remove the `cases.project.project_id` field from the current cohort's filter. The `clearCohortFilters` action can be used to clear all the filters from the current cohort. The `clearCohortFilters` action takes no arguments. For example to clear all the filters from the current cohort the following code can be used:

```
import { useCoreDispatch, clearCohortFilters } from '@gff/core';

const coreDispatch = useCoreDispatch();

coreDispatch(clearCohortFilters());
```

This will clear all the filters from the current cohort.

Updating the cohort name

The cohort name can be updated using the `updateCohortName` action. The `updateCohortName` action takes a single argument:

```

interface UpdateCohortNameParams {
  name: string;
}
```

where `name` is the new name for the cohort. For example, to update the current cohort's name to `My Cohort` the following code can be used:

```
import { useCoreDispatch, updateCohortName } from '@gff/core';

const coreDispatch = useCoreDispatch();

coreDispatch(updateCohortName({
  name: "My Cohort",
}));
```

This will update the current cohort's name to `My Cohort`.

Setting the current cohort

The current cohort can be set using the `setCurrentCohort` action. The `setCurrentCohort` action takes a single argument:

```

interface SetCurrentCohortParams {
  cohortId: string;
}
```

where `cohortId` is the id of the cohort to set as the current cohort. For example to set the cohort with id `1234` as the current cohort the following code can be used:

```
import { useCoreDispatch, setCurrentCohort } from '@gff/core';

const coreDispatch = useCoreDispatch();

coreDispatch(setCurrentCohort({
  cohortId: "1234",
}));
```

This will set the cohort with ID `1234` as the current cohort.

Count Information

Counts information can be queried using the `useTotalCounts` hook. This hook takes a number of arguments:

```
import { useTotalCounts } from "@gff/core";

const { data, isFetching, isSuccess, isError } = useTotalCounts();
```

this will return the total counts for the GDC. The data in the response is of the form:

```
interface TotalCounts {
  counts: {
    caseCounts: number;
    fileCounts: number;
    genesCounts: number;
    mutationCounts: number;
    repositoryCaseCounts: number;
    projectsCounts: number;
    primarySiteCounts: number;
  },
  status: DataStatus;
}
```

where `DataStatus` is defined as:

```
export type DataStatus = "uninitialized" | "pending" | "fulfilled" | "rejected";
```

Component Library

As a developer you will likely want to use the components provided by the GDC Portal. The GDC Portal provides a number of components that make it easy to develop applications for the GDC Portal. These components are located in the `@gff/portal-proto` package. In several components, the GDC Portal uses the Mantine component library but base components and encapsulates calls to the GDC API so that you do not have to.

Buttons

The GDC Portal provides a number of buttons that can be used for various purposes. These buttons are located in the `@gff/portal-proto` package. The buttons are:

- `DownloadButton` - a button that can be used to download data from the GDC API.
- `SaveCohortButton` - a button that can be used to save a cohort.

The `DownloadButton` component is used in the repository application to download data from the GDC API. The `DownloadButton` component takes a number of arguments:

```

<DownloadButton
        inactiveText={`Download ${numFilesCanAccess} Authorized File${
                numFilesCanAccess !== 1 ? "s" : ""
        }`}
        activeText=""
        disabled={
                numFilesCanAccess === 0 ||
                (user.username && dbGapList.length > 0 && !checked)
        }
        endpoint="data"
        extraParams={{
          ids: (filesByCanAccess?.true || []).map((file) => file.file_id),
          annotations: true,
          related_files: true,
```

```
        }}
        method="POST"
        setActive={setActive}
/>
```

The parameters for the DownloadButton are defined in the Portal 2.0 SDK API documentation. The `DownloadButton` component will take care of calling the GDC API and downloading the data. The `DownloadButton` component will also provide status that can be used with a progress bar or spinner to display the progress of the download.

The `SaveCohortButton` component is used in the repository application to save a cohort.

Figure 16.4: img.png

The `SaveCohortButton` component takes a number of arguments:

```

<CohortCreationButton
        numCases={cohort1Count}
        label={cohort1Count.toLocaleString()}
        filtersCallback={async () =>
                generateFilters(caseSetIds[0], caseSetIds[1])
        }
/>
```

The CohortCreationButton component will show the number of selected cases and will create a new saved cohort when the button is clicked. The `filtersCallback` is a function that returns the filters for the cohort.

Modals

Modals are used to show transitory information or obtain information from the user. The GDC Portal provides many modals that can be used for various purposes. One such modal is the `SaveCohortModal` component mentioned previously.

- `SaveCohortModal` - a modal that can be used to save a cohort.
- Various modals for displaying information on Sets:
  - `CaseSetModal`
  - `GeneSetModal`
  - `MutationSetModal`
- `SaveOrCreateEntityModal` - a modal that can be used to save or create a new entity.

These modals and others, are documented in the Portal 2.0 SDK API documentation.

Charts

Basic charts are provided for use in your application, although applications are free to use any charting library they wish. The charts provided are:

- `BarChart` - a bar chart

The `BarChart` component (based on Plotly) is passed data in the form:

```
import { PlotData } from "plotly.js";


export interface BarChartData {
  datasets: Partial<PlotData>[];
  yAxisTitle?: string;
```

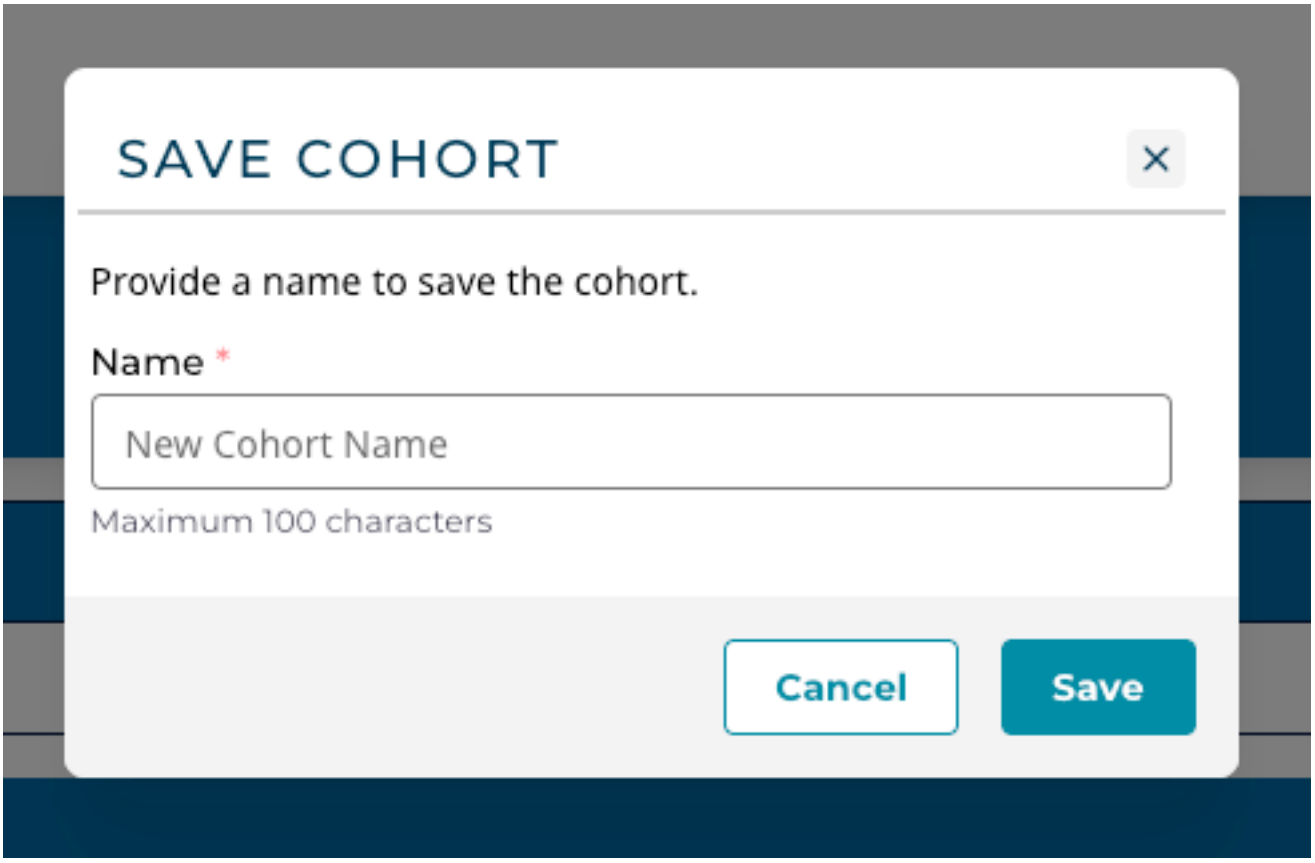


Figure 16.5: img.png

```
  tickvals?: number[];
  ticktext?: string[];
  label_text?: string[] | number[];
  title?: string;
  filename?: string;
}
```

Where `datasets` is an array of PlotData objects, which at the minimum contain the `x` and `y` fields.

The `yAxisTitle` is the title for the y-axis, the `tickvals` and `ticktext` are the tick values and text for the x-axis, the `label_text` is the text for the labels, the `title` is the title for the chart, and the `filename` is the filename to use when downloading the chart.

Note that BarChart needs to be imported as a dynamic component:

```
import dynamic from "next/dynamic";

const BarChart = dynamic(() => import("@/components/charts/BarChart"), {
  ssr: false,
});
```

- `Cancer Distribution` - a cancer distribution chart

The `CancerDistribution` component (based on Plotly) is different as it passed the Gene Symbol and optionally cohort and gene filters.

```
interface CNVPlotProps {
  readonly gene: string;
  readonly height?: number;
  readonly genomicFilters?: FilterSet;
  readonly cohortFilters?: FilterSet;
```

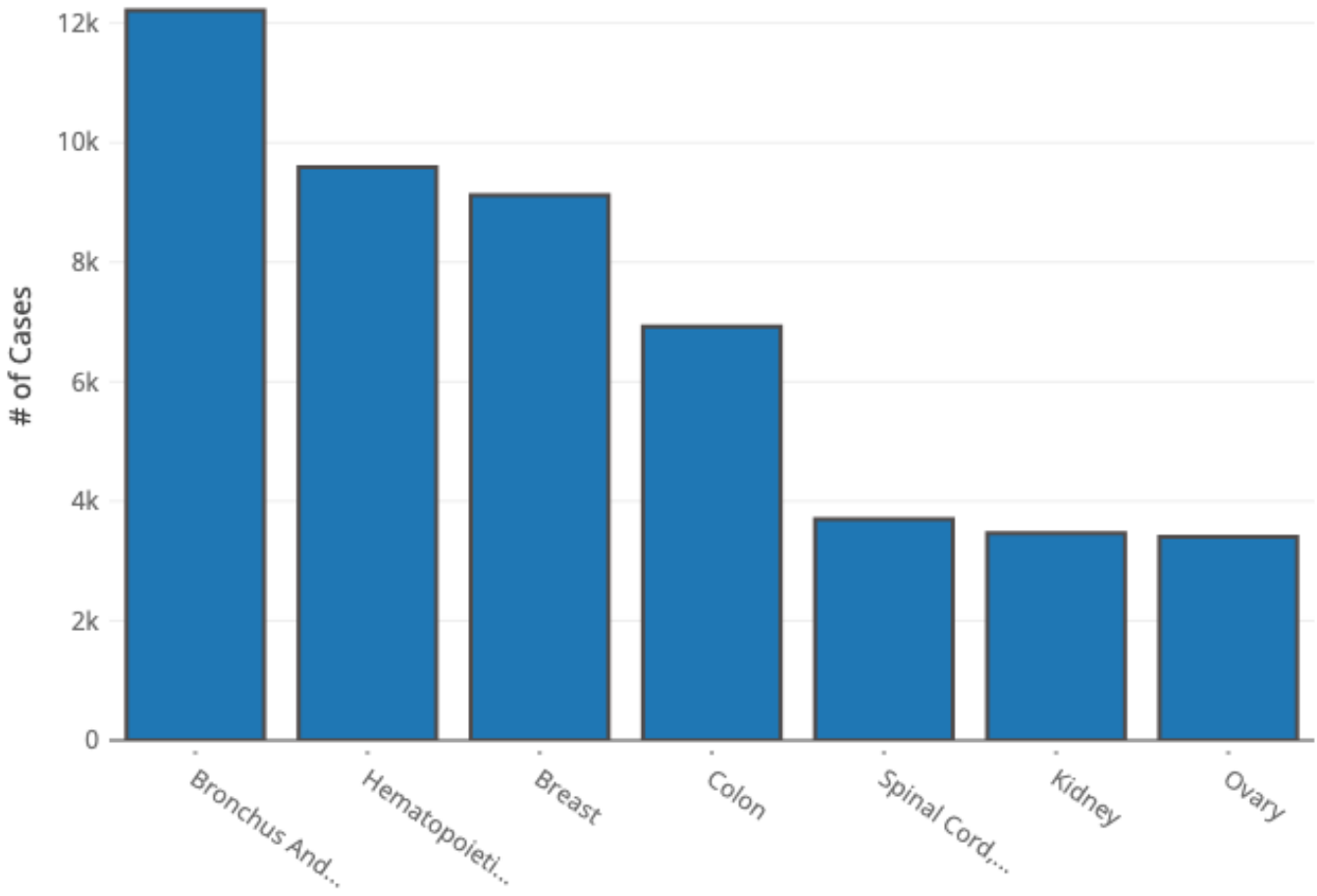


Figure 16.6: Bar Chart

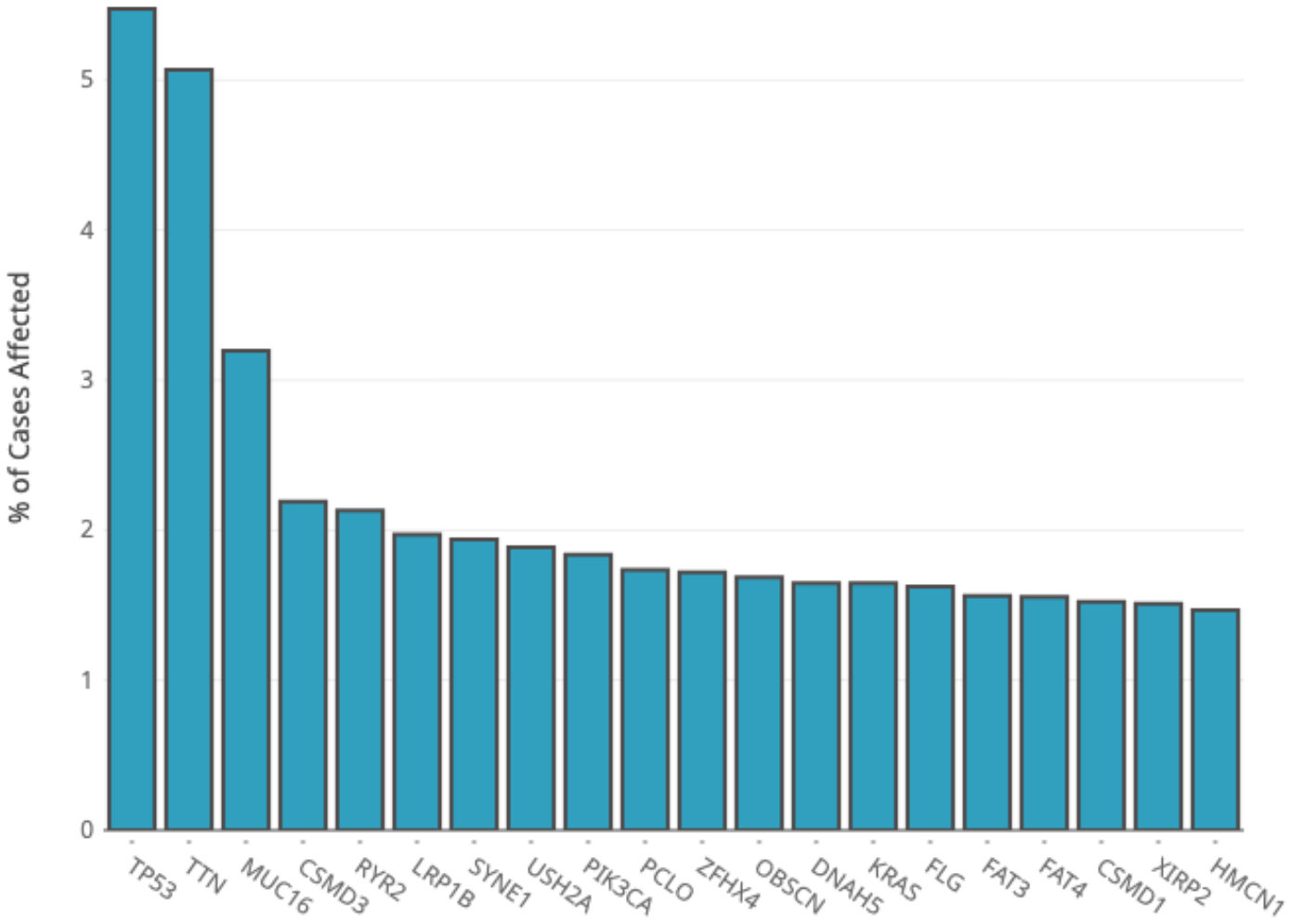


Figure 16.7: cancer distribution

```
}
```

These charts and others are documented in the Portal 2.0 SDK API documentation.

Facets

Facet components are provided for use in building local filters for your application. There are several types of facet components:

- `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 a text field
- `BooleanFacet` - a facet that is used to filter on a boolean field

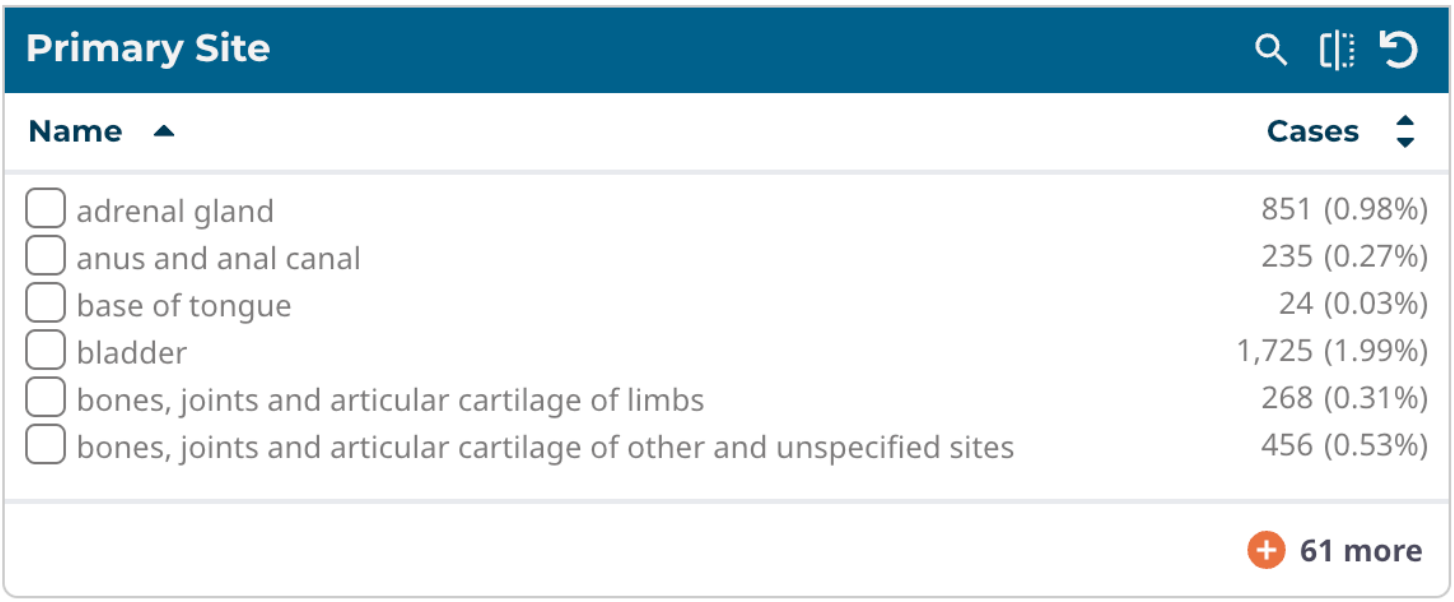


Figure 16.8: Enum Facet Component

Enum Facet

Range Facet

Date Range Facet

Number Range Facet

Percentile Facet

Age Range Facet

Exact Value Facet

Toggle Facet

The facet components are documented in the Portal 2.0 SDK API documentation. As these components are passed data fetcher and filter management hooks, they can be used for both cohort and local filters in an application.

VerticalTable

The VerticalTable component is used to display data in a table format. The VerticalTable component is a Mantine component implementing React table version 8. The VerticalTable component has a number of parameters, the most important being data, columns, and filters. The data is the data to display in the table, the columns are the columns to display in the table, and is

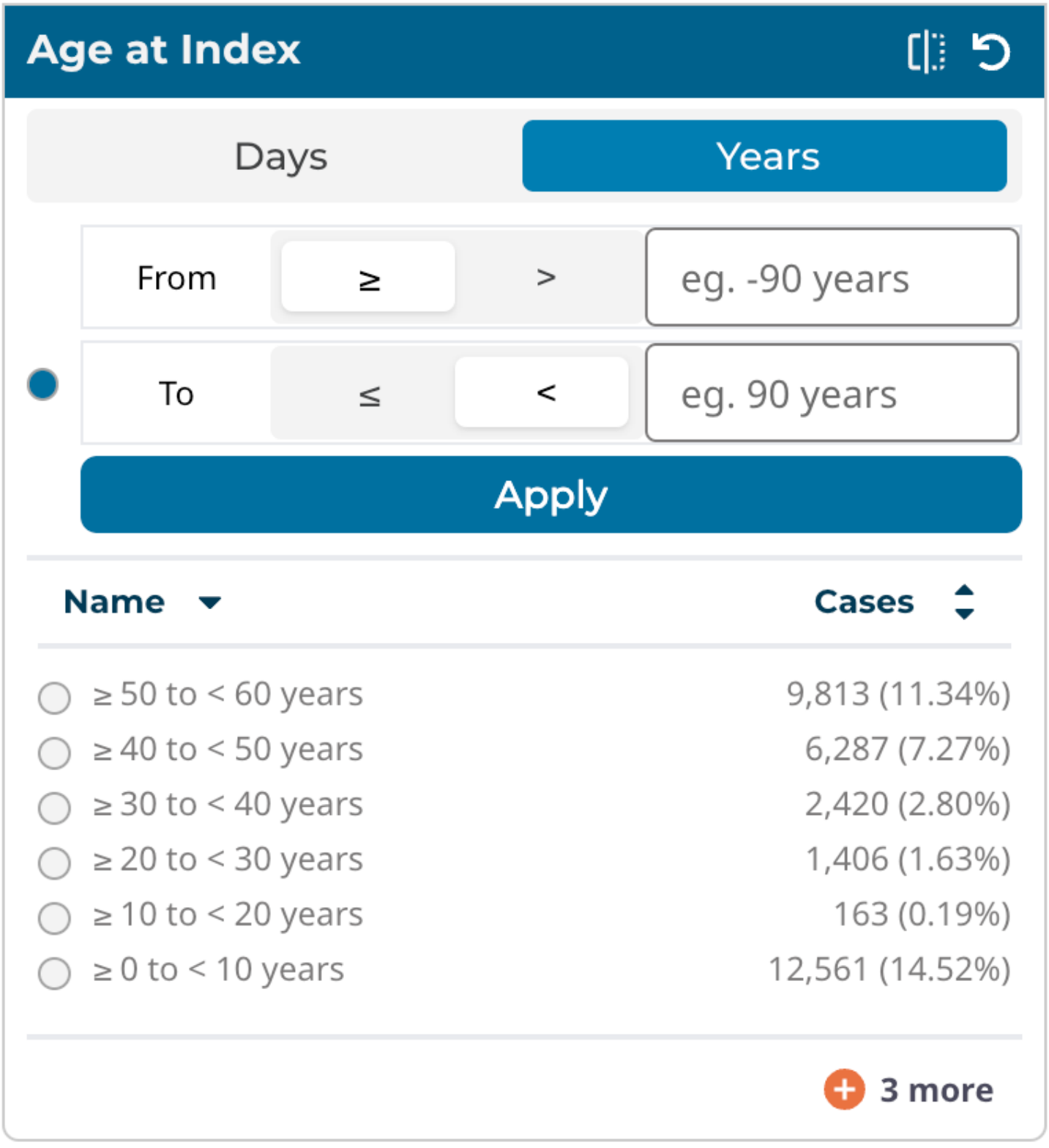


Figure 16.9: Range Facet Component

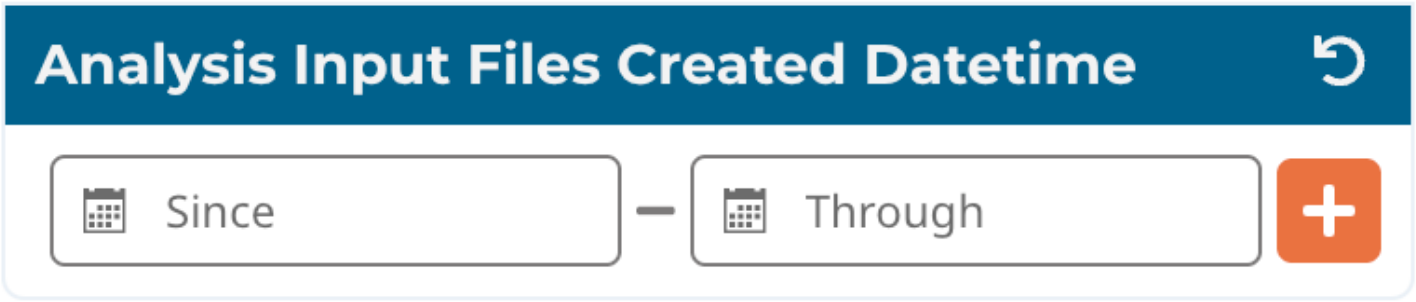


Figure 16.10: Date Range Facet Component

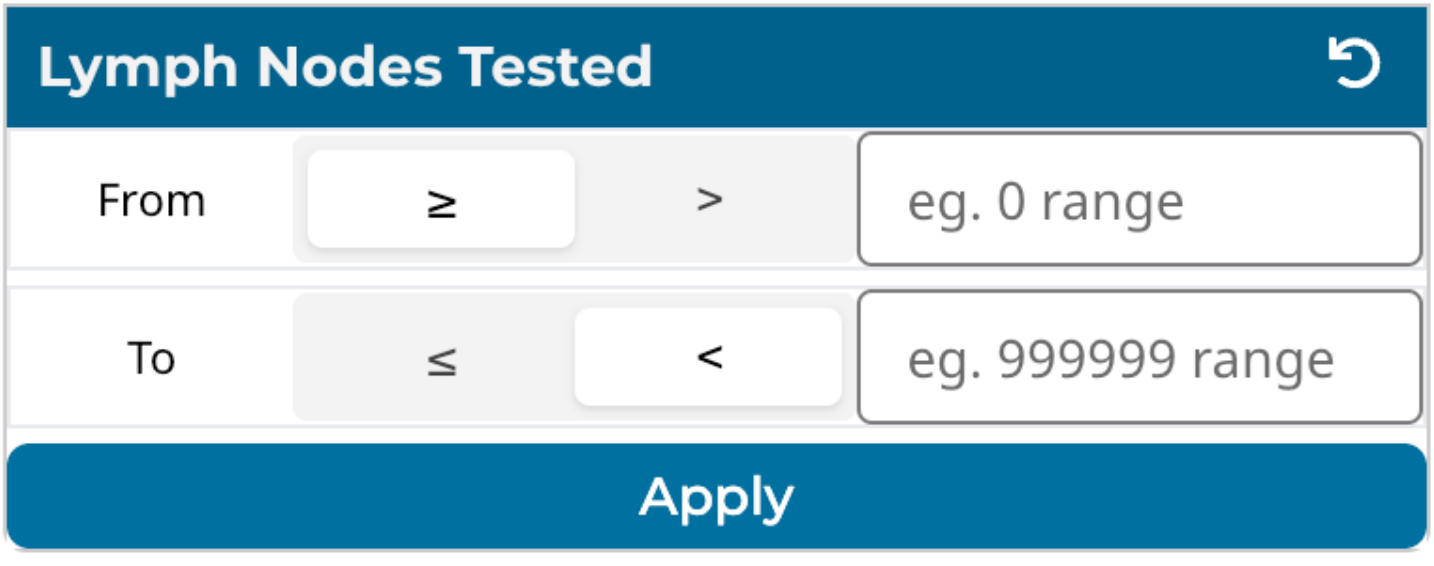


Figure 16.11: Number Range Facet Component

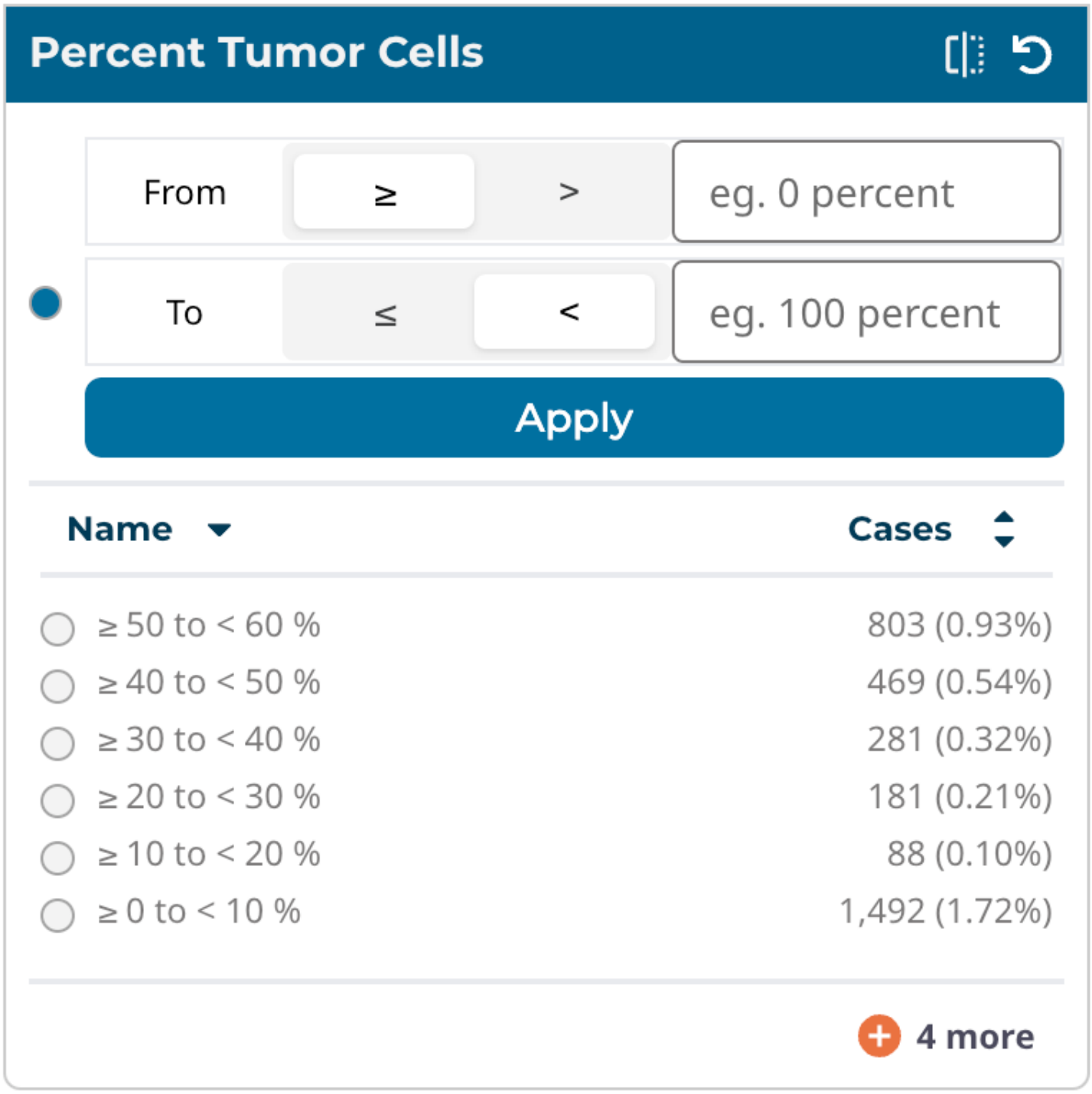


Figure 16.12: Percent Range Facet Component

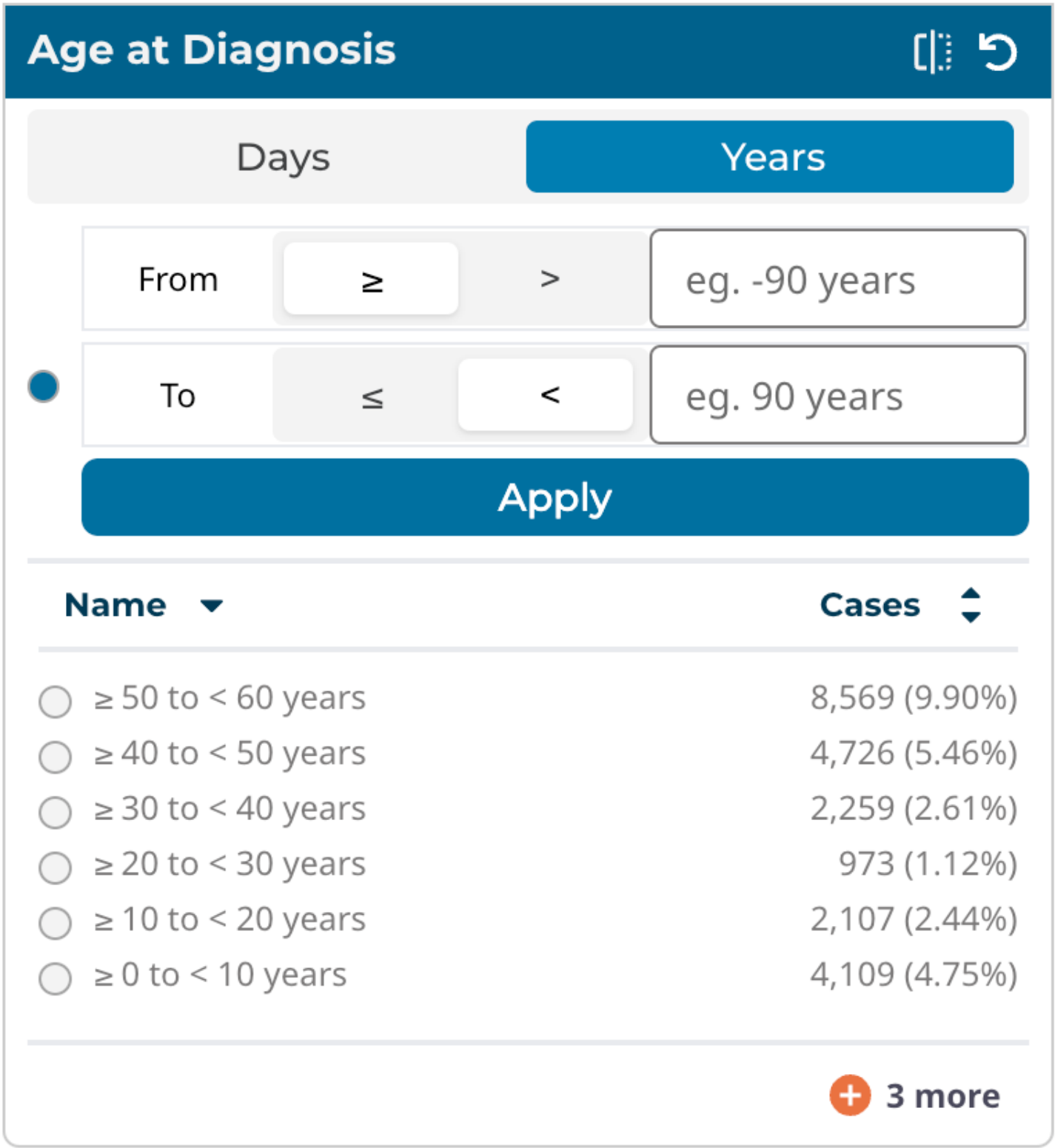


Figure 16.13: Age Range Facet Component

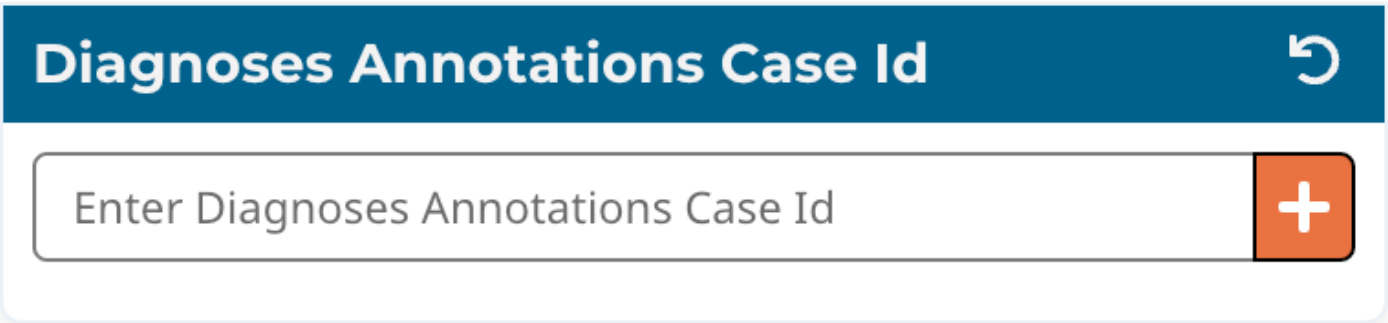


Figure 16.14: Exact Value Facet Component

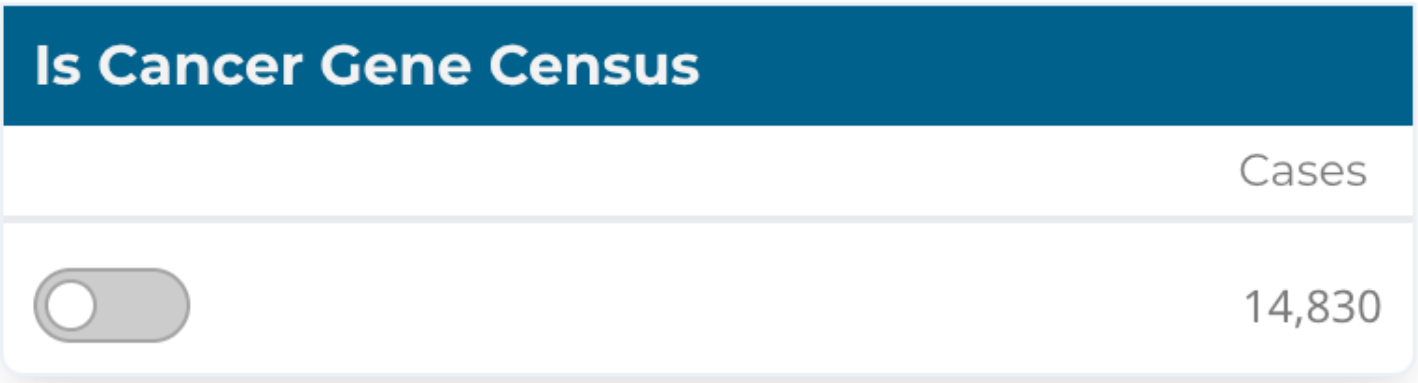
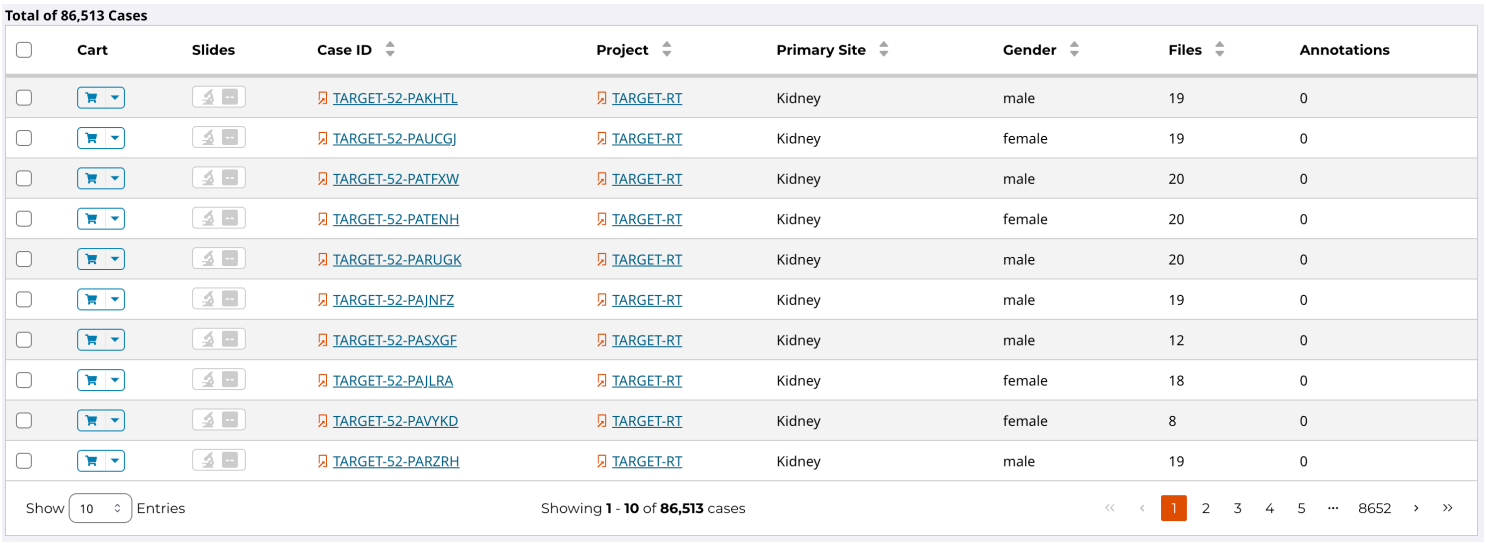


Figure 16.15: Boolean Toggle Facet Component

where the fields of the data tp be rendered are. The table has support for searching, sorting, and pagination. It can be configured to render many different types of columns, including text, numeric, and date. The table can also be configured to use React components for rendering columns. The Vertical Table is used for most of the table views in the GDC Portal. There are a number of examples of its use and is documented in the Portal 2.0 SDK API documentation.



Total of 86,513 Cases

Cart	Slides	Case ID	Project	Primary Site	Gender	Files	Annotations
		TARGET-52-PAKHTL	TARGET-RT	Kidney	male	19	0
		TARGET-52-PAUCGJ	TARGET-RT	Kidney	female	19	0
		TARGET-52-PATFXW	TARGET-RT	Kidney	male	20	0
		TARGET-52-PATENH	TARGET-RT	Kidney	female	20	0
		TARGET-52-PARUGK	TARGET-RT	Kidney	male	20	0
		TARGET-52-PAJNFZ	TARGET-RT	Kidney	male	19	0
		TARGET-52-PASXGF	TARGET-RT	Kidney	male	12	0
		TARGET-52-PAJLRA	TARGET-RT	Kidney	female	18	0
		TARGET-52-PAVYKD	TARGET-RT	Kidney	female	8	0
		TARGET-52-PARZRH	TARGET-RT	Kidney	male	19	0

Show 10 Entries Showing 1 - 10 of 86,513 cases 1 2 3 4 5 ... 8652

Vertical Table

Application Development

Getting Started

The GDC Portal 2.0 is a monorepo that contains all the code for the GDC Portal. The monorepo is managed using lerna and npm. The monorepo contains the following packages:

- `@gff/core` - contains the core components and hooks for the GDC Portal.
- `@gff/portal-proto` - contains the UI components and application framework (using NextJS) for the GDC Portal.

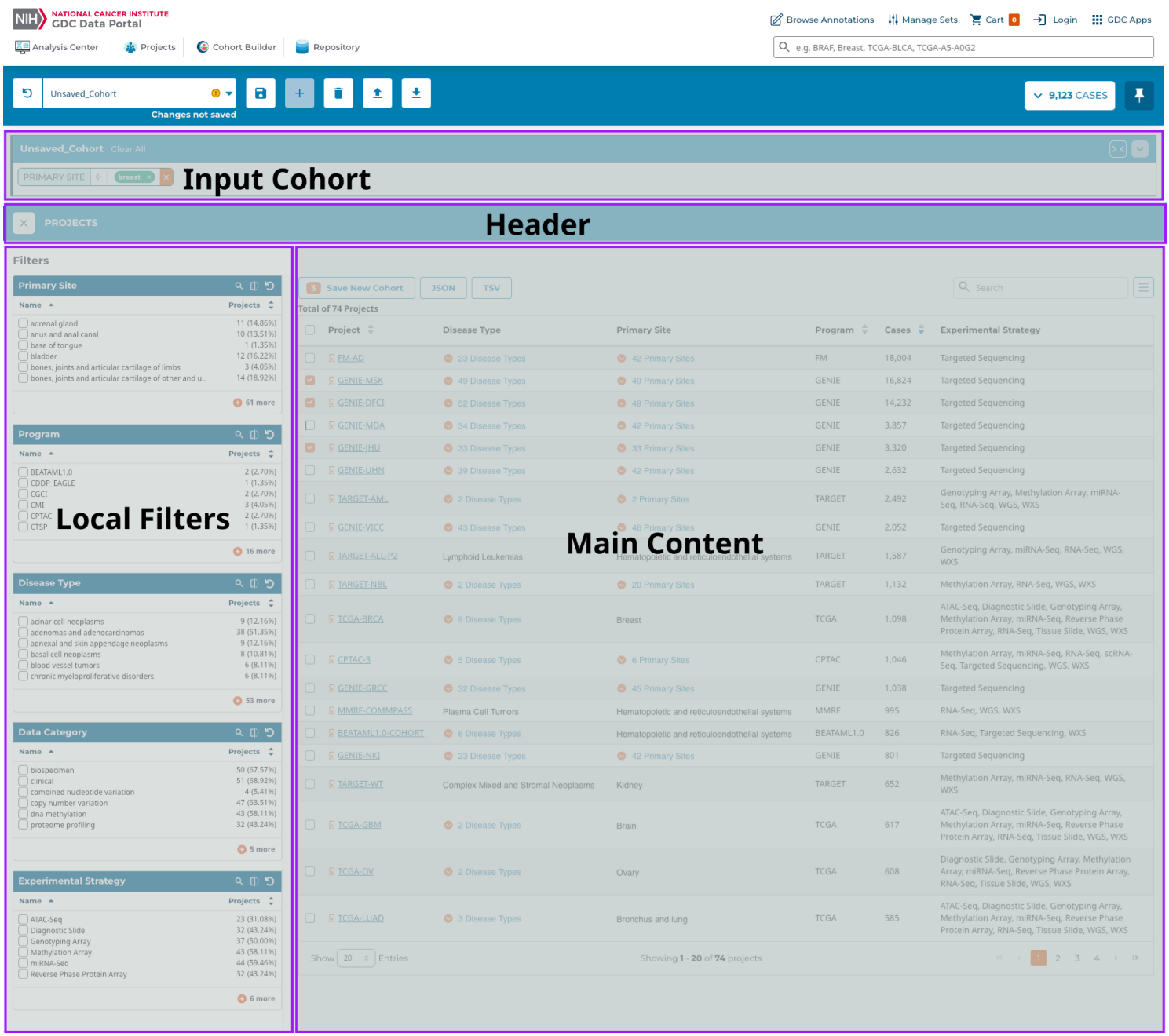
Note that the UI components located in the `@gff/portal-proto` package will be refactored into a separate package in the future, and `@gff/portal-proto` will be renamed to `@gff/portal`.

You can get started by cloning the repo and following the instructions in the README.md file.

Application Layout

A typical application will have the following layout. The main section of the application is the area where components like tables, graphs, and other components are displayed. Local filters are displayed on the left side and depending on the numbers will scroll vertically. This is a typical layout but other layouts are possible, like in the case of Protein Paint. Applications are encouraged to use vertical space as much as possible, as horizontal scrolling can be a poor user experience.

This section will describe parts of the Project application and how it is structured. The Project application is a simple application that displays a table of projects and allows the user to filter the projects by a number of filters. As the local filters are selected the table display is updated, but the cohort is not changed (e.i cohort filters are not updated). The Project application is located in the `@gff/portal-proto` package in the `src/features/projectsCenter` directory. The user can create a new saved cohort by selecting projects and clicking the “Save New Cohort” button. This will open a modal that will allow the user to name the cohort and save it. The Project application is a good example of how to use the GDC Portal SDK to create an application.



Major Sections of an Application

Local State

Depending on the application, it may be necessary to maintain the local state. For example, in the Projects application, the selected local filters, in this case, represented as Enumeration Facets, are stored in the local state. This allows the application to remember the selected filters when the user navigates away from the page and then returns. Persisting the state uses [Redux Toolkit] and [Redux Persist] to store the state in local storage. While the CoreState is managed by the portal core, the local state is managed by the application. Using a separate store for the local state allows the application to manage the state without having to worry about affecting the core state.

The Portal core provides a number of functions to assist in the creation and persisting of the redux store and will create handlers such as AppState, AppDispatch, and AppSelector. The AppState is the type of the local state, the AppDispatch is the type of the dispatch function, and the AppSelector is the type of the selector function.

An application can create all of the them using the `createAppStore` function:

```
import { createAppStore } from "@gff/core";
import { projectCenterFiltersReducer } from "./projectCenterFiltersSlice";

const PROJECT_APP_NAME = "ProjectCenter";

// create the store, context and selector for the ProjectsCenter
// Note the project app has a local store and context which isolates
// the filters and other store/cache values

const reducers = combineReducers({
  projectApp: projectCenterFiltersReducer, // Your application might have more that one reducer
});

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

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

Call this function will create the local store given the reducers and the name, and version of the application. The name of the application is used to create the local storage key for the application. The `id` is the id of the application and is used to create the local storage key. The `AppStore` is the local store, the `AppContext` is the local context, the `useAppSelector` is the selector hook, and the `useAppDispatch` is the dispatch hook.

Since you now have a local store, you can create a slice for the local state. The slice is a standard Redux Toolkit slice and will contain the reducer, actions, and selectors for the local state.

Persisting the local state

If is desirable to persist the local state. This can be done using the `persistReducer` function from the redux-persist package. Any reducer can be persisted by creating a persisted store and passing the reducer to the `persistReducer` function. For example, the `createAppStore` function can be modified to persist the local filter state as:

```
import { combineReducers } from "redux";
import { persistReducer } from "redux-persist";
import storage from "redux-persist/lib/storage";
import { createAppStore } from "@gff/core";
import { projectCenterFiltersReducer } from "./projectCenterFiltersSlice";

const PROJECT_APP_NAME = "ProjectCenter";

const persistConfig = {
  key: PROJECT_APP_NAME,
  version: 1,
  storage,
  whitelist: ["projectApp"],
};

// create the store, context and selector for the ProjectsCenter
// Note the project app has a local store and context which isolates
// the filters and other store/cache values

const reducers = combineReducers({
  projectApp: projectCenterFiltersReducer,
});
```

```

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

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

For example, the `projectCenterFiltersSlice.ts` which handles the local filters, is defined as:

```
import { createSlice, PayloadAction } from "@reduxjs/toolkit";
import { Operation, FilterSet } from "@gff/core";
import { AppState } from "./appApi";

export interface ProjectCenterFiltersState {
  readonly filters: FilterSet;
}

const initialState: ProjectCenterFiltersState = {
  filters: { mode: "and", root: {} },
};

const slice = createSlice({
  name: "projectCenter/filters",
  initialState,
  reducers: {
    updateProjectFilter: (
            state,
            action: PayloadAction<{ field: string; operation: Operation }>,
    ) => {
      return {
        ...state,
        filters: {
          mode: "and",
          root: {
            ...state.filters.root,
            [action.payload.field]: action.payload.operation,
          },
        },
      };
    },
    removeProjectFilter: (state, action: PayloadAction<string>) => {
      // eslint-disable-next-line @typescript-eslint/no-unused-vars
      const { [action.payload]: _, ...updated } = state.filters.root;
      return {
        ...state,
        filters: {
          mode: "and",
          root: updated,
        },
      };
    },
    clearProjectFilters: () => {
      return { filters: { mode: "and", root: {} } };
    },
  },
```

```
  extraReducers: {},
});

export const projectCenterFiltersReducer = slice.reducer;
export const { updateProjectFilter, removeProjectFilter, clearProjectFilters } =
        slice.actions;

export const selectFilters = (state: AppState): FilterSet | undefined =>
        state.projectApp.filters;

export const selectProjectFiltersByName = (
        state: AppState,
        name: string,
): Operation | undefined => {
  return state.projectApp.filters.root[name];
};
```

The above code creates a slice for the local state. The slice contains the reducer and the actions for the local state.

The reducer is `projectCenterFiltersReducer` and the actions are `updateProjectFilter`, `removeProjectFilter`, and `clearProjectFilters`. The selectors are `selectFilters` and `selectProjectFiltersByName`. The `selectFilters` selector returns the filters for the application, while the `selectProjectFiltersByName` selector returns the filter for a given name.

Application Hooks

The above can be used to define hooks for use in the local filter EnumFacet component. For example, the `useProjectFiltersByName` hook is implemented as:

```
export const useUpdateProjectsFacetFilter = (): UpdateFacetFilterFunction => {
  const dispatch = useAppDispatch();
  // update the filter for this facet

  return (field: string, operation: Operation) => {
    dispatch(updateProjectFilter({ field: field, operation: operation }));
  };
};
```

Clearing the local filters is done using the `clearProjectFilters` action:

```
export const useClearProjectsFacetFilters = (): ClearFacetFiltersFunction => {
  const dispatch = useAppDispatch();
  // clear the filters for this facet

  return () => {
    dispatch(clearProjectFilters());
  };
};
```

Note that the dispatch function is the `useAppDispatch` hook returned by the `createAppStore` function. The user-selected local filters can be retrieved using the `useProjectsFilters` hook created by combining the `useAppSelector` hook and the `selectFilters selector`:

```
export const useProjectsFilters = (): FilterSet => {
  return useAppSelector((state) => selectFilters(state));
};
```

Creating a new cohort

The project application allows users to create a new cohort from the selected projects. The cohort is created using the `SaveCohortModal` component. The `SaveCohortModal` component passes the current cohort filters and the local project filters to create a new saved cohort. In the case of the project application, the `SaveCohortModal` component is used in a button component. The button component is passed the selected projects and the `SaveCohortModal` component is rendered when the button is clicked. The `SaveCohortModal` component passes the current cohort filters and the local project filters to create a new saved cohort. The `SaveCohortModal` component is used in the project application as:

```
import React, { useState } from "react";
import { Button, Tooltip } from "@mantine/core";
import { CountsIcon } from "@/components/tailwindComponents";
import SaveCohortModal from "@/components/Modals/SaveCohortModal";

const ProjectsCohortButton = ({ pickedProjects, }: { pickedProjects: string[]; }): JSX.Element => {
  const [showSaveCohort, setShowSaveCohort] = useState(false);

  return (
          <>
            <Tooltip
                    label="Save a new cohort of cases in selected project(s)"
                    withArrow
            >
        <span>
          <Button
                  data-testid="button-create-new-cohort-projects-table"
                  variant="outline"
                  color="primary"
                  disabled={pickedProjects.length == 0}
                  leftIcon={
                    pickedProjects.length ? (
                            <CountsIcon $count={pickedProjects.length}>
                              {pickedProjects.length}{" "}
                            </CountsIcon>
                    ) : null
                  }
                  onClick={() => setShowSaveCohort(true)}
                  className="border-primary data-disabled:opacity-50 data-disabled:bg-base-max
                      data-disabled:text-primary"
          >
            Save New Cohort
          </Button>
        </span>
            </Tooltip>
            {showSaveCohort && (
                    <SaveCohortModal
                            onClose={() => setShowSaveCohort(false)}
                            filters={{
                              mode: "and",
                              root: {
                                "cases.project.project_id": {
                                  operator: "includes",
                                  field: "cases.project.project_id",
                                  operands: pickedProjects,
                                },
                              },
                            }}
                    />
```

```
            )}
          </>
  );
};

export default ProjectsCohortButton;
```

This custom button component used the state variable `showSaveCohort` to determine if the `SaveCohortModal` component needs to be shown. The `SaveCohortModal` component is passed the current list of projects selected by the user and handles the creation of the cohort and saving it.

Application Demo

In addition to a application that works on cohorts, an application can have a demo. This demo can be used to show the application's functionality. The demo is shown when the demo button is clicked. The demo button is shown when the application is registered with `hasDemo: false,` as described in the Application Registration section.

The application can determine if the demo button should be shown by using the `useHasDemo` hook. The `useHasDemo` hook returns a boolean indicating if the demo button should be shown. The demo button can be shown using the following code:

```
import { useIsDemoApp } from "@/hooks/useIsDemoApp";

const GenesAndMutationFrequencyAnalysisTool: React.FC = () => {
  const isDemoMode = useIsDemoApp();
  ...
```

Application Registration

An application needs to be "registered" to be used in the GDC Portal. Registration is done by adding the application using `createGdcAppWithOwnStore`function. If the app is not using its store, then the `createGdcApp` function can be used.

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

export default createGdcAppWithOwnStore({
  App: ProjectsCenter,
  id: id,
  name: "Projects Center",
  version: "v1.0.0",
  requiredEntityTypes: [],
  store: AppStore,
  context: AppContext,
});

export const ProjectsCenterAppId: string = id;
```

The above code registers the application with the GDC Portal. The `createGdcAppWithOwnStore` function takes a number of arguments:

- `App`: React.ComponentType - the application component
- `id`: string - the id of the application
- `name`: string - the name of the application
- `version`: string - the version of the application
- `requiredEntityTypes`: string[] - the required entity types for the application
- `store`: Store - the store for the application
- `context`: Context - the context for the application

The required entity types are the entity types that the application requires to function. For example, the Mutation Frequency application requires the `ssms` entity type. While this value is not currently used, it will be used in the future to determine if the application can be used.

The other registration needed for your app is in packages/portal-proto/src/features/user-flow/workflow/registeredApps.tsx This file contains an array of registered applications. For example the entry for the Project Center is:

```
import ProjectsIcon from "public/user-flow/icons/crowd-of-users.svg";

...
{
  name: "Projects",
          icon: (
          <ProjectsIcon
                  width={64}
                  height={64}
                  viewBox="0 -20 128 128"
                  role="img"
                  aria-label="Projects icon"
          />
          ),
  tags: [],
          hasDemo: false,
          id: "Projects",
          countsField: "repositoryCaseCount",
          description:"View the Projects available within the GDC and select them for further exploration
              and analysis.",
},
...
```

The above code registers the Project Center application with the GDC Portal. The members of the object are:

name *is the name of the application* `icon` is the icon as an SVH file, it size and position can be adjusted using the `width`, `height`, and `viewBox` properties ***tags*** *are the tags for the application used for searching (which is not currently active)* `hasDemo` is a boolean indicating if the application has a demo, if so the demo button will be shown ***id*** *is the id of the application and needs to match the id of the application registered in the* ***createGdcAppWithOwnStore*** *function* `countsField` is the field to use for the counts in the application, this is used to determine if the application can be used ***description*** *is the description of the application* `noDataTooltip` is the tooltip to show if the application has no data

When the app is registered, it will be available in the GDC Portal. The application can be accessed by clicking on the app card. The visual elements of the card are:

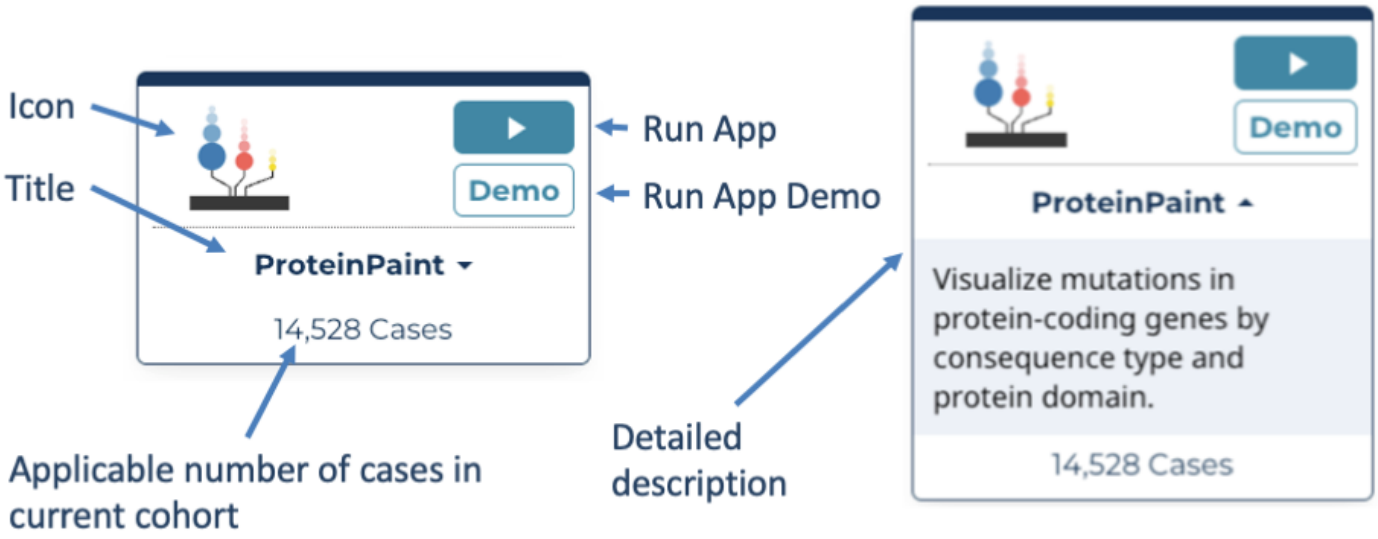


Figure 16.16: application_card.png

Application Card and it's elements

Source code layout

While you are free to structure your application code as you with, the following is a recommended layout for your application's source code:

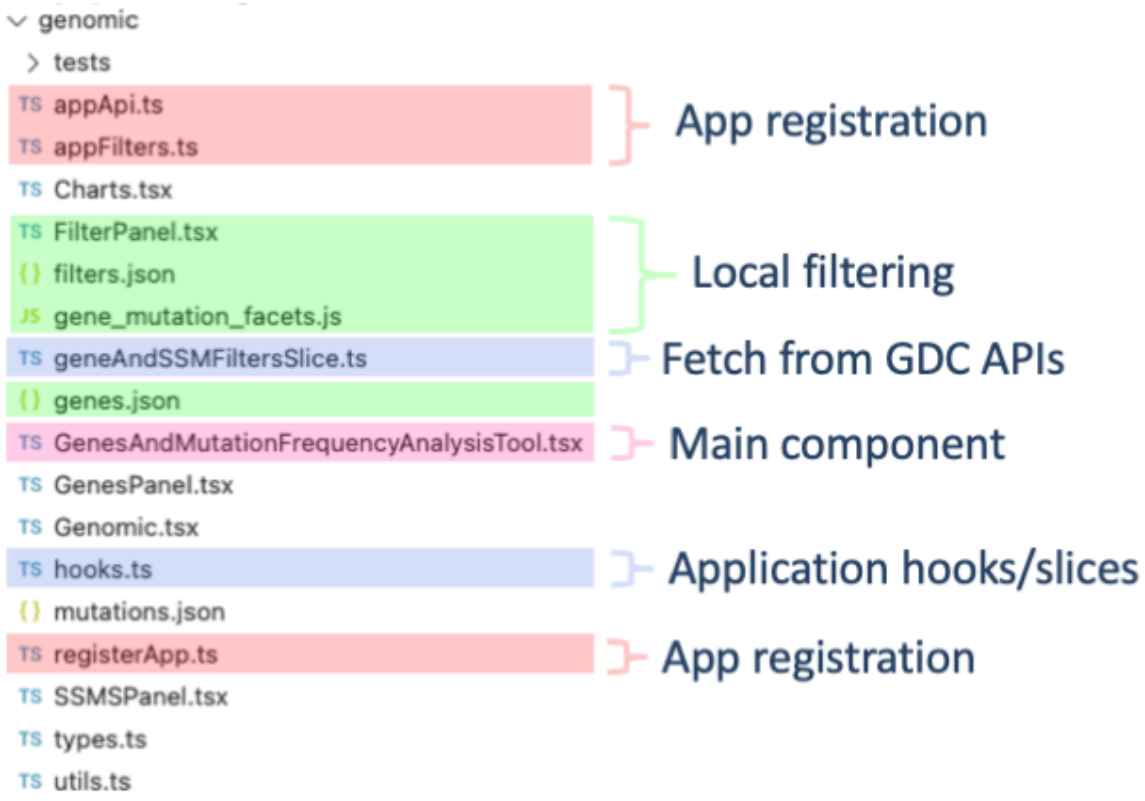


Figure 16.17: source code layout

Application source code layout

Appendix

Using selectors and hooks

Although a complete guide to react hooks and selectors is out of the scope of this document, we will provide a brief overview of how to use them for application development. For more information on hooks and selectors please see the React Hooks. As we are using Redux-toolkit, we will be using the calls described in the Redux Toolkit documentation.

Selectors

Selectors are used to access the state of the GDC Portal's main redux store. Using selectors is the preferred method for accessing the state of the GDC Portal. Selectors are functions that take the state as an argument and return a value.

```
1 import {useCoreSelector,  selectCurrentCohort } from '@gff/core';
2
3 const currentCohort = useSelector(selectCurrentCohort);
```

The selector will return the current value of the item in the store. Consult the GDC 2.0 API documentation for a complete list of selectors.

Hooks

Fetching data from the GDC API is done via hooks. Hooks are functions that take arguments and return a value. The value returned is typically a promise that resolves to the data requested. The GDC Portal provides a number of hooks for fetching data from the GDC API. These hooks are located in the `@gff/core` package.

```
import { useGeneSymbol } from '@gff/core';

const { data: geneSymbolDict, isSuccess } = useGeneSymbol(
        field === "genes.gene_id" ? facetValues.map((x) => x.toString()) : [],
);
```

GDC Portal hooks are designed to work similarly to the RTL Query hooks. The hooks take arguments and return a object. The object contains the data and the status of the query. The status of the query is stored in the `isSuccess` variable. The data returned from the query is stored in the `data` variable. The object returned from a GDC hook is of the form:

```
{
  data: any;
  isSuccess: boolean;
  isLoading: boolean;
  isError: boolean;
  error: Error;
}
```

where `data` is the data returned from the query, `isSuccess` is a boolean indicating if the query was successful, `isLoading` is a boolean indicating if the query is currently loading, `isError` is a boolean indicating if the query resulted in an error, and `error` is the error returned from the query.

Querying the GDC API Directly

There may be cases where you need to query the GDC API directly. The GDC Portal provides a number of functions for querying the GDC API. These functions are located in the `@gff/core` package. The functions are: * `fetchGdcProjects` - fetches project data * `fetchGdcAnnotations` - fetches annotation data * `fetchGdcSsms` - fetches ssms data * `fetchGdcCases` - fetches cases data * `fetchGdcFiles` - fetches files data

which are wrappers around `fetchGdcEntities` function. The `fetchGdcEntities` function takes a number of arguments:

```
export interface GdcApiRequest {
  readonly filters?: GqlOperation;
  readonly case_filters?: GqlOperation;
  readonly fields?: ReadonlyArray<string>;
  readonly expand?: ReadonlyArray<string>;
  readonly format?: "JSON" | "TSV" | "XML";
  readonly size?: number;
  readonly from?: number;
  readonly sortBy?: ReadonlyArray<SortBy>;
  readonly facets?: ReadonlyArray<string>;
}
```

There is also support for the GraphQL API. The `fetchGdcGraphQL` function takes two arguments:

```
export const graphqlAPI = async <T>(
  query: string,
  variables: Record<string, unknown>,
): Promise<GraphQLApiResponse<T>> =>
```

where `query` is the GraphQL query and `variables` are the variables for the query.

API Documentation

To access the Developers documentation for the GDC API, use the following commands in your terminal:

```
git clone git@github.com:NCI-GDC/gdc-frontend-framework.git
cd gdc-frontend-framework/
git checkout feat/with_api_docs
```

Next, within the repo, open `docs/api/index.html` in your browser.

Chapter 17

Release Notes

Data Portal Release Notes

Version	Date
v2.0.0	February 8, 2024
v1.30.4	May 11, 2023
v1.30.0	July 8, 2022
v1.29.0	August 23, 2021
v1.28.0	May 17, 2021
v1.25.1	August 14, 2020
v1.25.0	July 2, 2020
v1.24.1	March 10, 2020
v1.23.1	December 10, 2019
v1.23.0	November 6, 2019
v1.22.0	July 31, 2019
v1.21.0	June 5, 2019
v1.20.0	April 17, 2019
v1.19.0	February 20, 2019
v1.18.0	December 18, 2018
v1.17.0	November 7, 2018
v1.16.0	September 27, 2018
v1.15.0	August 23, 2018
v1.14.0	June 13, 2018
v1.13.0	May 21, 2018
v1.12.0	February 15, 2018
v1.11.0	December 21, 2017
v1.10.0	November 16, 2017
v1.9.0	October 24, 2017
v1.8.0	August 22, 2017
v1.6.0	June 29, 2017
v1.5.2	May 9, 2017
v1.4.1	October 31, 2016
v1.3.0	September 7, 2016
v1.2.0	August 9th, 2016
v1.1.0	June 1st, 2016
v1.0.1	May 18, 2016

Release 2.0.0

- **GDC Product**: GDC Data Portal
- **Release Date**: February 8, 2024

New Features and Changes

GDC 2.0 is a major update to the original GDC Data Portal introduced in 2016. This latest version adopts a "cohort-centric" workflow, in which users build custom sets of cases to analyze, and introduces several new analysis tools. New features of GDC 2.0 include:

- A cohort-centric workflow in which a cohort is first built and then analyzed using tools on the Data Portal. All of these functionalities can be reached from the Analysis Center.
  - This includes a toolbar, that can be used to view or modify an existing cohort while using any analysis tool.
- Core tools that compose the main functionalities of the GDC Data Portal:
  - **Cohort Builder:** Build a cohort of cases using clinical and biospecimen properties
  - **Repository:** Download files based on a specific cohort
  - **Projects:** Browse, filter, and create cohorts based on GDC projects
- Analysis tools that analyze specific cohorts:
  - **Mutation Frequency:** Analyze somatic mutations that were called in the WXS and Targeted Sequencing pipelines and their associated genes
  - **Clinical Data Analysis:** Analyze and visualize clinical data associated with your cohort
  - **Cohort Comparison:** Analyze the properties of multiple cohorts
  - **Set Operations:** Display a Venn diagram and compare/contrast cohorts or gene/mutation sets
  - **BAM Slicing Download:** Download a specific region of a BAM file created by the GDC
  - **ProteinPaint:** Visualize somatic mutations on a specific linear gene or chromosomal region
  - **Gene Expression Clustering:** Visualizes gene expression clustering for a specific cohort
  - **Sequence Reads:** Visualize the reads within a specific BAM file
  - **OncoMatrix:** Visualize the most commonly mutated genes across a cohort

Bugs Fixed Since Last Release

Not applicable as this is the initial release of GDC 2.0.

Known Issues and Workarounds

- **Section 508 Accessibility**:
  - There are known Section 508 accessibility issues that the GDC plans to address in subsequent releases. If a user encounters a Section 508 barrier, please contact GDC Support (support@nci-gdc.datacommons.io) for assistance. Known Section 508 issues are identified below.
    - There are keyboard focus and navigation issues in analysis tools that use popup windows/overlays for custom user selections. Impacted analysis tools include BAM Slicing, Sequence Reads, Gene Expression Clustering, OncoMatrix, and ProteinPaint.
    - Heatmaps within the Sequence Reads tool do not contain concise alternative text or equivalent alternatives. Additionally, equivalent alternatives to the Box plots, QQ plots, Venn diagrams, and the body plot on the home page are not available.
    - In the Gene Expression Clustering tool and OncoMatrix, there are no headers for genes, clusters, and/or cases in the heatmap.
    - In the Gene Expression Clustering tool, color is used to convey gene expression values but there are no patterns to convey the same information as color. Color is also used in ProteinPaint and the Sequence Reads tool to convey consequence type but there are no distinguishing patterns.
    - Some text can be difficult to read on a small screen at a 200% zoom level.
    - Keyboard focus is not returned to the triggering element when modals are closed.
    - Assistive technologies may not behave correctly with some controls due to incorrect, missing, or redundant labels, attributes, or roles.
- **Cohorts**:
  - Cohorts are under active development and their behavior may change in the first several months after the release of GDC Portal 2.0. As this process may result in the loss of saved cohorts on the portal, we highly recommend exporting cohorts locally.

  - Cohorts created based on CNV losses or gains may not have the correct composition when filtered by additional mutated genes. As a workaround, first filter by the mutated genes before creating cohorts based on CNV losses and gains.
  - Cohorts filtered by mutated genes and SSMs not in those genes may unexpectedly result in 0 cases.
  - When saving a cohort, the confirmation notification may be automatically dismissed before the saving dialog has closed.
  - Using "Save As" to replace a cohort with itself will result in an error notification despite the replacement being successful.
  - Cohorts containing FM-AD cases may not update correctly when users with dbGaP access to FM-AD (phs001179) log in or out. As a workaround, logging in before creating cohorts with FM-AD cases is recommended.
  - If removing gene/mutation filters from a cohort temporarily results in 0 cases, cohorts may not display data in Mutation Frequency, Cohort Builder, and the summary charts. As a workaround, remove the gene and mutation filters, then add them back.
- **Survival Plot**:
  - The survival plot in Cohort Comparison does not display text indicating that there is insufficient survival data to plot.
  - The survival plot in Mutation Frequency may flicker when the cohort has 0 cases.
  - In Mutation Frequency, the downloaded image may display a survival curve when none is plotted within the portal.
  - When the survival plot is zoomed in and an image is downloaded, the curves within the image may extend beyond the y-axis.
- **Cart**:
  - Spinners on the Download Cart and Download Associated Data buttons may be displayed longer than expected. This is a visual issue and does not affect the use of these buttons.
  - More than 5 GB of files in total may be downloaded at a time via the browser if the user first attempts to download controlled access data without being logged in, then logs in via the information dialog displayed before continuing with the download.
  - Using multiple browser tabs with the portal when adding or removing files from the cart may result in the cart not being updated as expected.
- **Mutation Frequency**:
  - Gene/mutation sets created from the tables in Mutation Frequency may contain 0 genes/mutations if the cohort has Available Data filters or Biospecimen filters.
  - Attempting to download a TSV of all the mutations in the GDC may result in an error due to the length of time needed to generate the TSV. As a workaround, limit the number of mutations downloaded to 1.5 million.
- **Main Toolbar**:
  - Attempting to download the **Clinical/Biospecimen TSV or JSON** before the cohort has fully loaded may result in an error.
  - The TSV of the **cases table** may not contain the expected tabs.
- **OncoMatrix**:
  - Manually deleting all genes will result in an error message "Error: Cannot read properties of undefined (reading 'lst')". The user can close and re-open OncoMatrix for use.
  - Dragging genes only works once. After one gene is dragged to a new position, no genes can be dragged to new positions.
- **ProteinPaint**:
  - A nested filter may be constructed for a Lollipop subtrack, e.g. sex=male AND (primarysite=aa OR disease=bb), but cannot be translated into GDC cohort filters. The translation code has a preliminary implementation that only works for "flat" filters without nesting.
  - Cohorts cannot be created using the Create Cohort button in ProteinPaint for a single sample.
- In the **Gene Expression Clustering** tool, if any part of the dendrogram is selected and the current cohort is modified, then the new dendrogram will render with scattered subtrees selected.
- The "A" in the Allele Summary text is cut off in the **Sequence Reads** tool.
- **Quick Search** may not display results if the the same search input is applied twice quickly. As a workaround, temporarily change the input before reentering the intended search.
- Filters related to numeric values may display a smaller number than what the user entered within the **Cohort Builder**. This is a visual issue and does not affect the filters applied to the cohort.
- When the **Cohort Comparison** tool is loading, the loading spinner may be displayed above the other areas of the Analysis Center.
- The **Repository** tool may display an incorrect file size total of 0 bytes when filtering is applied within the tool and the active cohort contains Available Data filters.
- The **Slide Image Viewer** will display a black image temporarily if a user zooms in on a slide then switches to another slide.

- In **Set Operations**, the saving of gene and mutation sets may be unsuccessful if the saving dialog is manually dismissed after the Save button is clicked.
- Clicking the X button on the **Unexpected Error dialog box** does not dismiss it. The workaround is to click the OK button.

Release 1.30.4

- **GDC Product**: GDC Data Portal
- **Release Date**: May 11, 2023

New Features and Changes

- The GDC Legacy Archive has officially been retired.
  - The Legacy Archive Portal can no longer be reached.
  - Any API call to query files from the Legacy Archive will no longer work.
  - Downloads for files from the Legacy Archive will work normally with manifests that were generated previously.

Bugs Fixed Since Last Release

- The Clinical TSV download in the case entity page and cart now contain TSVs for pathology detail, follow up, and molecular test entities.
- Fixed bug in which demographic information would not download as a TSV when diagnosis information was not available.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.30.0

- **GDC Product**: GDC Data Portal
- **Release Date**: July 8, 2022

New Features and Changes

- None

Bugs Fixed Since Last Release

- Fixed error in which the manifest could not be downloaded directly from the GDC Data Portal in certain instances.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.29.0

- **GDC Product**: GDC Data Portal
- **Release Date**: August 23, 2021

New Features and Changes

- None

Bugs Fixed Since Last Release

- Fixed error in which the data summaries in the clinical analysis and cart pages were only partially displayed when viewed in Chrome v.91.0.4472.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations

  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.28.0

- **GDC Product**: GDC Data Portal
- **Release Date**: May 17, 2021

New Features and Changes

- New columns were added to the "molecular test" table at the bottom of the case entity page to display additional molecular test fields.

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- When accessing the data portal with Chrome v.91.0.4472, users may experience some display errors. This includes the data summary in the clinical analysis and cart pages being only partially displayed.
- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.

- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.25.1

- **GDC Product**: GDC Data Portal
- **Release Date**: August 14, 2020

New Features and Changes

- API improvements were made to increase portal performance.

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.

  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.25.0

- **GDC Product**: GDC Data Portal
- **Release Date**: July 2, 2020

New Features and Changes

- Suppressed Experimental Strategy filter on the Exploration page as this currently filters for files with a particular strategy, not for cases. This may cause confusion amongst users. The filter will be re-instated in a future release once the logic is available to filter more appropriately for cases tied to a specific strategy.
- Updated the filter control panel styling across the Portal to have clearer titles (e.g. “Search Cases” instead of “Cases” in the quick search box).
- Made minor updates to the styling of the filter query display at the top of the Exploration page (spacing, borders).
- Added an expand/collapse control to the quick search bar of Clinical tab on the Exploration page, to be consistent with other Exploration tabs.
- Added a clear title above the counts in each filter control panel across the Portal (e.g. " Cases“," Genes", etc.).
- Moved various action buttons above the results table on the Repository Page to more accessible locations.
- Improved load time of the initial custom filter list on the Repository Page, when clicking “Add a Filter Filter” or “Add a Case/Biospecimen Filter”.

Bugs Fixed Since Last Release

- Fixed a bug in the Age at Diagnosis table on the Cohort Comparison page, where the of cases in the table was not consistent with the of cases shown when clicking the link to the Exploration page.
- Fixed minor positional accuracy issue of the lollipop data points on the Protein Viewer.
- Fixed bug on the Protein Viewer where, if clicking to switch between different lollipop data points, details of the previous lollipop was not closing.
- Fixed bug where the quick search bar on the Exploratin Page’s Genes filter tab was not expanding/collapsing properly.
- Fixed bug in the pop-up warning message when adding or removing items from the Cart, where long filenames were spilling outside the border of the pop-up.
- Fixed typo in the “View Cases in Exploration” button on the Repository page.
- Fixed typo in the pop-up user consent message when downloading controlled files from the Cart.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.

  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.24.1

- **GDC Product**: GDC Data Portal
- **Release Date**: March 10, 2020

New Features and Changes

- Removed unnecessary comma and y-axis value from title of the mutation details pop-up in the Protein Viewer.
- Added Tobacco Smoking Status field to the Exposures tab on the Case entity page.
- Added a link to the Cart where users can access instructions for downloading the GDC Genome Build reference files.
- Added logic to prevent duplicate fetching of data for Clinical Analysis survival plots and optimize rendering.
- Added a button to clear searches for certain Portal search controls that were previously missing this ability.
- Reduced whitespace between Oncogrid and its control panel to optimize spacing and layout.
- Made entire Clinical Analysis results page responsive (card columns now scale & stack in response to the size of the browser window).
- Replaced Clinical Analysis function for printing clinical cards to a single PDF file, with more flexible functionality to instead download all the cards in SVG and/or PNG format.
- Added message to notify users when they try to access the Portal using Microsoft Internet Explorer, indicating which browsers are officially supported.
- Added arrow icon to sortable columns across the Portal to indicate the current sort direction.

Bugs Fixed Since Last Release

- Fixed bug where clicking a primary site on the Human Body Image was not re-directing to the Exploration page.
- Fixed layout issue where long Annotation Notes were exceeding the border of the text box.
- Fixed layout issue where the Repository header and action buttons were scaling and wrapping incorrectly if the browser window is shrunk beyond a certain threshold.
- Fixed layout issue where the responsive Clinical Analysis Cards were clipping improperly as the browser window is shrunk beyond a certain threshold.
- Fixed bug where the Clinical Tab on the Exploration page was crashing when entering a custom range of Years for the Age at Diagnosis facet.
- Fixed various minor cosmetic and color issues in PNG, SVG downloads of the Clinical Analysis survival plots.
- Fixed bug where the x-axis in PNG, SVG downloads of histograms across the Portal was being bolded incorrectly.
- Fixed bug where the expand/collapse symbols in the UI were incorrectly being exported in the TSV download of the Projects table.
- Fixed bug where Oncogrid's modal for customizing colors could not be scrolled below the fold if it was shrunk beyond a certain threshold.
- Fixed incorrect DTT hyperlink in the GDC Apps menu.
- Fixed bug where the "dbSNP rs ID" facet could not be minimized in the Exploration page's Mutations facet tab.
- Fixed layout issue where the Portal's header incorrectly overlaps some content when a notification banner is displayed.
- Fixed some minor layout & styling issues in the Exploration page's facets panel.
- Fixed bug where the Case ID on the Exploration page's Cases facet tab was not searchable in certain scenarios.
- Fixed bug where the Expand/Collapse button state was not changing properly when being used in the Biospecimen section of the Case entity page.
- Fixed incorrect capitalization of "dbGaP" in the Summary section of the Project entity page.
- Fixed layout issue where the Advanced Search query box on the Repository page could expanded beyond the margins of the box's border.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.23.1

- **GDC Product**: GDC Data Portal
- **Release Date**: December 10, 2019

New Features and Changes

- Updated display of x-axis units on the homepage Human Body chart to more easily display increased case counts for newly-added projects

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.

  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.23.0

- **GDC Product**: GDC Data Portal
- **Release Date**: November 6, 2019

New Features and Changes

- Added Clinical Data Analysis feature that allows Users to:
  - Explore clinical data via the new Clinical Tab on the Exploration page.
  - Build custom Case sets based on that clinical data for later analysis.
  - Create an analysis to examine the clinical variables in a Case set, using various tools including histograms, survival plots, box plots, QQ plots, and custom binning.
  - Download the data (as TSV, JSON) and plots (as PNG, SVG) of each clinical variable in an anlysis.
  - Save an analysis to local storage to resume later (as long as storage is not cleared).
- Added links to CIViC annotations on the Gene and Mutation entity pages.
- Updated the default Top Mutated Genes histogram on the Exploration page to display only COSMIC Genes by default.
- Added Follow-Ups tab and nested Molecular Tests to Case entity page.
- Added text to BAM slicing modal to instruct Users how to access unmapped reads.

Bugs Fixed Since Last Release

- Fixed font in exported PNGs, SVGs to be consistent with the Portal UI.
- Made custom Case and File filters in the Repository page case insensitive.
- Fixed bug where pfam domains in Protein Viewer could not be clicked in Firefox.
- Fixed bug where TSV download button could not be clicked in MS Edge.
- Fixed controlled access alert pop-up in the Cart so that the modal disappears correctly once the User has successfully logged in and initiated the download.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - Negative numbers may be displayed for the Missing value category in the Treatment node within a Clinical Analysis. This occurs with projects that have multiple treatment nodes per case. All other values should be accurate.

  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - The footer says version 1.9, but it is actually 1.13
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.22.0

- **GDC Product**: GDC Data Portal
- **Release Date**: July 31, 2019

New Features and Changes

- Replaced existing Clinical, Biospecimen columns on the Projects page with 4 columns: Clinical, Clinical Supplement, Biospecimen, Biospecimen Supplement. The Clinical and Biospecimen columns now link directly to the project page, and their counts indicate the total cases in the project. The Clinical Supplement and Biospecimen Supplement columns work the same as the old Clinical and Biospecimen columns - They link to the Repository page with Files filtered based on the Project and Data Category (Clinical or Biospecimen).
- Added a new icon to the GDC Apps menu, which links to the GDC Publications website page.
- Added the Synchronous Malignancy field to the Diagnoses / Treatments tab on the Case entity page.
- Added the Pack Years Smoked field to the Exposures tab on the Case entity page.
- Increased length of x-axis labels on histograms to 10 characters so that projects with names that are typically standard 10 chars will display fully (e.g. most TCGA projects like TCGA-BRCA).

Bugs Fixed Since Last Release

- Fixed bug where the PNG, SVG files for the Overall Survival Plot could not be downloaded.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart

  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Filtering by vital_status does not function in the Legacy Archive due to updates in how this property has been indexed. A workaround is to perform the case level filtering in the GDC Data Portal and copy the filter string for use in the Legacy Archive or the legacy API.
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.21.0

- **GDC Product**: GDC Data Portal
- **Release Date**: June 5, 2019

New Features and Changes

- Changed all Survival Plots to display the Duration (x-axis) in years instead of days.
- Updated data references to clinical properties throughout the Portal to match the underlying changes in the GDC data dictionary.

Bugs Fixed Since Last Release

- Fixed bug where X-axis labels in histograms were cut off when displayed.
- Renamed the 'Experimental Strategies' facet on the Projects page to singular form.
- Fixed bug where columns with a % value of infinity (due to division by zero) show as 'NaN%'. Replaced instead with a label of '–'.
- Fixed bug where the download button in the cart access banner was still disabled after a user logged in from the banner. Instead, the experience is now improved so that after login, the banner is closed and the user must explicitly click 'Download' again.
- Fixed bug where if a new user logs into the Portal and views their profile, the app crashes if the user has no projects assigned yet.
- Fixed bug where Survival Rate numbers in the Survival Plot plot y-axis did not scale properly and overlapped into the axis lines.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive

  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.20.0

- **GDC Product**: GDC Data Portal
- **Release Date**: April 17, 2019

New Features and Changes

- Upgraded the Portal to use the latest React Javascript library (version 16.8)

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.19.0

- **GDC Product**: GDC Data Portal
- **Release Date**: February 20, 2019

New Features and Changes

- Added support for viewing of controlled-access mutations in the Data Portal

- Added a new data access notification to remind logged-in users with access to controlled data that they need to follow their data use agreement. The message is fixed at the top of the Portal.
- Added the ability to search for previous versions of files. If the user enters the UUID of a previous version that cannot be found, the Portal returns the UUID of the latest version available.
- Renamed the Data Category for "Raw Sequencing Data" to "Sequencing Reads" throughout the portal where this appears, to be consistent with the Data Dictionary.
- Added a link in the Portal footer to the GDC support page.

Bugs Fixed Since Last Release

- Fixed bug where Survival Plot button never stops loading if plotting mutated vs. non-mutated cases for a single Gene.
- Fixed inconsistent button styling when downloading controlled Downstream Analyses Files from File Entity page.
- Removed unnecessary Survival column from Arrange Columns button on Case Entity, Gene Entity pages.
- Removed unnecessary whitespace from pie charts on Repository page.
- Added missing File Size unit to Clinical Supplement File, Biospecimen Supplement File tables on Case Entity page.
- Fixed bug where clicking on Case Counts in Projects Graph tab was going to the Repository Files tab instead of the Cases tab.
- Fixed bug where the counts shown beside customer filters on the Repository Cases tab were not updating when filtering on other facets.
- Fixed bug where clicking the of Affected Cases denominator on the Gene page's Most Frequent Somatic Mutations table displayed an incorrect number of Cases.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.18.0

- **GDC Product**: GDC Data Portal
- **Release Date**: December 18, 2018

New Features and Changes

- A new data access message has been added when downloading controlled data. Users must agree to abide by data access control policies when downloading controlled data.

- In the Mutation free-text search in Exploration, mutation display now includes the UUID, genomic location, and matched search term for easier mutation searching.
- The ability to sort on ranked columns has been made available.

Bugs Fixed Since Last Release

- In some cases, text was being cut off on the Project page visualization tab. Text is no longer cut off.
- HGNC link on Gene page broke as the source format url changed; The format was updated and the link is now functional
- In the biospecimen details on the Case page, the cart icon would disappear once clicked. It now is always visible.

Known Issues and Workarounds

- Pre-release Data Portal login is not supported on Internet Explorer or the last version of Edge (42). Edge 41 does login successfully.
- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.17.0

- **GDC Product**: GDC Data Portal
- **Release Date**: November 7, 2018

New Features and Changes

- Copy Number Variation (CNV) data derived from GISTIC results are now available in the portal:
  - View number of CNV events on a gene in a cohort in the Explore Gene table tab
  - Explore CNVs associated with a gene on the Gene Entity Page
  - Explore CNVs concurrently with mutations on the Oncogrid with new visualization

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- Custom Facet Filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations

  - SIFT and PolyPhen annotations are missing from the export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.16.0

- **GDC Product**: GDC Data Portal
- **Release Date**: September 27, 2018

New Features and Changes

- Updated Human Body Image to aggregate all current primary sites to available Major Primary Sites

Bugs Fixed Since Last Release

- Fixed link on cart download error popup
- Updated Cancer Distribution table to have dropdown menus for primary_site and disease_type
- Updated Y-axis label on `Top Mutated Cancer Genes in Selected Projects Graph`
- Updated Set Operation Image to remove stray text

Known Issues and Workarounds

- Advanced Search
  - For advanced search and custom file facet filtering there are some properties that will appear as options that are no longer supported (e.g. file_state).
- Custom facet filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the Export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers

  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.15.0

- **GDC Product**: GDC Data Portal
- **Release Date**: August 23, 2018

New Features and Changes

- File Versions are now visible in the "File Versions" section on the File Entity Page.
- "View Files in Repository" and "View Cases in Repository" button methods were updated to work faster.

Bugs Fixed Since Last Release

- Fixed warning messages that prompted users to login even when already logged in. Error warnings now correctly prompt users to reference dbGAP for data access if already signed in.
- Fixed error where you could click Go on Case ID wildcard facet before inputting any data.
- Fixed cart header to be a consistent color for the whole table.
- Fixed error where you could save a set with no name or items, which resulted in an infinite spinner.
- Fixed table width issue when FM-AD was selected as a filter.
- Updated broken help link on Advanced Query.

Known Issues and Workarounds

- Advanced Search
  - For advanced search and custom file facet filtering there are some properties that will appear as options that are no longer supported (e.g. file_state).
- Custom facet filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the Export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release 1.14.0

- **GDC Product**: GDC Data Portal
- **Release Date**: June 13, 2018

New Features and Changes

- Added new Experimental Strategies Diagnostic Slide Image, Bisulfite-Seq, ChIP-Seq, and ATAC-Seq to Case and Project entity pages.

Bugs Fixed Since Last Release

- Fixed download of clinical and biospecimen data from the Repository when Case table rows are selected.

Known Issues and Workarounds

- Custom facet filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the Export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - When user is logged in and try to download a controlled file he does not have access to, he's prompted to log in. He should be promted to request access.
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.13.0

- **GDC Product**: GDC Data Portal
- **Release Date**: May 21, 2018

New Features and Changes

- Added new image viewer functionality for viewing tissue slide images

Bugs Fixed Since Last Release

- Updated gene reference labels on gene entity page to adhere to preferred usage
- Fixed issue with user profile displaying all projects twice

Known Issues and Workarounds

- Custom facet filters
  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the Export JSON of the mutation table. They are present in the export TSV.

  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - When user is logged in and try to download a controlled file he does not have access to, he's prompted to log in. He should be promted to request access.
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.12.0

- **GDC Product**: GDC Data Portal
- **Release Date**: February 15, 2018

New Features and Changes

- Provided the ability to export clinical and biospecimen data in a TSV format from the Case, Project, Exploration, Repository and Cart pages.
- Removed from the Project entity page the sections about mutated genes, somatic mutations and affected Cases and replaced with a button "Explore data" that will open the Exploration page filtered on the project. Indeed the Exploration page provides the same information. Added a breakdown of cases per primary site for a Project entity page with multiple primary sites (e.g. FM-AD).
- Added display of coding DNA change and impacts for all the transcripts (instead of canonical transcript only) in the Mutation entity page - Consequences section. In the mutation table (e.g. in Repository), the impacts and consequences are displayed for the canonical transcript only.

Bugs Fixed Since Last Release

- Replaced the suggested set name when saving a set with selected items, e.g. for case set the suggested name is now "Custom Case selection".
- Fixed the protein viewer to indicate when there are overlapping mutations. Mousing over the dot showing multiple mutations will open a right panel with the list of all the corresponding mutations.
- Fixed Mutation entity page - Consequences table: the "Coding DNA Change" column is now populated for all the transcripts.
- Fixed download clinical and download biospecimen actions from TCGA-BRCA project.
- Fixed facet behavior that did not reset back to showing all options after pressing reset-arrow.
- Fixed error when user was trying to save a set with no value in the textbox "Save top:".
- Removed somatic mutation section from Case entity page for cases with no open-access mutation data (e.g. FM-AD or TARGET cases).
- Fixed error where a blank page appears after unselecting `Cancer Gene Census` mutation facet.
- Fixed duplicated date in sample sheet name (e.g. gdc_sample_sheet_YYYY-MM-DD_HH-MM.tsv.YYYY-MM-DD_HH-MM.tsv).
- Fixed error when annotations were not downloaded along with the file (in File entity page and Cart).

Known Issues and Workarounds

- Custom facet filters

  - Some definitions are missing from the property list when adding custom facet file or case filters.
- Visualizations
  - SIFT and PolyPhen annotations are missing from the Export JSON of the mutation table. They are present in the export TSV.
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
- Repository and Cart
  - When user is logged in and try to download a controlled file he does not have access to, he's prompted to log in. He should be promted to request access.
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.11.0

- **GDC Product**: GDC Data Portal
- **Release Date**: December 21, 2017

New Features and Changes

- Updated UI to support SIFT and Polyphen annotations
- A `Sample Sheet` can now be created which allows easy association between file names and the case and sample submitter_id
- Updated Advanced Search page to include options to `Add All Files to Cart`, `Download Manifest`, and `View X Cases in Exploration`
- Provide clear message rather than blank screen if survival plots cannot be calculated for particular cohort comparison
- Display sample_type on associated entities section on file page
- Allows for special characters in case, gene, and mutation set upload (`-, :, >, .`)

Bugs Fixed Since Last Release

- Fixed error when trying to download large number of files from the Legacy Archive cart
- Fixed number of annotations displayed in Legacy Archive for particular entities
- Replaced missing bars to indicate proportion of applicable files and cases on project entity page in Cases and File Counts by Data Category table
- Fixed project page display when projects are selected that contain no mutation data in the facet panel
- Fixed error where exporting case sets as TSV included fewer cases than the total
- Fixed error in exploration section when adding custom facets. Previously selecting 'Only show fields with values' did not result in the expected behavior
- Fixed error where number of associated entities for a file was showing an incorrect number

Known Issues and Workarounds

- Sample sheet will download with a file name including the date duplicated (e.g. gdc_sample_sheet_YYYY-MM-DD_HH-MM.tsv.YYYY-MM-DD_HH-MM.tsv)
- Custom facet filters
  - Definitions are missing from the property list when adding custom facet file or case filters

- Visualizations
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
  - In the protein viewer there may be overlapping mutations. In this case mousing over a point will just show a single mutation and the other mutations at this location will not be apparent.
- Entity page
  - On the mutation entity page, in the Consequences Table, the "Coding DNA Change" column is not populated for rows that do not correspond to the canonical mutation.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.10.0

- **GDC Product**: GDC Data Portal
- **Release Date**: November 16, 2017

New Features and Changes

- Support for uploading Case and Mutation sets in Exploration page
- Support for saving, editing, removing Case, Gene and Mutation sets in the Exploration page
- Added a Managed Sets menu where the user can see their saved sets
- Added an Analysis menu with two analyses: Set Operation and Cohort Comparison
- Added a User Profile page that shows all the projects and permissions assigned to the user: available in the username dropdown after the user logs in

Bugs Fixed Since Last Release

- Project page
  - On the project page, the Summary Case Count link should open the case tab on the Repository page - instead it opens the file page

Known Issues and Workarounds

- Custom facet filters
  - Definitions are missing from the property list when adding custom facet file or case filters
  - Selecting 'Only show fields with values' will show some fields without values in the Repository section. This works correctly under the Exploration section.
- Visualizations
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
  - In the protein viewer there may be overlapping mutations. In this case mousing over a point will just show a single mutation and the other mutations at this location will not be apparent.
- Entity page

  - On the mutation entity page, in the Consequences Table, the "Coding DNA Change" column is not populated for rows that do not correspond to the canonical mutation.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.9.0

- **GDC Product**: GDC Data Portal
- **Release Date**: October 24, 2017

New Features and Changes

- Support for projects with multiple primary sites per project
- Support for slides that are linked to `sample` rather than `portion`

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- Visualizations
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
  - In the protein viewer there may be overlapping mutations. In this case mousing over a point will just show a single mutation and the other mutations at this location will not be apparent.
- Project page
  - On the project page, the Summary Case Count link should open the case tab on the Repository page - instead it opens the file page
- Entity page
  - On the mutation entity page, in the Consequences Table, the "Coding DNA Change" column is not populated for rows that do not correspond to the canonical mutation.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.

  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.8.0

- **GDC Product**: GDC Data Portal
- **Release Date**: August 22, 2017

New Features and Changes

Major features/changes:

- A feature that links the exploration and repository pages was added. For example:
  - In the exploration page, cases with a specific mutation could be selected. This set could then be linked to the repository page to download the data files associated with these cases.
  - In the repository menu, the user can select cases associated with specific files. The set could then be linked to exploration page to view the variants associated with this set of cases.
- Users can now upload a custom gene list to the exploration page and leverage the GDC search and visualization features for cases and variants associated with the gene set.
- Filters added for the gene entity page. For example:
  - Clicking on a mutated gene from the project page will display mutations associated with the gene that are present in this project (filtered protein viewer, etc.).
  - Clicking on a mutated gene from the exploration page will display the mutations associated with the gene filtered by additional search criteria, such as "primary site is Kidney and mutation impact is high".
- UUIDs are now hidden from tables and charts to simplify readability. The UUIDs can still be exported and viewed in the tables using the "arrange columns" feature. In the mutation table, UUIDs are automatically exported.
- Mutation entity page - one consequence per transcript is shown (10 rows by default) in the consequence table. The user should display all rows before exporting the table.

Bugs Fixed Since Last Release

- Exploration
  - Combining "Variant Caller" mutation filter with a case filter will display incorrect counts in the mutation facet. The number of mutations in the resulting mutation table is correct.
  - Mutation table: it is difficult to click on the denominator in "#Affected Cases in Cohort" column displayed to the left side of the bar. The user should click at a specific position at the top of the number to be able to go to the corresponding link.

Known Issues and Workarounds

- Visualizations
  - Data Portal graphs cannot be exported as PNG images in Internet Explorer. Graphs can be exported in PNG or SVG format from Chrome or Firefox browsers . Internet Explorer does not display chart legend and title when re-opening previously downloaded SVG files, the recommendation is to open downloaded SVG files with another program.
  - In the protein viewer there may be overlapping mutations. In this case mousing over a point will just show a single mutation and the other mutations at this location will not be apparent.
- Project page
  - On the project page, the Summary Case Count link should open the case tab on the Repository page - instead it opens the file page
- Entity page
  - On the mutation entity page, in the Consequences Table, the "Coding DNA Change" column is not populated for rows that do not correspond to the canonical mutation.
- Repository and Cart

  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.6.0

- **GDC Product**: GDC Data Portal
- **Release Date**: June 29, 2017

New Features and Changes

There was a major new release of the GDC Data Portal focused on Data Analysis, Visualization, and Exploration (DAVE). Some important new features include the following:

- New visual for the Homepage: a human body provides the number of Cases per Primary Site with a link to an advanced Cancer Projects search
- The Projects menu provides the Top 20 Cancer Genes across the GDC Projects and the Case Distribution per Project
- A new menu "Exploration" is an advanced Cancer Projects search which provides the ability to apply Case, Gene, and Mutation filters to look for:
  - List of Cases with the largest number of Somatic Mutations
  - The most frequently mutated Genes
  - The most frequent Variants
  - Oncogrid view of mutation frequency
- Visualizations are provided across the Project, Case, Gene and Mutation entity pages:
  - List of most frequently mutated genes and most frequent variants
  - Survival plots for patients with or without specific variants
  - Survival plots for patients with or without variants in specific genes
  - Lollipop plots of mutation frequency across protein domains
- Links to external databases (COSMIC, dbSNP, Uniprot, Ensembl, OMIM, HGNC)
- Quick Search for Gene and Mutation entity pages
- The ability to export the current view of a table in TSV
- Retired GDC cBioPortal

For detailed updates please review the Data Portal User Guide.

Bugs Fixed Since Last Release

- BAM Slicing dialog box does not disappear automatically upon executing the BAM slicing function. The box can be closed manually.
- Very long URLs will produce a 400 error. Users may encounter this after clicking on "source files" on a file page where the target file is derived from hundreds of other files such as for MAF files.
- If bam slicing produces an error pop-up message it will be obscured behind the original dialog box.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - Exporting large tables in the Data Portal may produce a 500 error. Filtering this list to include fewer cases or files should eliminate the error

Known Issues and Workarounds

- New Visualizations
  - Cannot export Data Portal graphs in PNG in Internet Explorer. Graphs can be exported to PNG or SVG from Chrome or Firefox browsers . Internet would not display chart legend and title when re-opening previously downloaded SVG files, recommendation is to open downloaded SVG files with another software.
  - In the protein viewer there may be overlapping mutations. In this case mousing over a point will just show a single mutation and the other mutations at this location will not be apparent.
- Exploration
  - Combining "Variant Caller" mutation filter with a case filter will display wrong counts in the mutation facet. The number of mutations in the result mutation table is correct.
  - Mutation table: it is difficult to click on the denominator in "#Affected Cases in Cohort" column displayed to the left side of the bar. The user should click at a specific position at the top of the number to be able to go to the corresponding link.
- Entity page
  - On the mutation entity page, in the Consequences Table, the "Coding DNA Change" column is not populated for rows that do not correspond to the canonical mutation.
- Repository and Cart
  - The annotation count in File table of Repository and Cart does not link to the Annotations page anymore. The user can navigate to the annotations through the annotation count in Repository - Case table.
- Legacy Archive
  - Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.
  - Exporting the Cart table in JSON will export the GDC Archive file table instead of exporting the files in the Cart only.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.5.2

- **GDC Product**: GDC Data Portal
- **Release Date**: May 9, 2017

New Features and Changes

- Removed link to Data Download Statistics Report
- Updated version numbers of API, GDC Data Portal, and Data Release

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- General
  - Exporting large tables in the Data Portal may produce a 500 error. Filtering this list to include fewer cases or files should eliminate the error
  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.
  - BAM Slicing dialog box does not disappear automatically upon executing the BAM slicing function. The box can be closed manually.
  - Due to preceding issue, If bam slicing produces an error pop-up message it will be obscured behind the original dialog box.

  - Very long URLs will produce a 400 error. Users may encounter this after clicking on "source files" on a file page where the target file is derived from hundreds of other files such as for MAF files. To produce a list of source files an API call can be used with the search parameter "fields=analysis.input_files.file_name".
    * Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.

Example

```
https://api.gdc.cancer.gov/files/455e26f7-03f2-46f7-9e7a-9c51ac322461?pretty=true&fields=analysis.input_files.fi
```

- Cart
  - Counts displayed in the top right of the screen, next to the Cart icon, may become inconsistent if files are removed from the server.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.4.1

- **GDC Product**: GDC Data Portal
- **Release Date**: October 31, 2016

New Features and Changes

- Added a search feature to help users select values of interest in certain facets that have many values.
- Added support for annotation ID queries in quick search.
- Added a warning when a value greater than 90 is entered in the "Age at Diagnosis" facet.
- Added Sample Type column to file entity page.
- Authentication tokens are refreshed every time they are downloaded from the GDC Data Portal.
- Buttons are inactive when an action is in progress.
- Improved navigation features in the overview chart on portal homepage.
- Removed State/Status from File and Case entity pages
- Removed the "My Projects" feature.
- Removed "Created" and "Updated" dates from clinical and biospecimen entities.

Bugs Fixed Since Last Release

- Advanced search did not accept negative values for integer fields.
- Moving from facet search to advanced search resulted in an incorrect advanced search query.
- Some facets were cut off in Internet Explorer and Firefox.

Known Issues and Workarounds

- General
  - Exporting large tables in the Data Portal may produce a 500 error. Filtering this list to include fewer cases or files should eliminate the error
  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.
  - BAM Slicing dialog box does not disappear automatically upon executing the BAM slicing function. The box can be closed manually.
  - Due to preceding issue, If bam slicing produces an error pop-up message it will be obscured behind the original dialog box.

  - Very long URLs will produce a 400 error. Users may encounter this after clicking on "source files" on a file page where the target file is derived from hundreds of other files such as for MAF files. To produce a list of source files an API call can be used with the search parameter "fields=analysis.input_files.file_name".
    * Downloading a token in the GDC Legacy Archive does not refresh it. If a user downloads a token in the GDC Data Portal and then attempts to download a token in the GDC Legacy Archive, an old token may be provided. Reloading the Legacy Archive view will allow the user to download the updated token.

Example

```
https://api.gdc.cancer.gov/files/455e26f7-03f2-46f7-9e7a-9c51ac322461?pretty=true&fields=analysis.input_files.fi
```

- Cart
  - Counts displayed in the top right of the screen, next to the Cart icon, may become inconsistent if files are removed from the server.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.3.0

- **GDC Product**: GDC Data Portal
- **Release Date**: September 7, 2016

New Features and Changes

- A new "Metadata" button on the cart page to download merged clinical, biospecimen, and file metadata in a single consolidated JSON file. **May require clearing browser cache**
- Added a banner on the Data Portal to help users find data
- Added support for "Enter" key on login button
- On the Data page, the browser will remember which facet tab was selected when hitting the "Back" button
- In file entity page, if there is a link to one single file, redirect to this file's entity page instead of a list page.

Bugs Fixed Since Last Release

- Adding a mix of open and controlled files to the cart from any Case entity pages was creating authorization issues
- Opening multiple browser tabs and adding files in those browser tabs was not refreshing the cart in other tabs.
- When user logs in from the advanced search page, the login popup does not automatically close
- When removing a file from the cart and clicking undo, GDC loses track of permission status of the user towards this file and will ask for the user to log-in again.
- Download File Metadata button produces incomplete JSON output omitting such fields as file_name and submitter_id. The current workaround includes using the API to return file metadata.
- Annotations notes do not wrap to the next line at the beginning or the end of a word, some words might be split in two lines
- Sorting annotations by Case UUID causes error

Known Issues and Workarounds

- General
  - When no filters are engaged in the Legacy Archive or Data Portal, clicking the Download Manifest button may produce a 500 error and the message "We are currently experiencing issues. Please try again later.". To avoid this error the user can first filter by files or cases to reduce the number files added to the manifest.
  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.

  - BAM Slicing dialog box does not disappear automatically upon executing the BAM slicing function. The box can be closed manually.
  - Due to preceding issue, If bam slicing produces an error pop-up message it will be obscured behind the original dialog box.
  - Very long URLs will produce a 400 error. Users may encounter this after clicking on "source files" on a file page where the target file is derived from hundreds of other files such as for MAF files. To produce a list of source files an API call can be used with the search parameter "fields=analysis.input_files.file_name".
  - On the Legacy Archive, searches for "Case Submitter ID Prefix" containing special characters are not displayed correctly above the result list. The result list is correct, however.

Example

```
https://api.gdc.cancer.gov/files/455e26f7-03f2-46f7-9e7a-9c51ac322461?pretty=true&fields=analysis.input_files.fi
```

- Cart
  - Counts displayed in the top right of the screen, next to the Cart icon, may become inconsistent if files are removed from the server.
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibility mode.

Release details are maintained in the GDC Data Portal Change Log.

Release 1.2.0

- **GDC Product**: GDC Data Portal
- **Release Date**: August 9th, 2016

New Features and Changes

- Added a retry (1x) mechanism for API calls
- Added support for ID fields in custom facets
- Added Case Submitter ID to the Annotation entity page
- Added a link to Biospeciment in the Case entity page

Bugs Fixed Since Last Release

- General.
  - Not possible to use the browser's back button after hitting a 404 page
  - 404 page missing from Legacy Archive Portal
  - Table widget icon and export JSON icon should be different
  - Download SRA XML files from the legacy archive portal might not be possible in some context
- Data and facets
  - Default values for age at diagnosis is showing 0 to 89 instead of 0 to 90
  - Biospecimen search in the case entity page does not highlight (but does bold and filter) results in yellow when title case is not followed
  - Table sorting icon does not include numbers
  - '–' symbol is missing on empty fields (blank instead), additional missing fields identified since last release. ### Known Issues and Workarounds
- General
  - When no filters are engaged in the Legacy Archive or Data Portal, clicking the Download Manifest button may produce a 500 error and the message "We are currently experiencing issues. Please try again later.". To avoid this error the user can first filter by files or cases to reduce the number files added to the manifest.
  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". This only impact users at the NIH. Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.

  - When user login from the advanced search page, the login popup does not automatically close
- Cart
  - When removing a file from the cart and clicking undo, GDC looses track of permission status of the user towards this file and will ask for the user to log-in again.
  - Counts displayed in the top right of the screen, next to the Cart icon, might get inconsistent if files are removed from the server.
  - Download File Metadata button produces incomplete JSON output omitting such fields as file_name and submitter_id. The current workaround includes using the API to return file metadata.
- Annotations
  - Annotations notes do not wrap to the next line at the beginning or the end of a word, some words might be split in two lines
  - Sorting annotations by Case UUID causes error
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibilty mode

Release details are maintained in the GDC Data Portal Change Log.

Release 1.1.0

- **GDC Product**: GDC Data Portal
- **Release Date**: June 1st, 2016

New Features and Changes

- This is a bug-fixing release, no new features were added.

Bugs Fixed Since Last Release

- General
  - Fixed 508 compliance issues.
  - Disabled download manifest action on projects without files.
  - Updated the portal to indicate to the user that his session expired when he tries to download the authentication token.
  - Unselected "My project" filter after user logs-in.
  - Fixed missing padding when query includes "My Projects".
  - Enforced "Add to cart" limitation to 10,000 files everywhere on the Data Portal.
- Tables
  - Improved usability of the "Sort" feature
  - Updated the "Add all files to cart" button to add all files corresponding to the current query (and not only displayed files).
  - Fixed an issue where Platform would show "0" when selected platform is "Affymetrix SNP 6.0".
- Data
  - Corrected default values populated when adding a custom range facet.
  - Fixed an issue preventing the user to sort by File Submitter ID in data tables.
- File Entity Page
  - Improved "Associated Cases/Biospecimen" table for files associated to a lot of cases.
  - Fixed an error when performing BAM Slicing.

Known Issues and Workarounds

- General.
  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". This only impact users at the NIH. Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.
  - Download SRA XML files from the legacy archive portal might not be possible in some context

  - Not possible to use the browser's back button after hitting a 404 page
  - 404 page missing from Legacy Archive Portal
  - Table widget icon and export JSON icon should be different
- Data and facets
  - Default values for age at diagnosis is showing 0 to 89 instead of 0 to 90
  - Biospecimen search in the case entity page does not highlight (but does bold and filter) results in yellow when title case is not followed
  - Table sorting icon does not include numbers
  - '–' symbol is missing on empty fields (blank instead), additional missing fields identified since last release.
- Cart
  - When removing a file from the cart and clicking undo, GDC looses track of permission status of the user towards this file and will ask for the user to log-in again.
  - Counts displayed in the top right of the screen, next to the Cart icon, might get inconsistent if files are removed from the server.
- Annotations
  - Annotations notes do not wrap to the next line at the beginning or the end of a word, some words might be split in two lines
- Web Browsers
  - Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.
  - Internet Explorer users are not able to use the "Only show fields with no values" when adding custom facets
  - The GDC Portals are not compatible with Internet Explorer running in compatibility mode. Workaround is to disable compatibilty mode

Release details are maintained in the GDC Data Portal Change Log.

Release 1.0.1

- **GDC Product**: GDC Data Portal
- **Release Date**: May 18, 2016

New Features and Changes

- This is a bug-fixing release, no new features were added.

Bugs Fixed Since Last Release

- Tables and Export
  - Restore default table column arrangement does not restore to the default but it restores to the previous state
- Cart and Download
  - Make the cart limit warning message more explanatory
  - In some situations, adding filtered files to the cart might fail
- Layout, Browser specific and Accessibility
  - When disabling CSS, footer elements are displayed out of order
  - If javascript is disabled html tags are displayed in the warning message
  - Layout issues when using the browser zoom in function on tables
  - Cart download spinner not showing at the proper place
  - Not all facets are expanded by default when loading the app

Known Issues and Workarounds

- General
  - If a user has previously logged into the Portal and left a session without logging out, if the user returns to the Portal after the user's sessionID expires, it looks as if the user is still authenticated. The user cannot download the token and gets an error message that would not close. The user should clear the cache to properly log out.
  - '–' symbol is missing on empty fields (blank instead)
  - Download manifest button is available for TARGET projects with 0 files, resulting in error if user clic on button

  - After successful authentication, the authentication popup does not close for Internet Explorer users running in "Compatibility View". This only impact users at the NIH. Workaround is to uncheck "Display Intranet sites in Compatibility View" in Internet Explorer options. Alternatively, refreshing the portal will correctly display authentication status.
- Data
  - When adding a custom range facet, default values are incorrectly populated
  - The portal might return incorrect match between cases and files when using field cases.samples.portions.created_datetime (custom facet or advanced search). Note: this is not a UI issue.
  - Sorting File Submitter ID option on the file tab result in a Data Portal Error
- Tables and Export
  - Table sorting icon does not include numbers
- Browsers limit the number of concurrent downloads, it is generally recommended to add files to the cart and download large number of files through the GDC Data Transfer Tool, more details can be found on GDC Website.

Release details are maintained in the GDC Data Portal Change Log.