GDC V2 User's Guide

User Acceptance Testing: July 2023

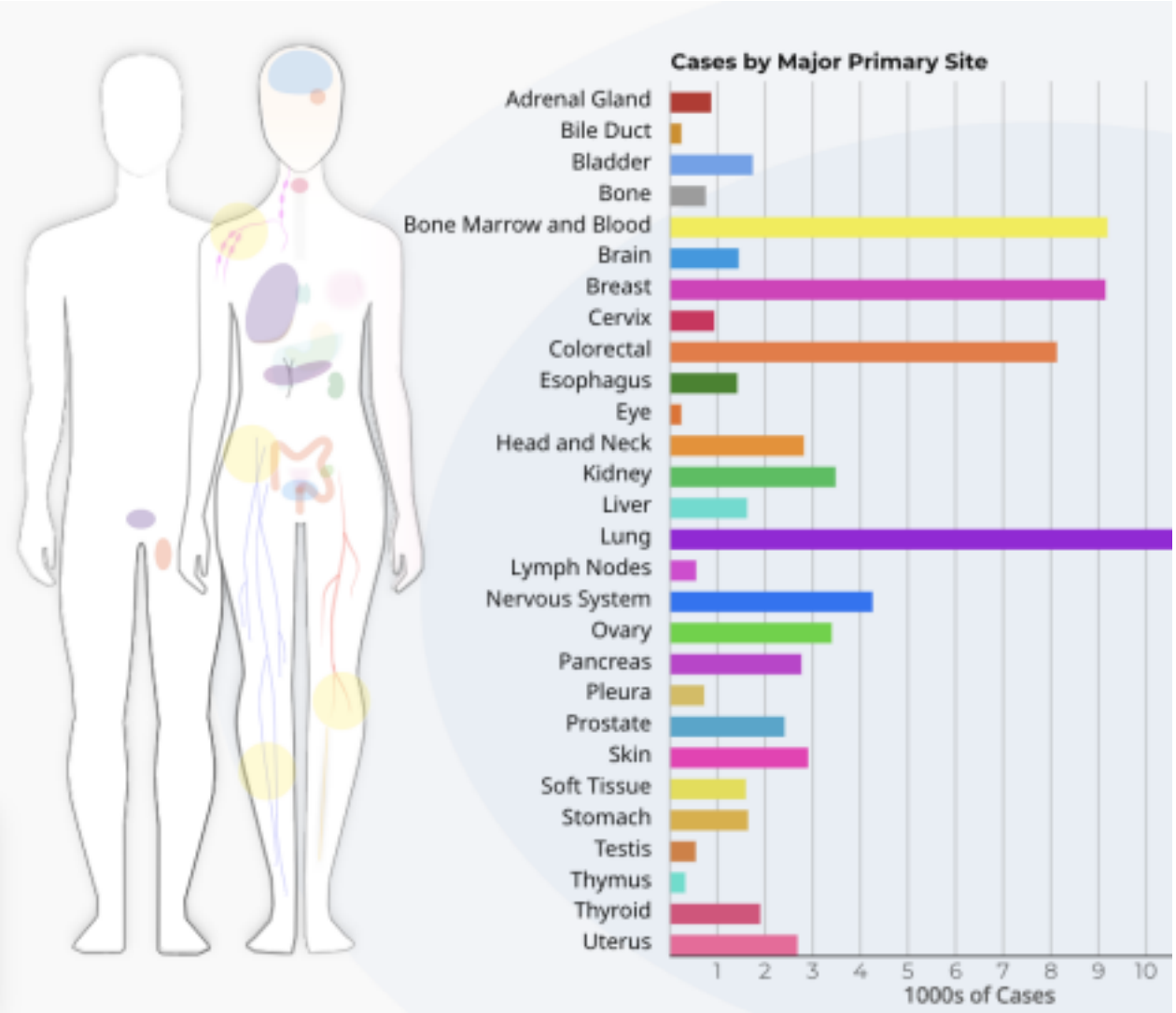


Table of Contents

Quick Start Page

Accessing the GDC Data Portal V2

First, go to GDC Portal V2.

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: links to Analysis Center page, the central hub for accessing all the tools in the GDC Data Portal. **Projects**: a shortcut to the Projects tool in the Analysis Center. The Projects tool allows exploration of all the projects within the GDC Data Portal.

Cohort Builder: a shortcut to the Cohort Builder tool in the Analysis Center. The Cohort Builder tool consists of a variety of clinical and biospecimen filters for building a custom cohort for analysis.

Repository: a shortcut to the Repository tool in the Analysis Center. The Repository tool allows exploration of files associated with a cohort.

On the right are the following features:

Browse Annotations: links to the Annotations Browser, where the user can view and search for annotations that may be of use when analyzing GDC data.

Manage Sets (Coming Soon!): allows the user to review gene and mutation sets that have been saved, upload new sets, and delete existing sets.

Cart: where data files of interest can be added for download.

Login: allows authentication for access to controlled access datasets.

GDC Applications: contains links to other GDC sites and applications.

Search: allows searching of projects, cases, files, genes, mutations, and annotations within the GDC Data Portal.

Cohorts

The GDC Data Portal V2 is a cohort-centric cancer research platform. Users can create custom cohorts of interest 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 ("New Unsaved Cohort") containing all the cases in the GDC will be automatically created for them. 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.

On the right side of the home page is a visualization that displays human figures with the available cancer primary sites for cases in the GDC Data Portal (Coming Soon!). Clicking on a colored bar or the cancer primary site will link out to the Cohort Builder tool with cases filtered with the primary site selected. Further filtering can then be performed on this cohort.

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 following sections:

Main Toolbar: contains functionality for managing and creating custom cohorts.
Query Expressions: displays the filters applied on 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. 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. Other cohorts which have been previously created can be accessed by clicking on 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 - Saves the active cohort and any changes made to it. 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. **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 - Allows for a set of cases to be imported. These can be imported as a plain text list of UUIDs or submitter_ids (barcodes).

Export Cohort - Allows for the current cohort to be exported to a file. A cohort will be exported as a list of UUIDs.

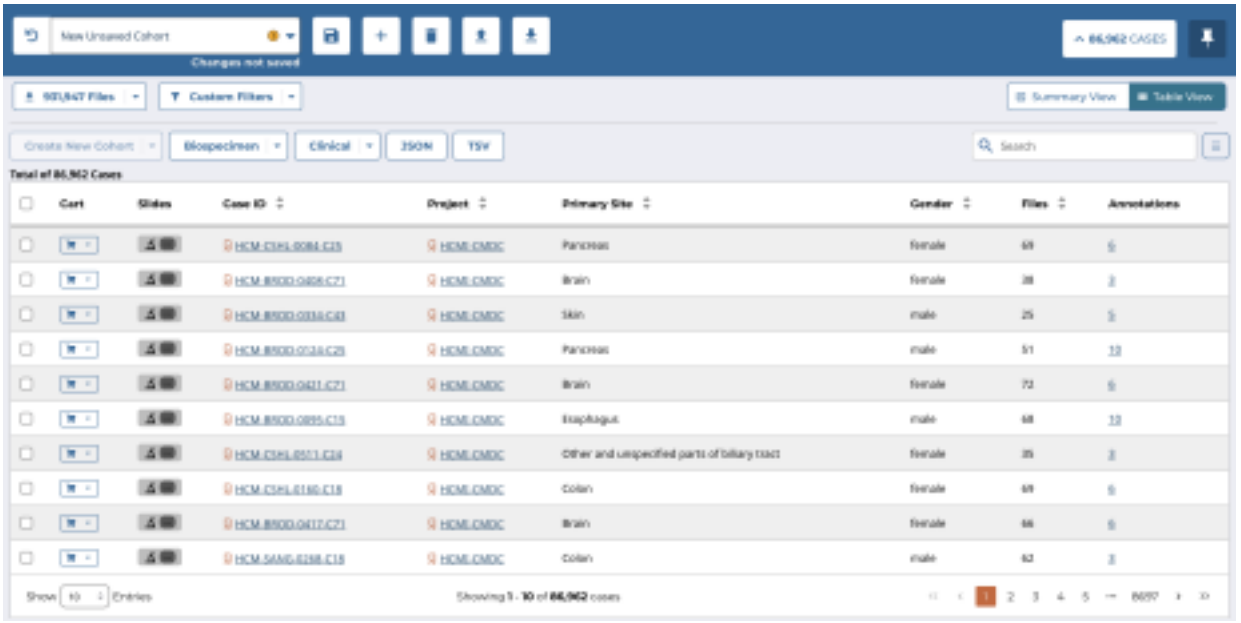
Two other buttons are located on the extreme right of the toolbar:

Expand/Collapse - Displays the number of cases associated with the current cohort. Allows access to 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.

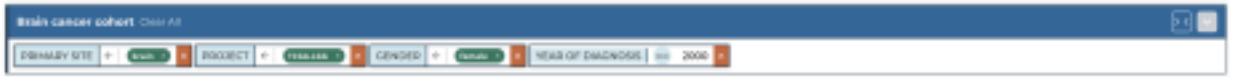
Cohort Summary Charts:

Cohort Case Table:



The case summary panel can be collapased by selecting the 'Collapse' button that replaces the 'Expand' button. **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 analysis tool is showcased within a tool 'card', which has several items related to the analysis tool such as:

A green '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

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.

File Download

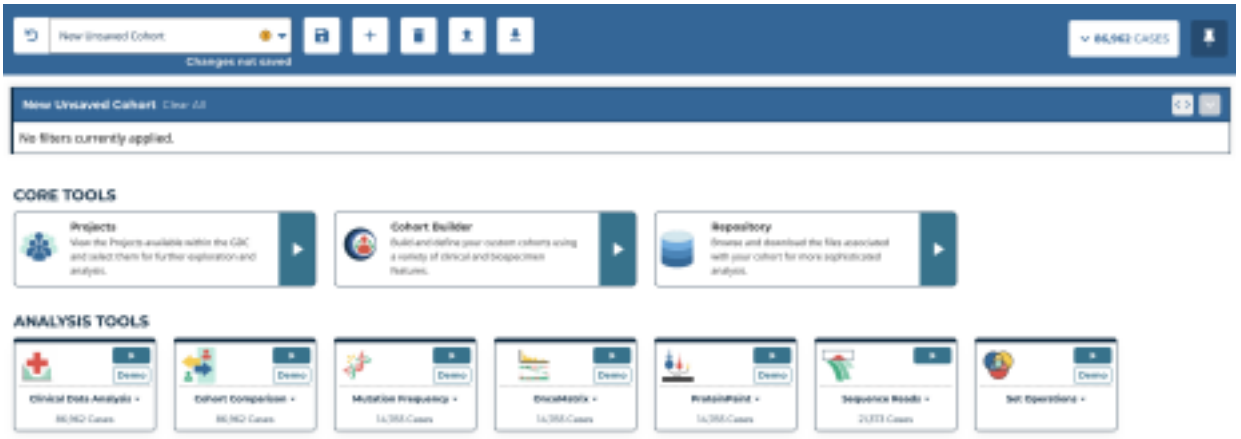
To download files from a cohort of interest:

1. Choose the Cohort Builder icon on 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 Repository icon on the header, or click on the "X" on the left of the Cohort Builder header and then click the Repository tool card. The Repository will appear on the screen.

4. The files in the Repository will automatically be filtered based on the current cohort. They can be further filtered using the available filters on the left panel.
5. Add files to the Cart, then go to the Cart to download files of interest directly or a manifest file to be used with the GDC Data Transfer Tool.

GDC Analysis Center

The Analysis Center is the central hub for accessing tools to support cohort analysis. To access the Analysis Center, click on the Analysis Center button from the GDC Data Portal's main page.



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.

A variety of cards are displayed in the Analysis Center. Each card represents an individual tool that is available in the GDC Data Portal.

Core Tools

This section contains the core GDC tools. This includes the Projects tool, the Cohort Builder, and the Repository. These can be selected for use with your current cohort by clicking on each from the Core Tools section.

Analysis Tools

The Analysis Tools section contains the analysis tools available in the Analysis Center: Clinical Data Analysis, Cohort Comparison, Mutation Frequency, OncoMatrix, ProteinPaint, Sequence Reads, and Set Operations.

These can be used by directly selecting 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.

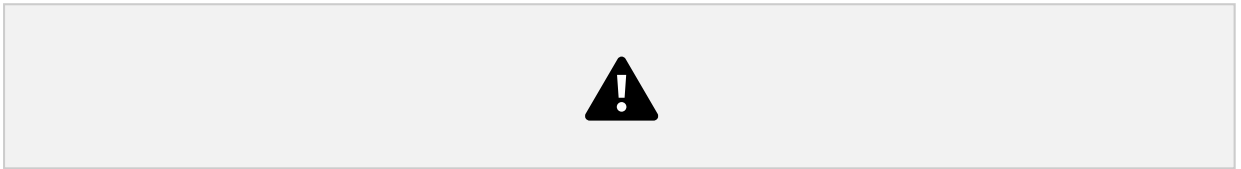
Cohort Builder

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.

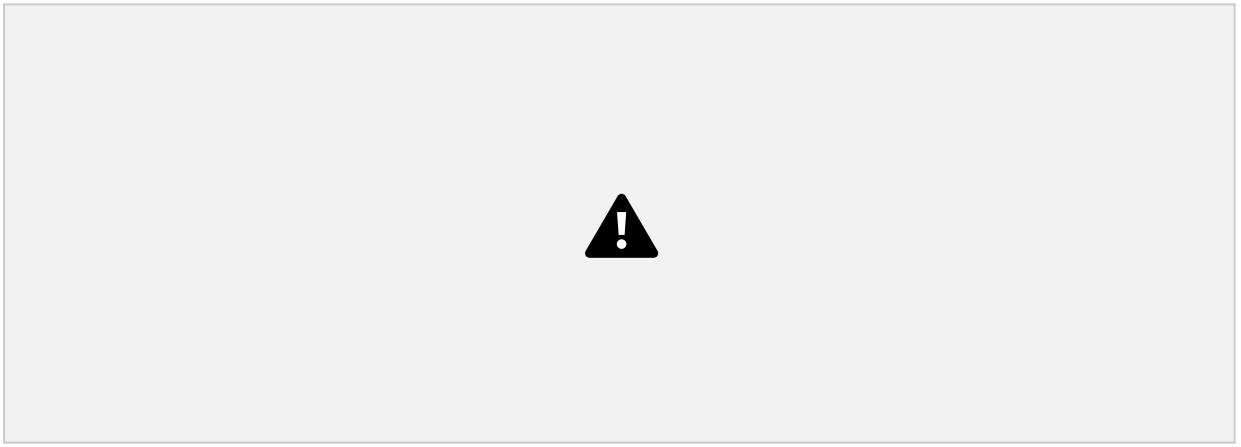
The Cohort Builder can be accessed in one of the following ways:

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.
The Cohort Builder 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

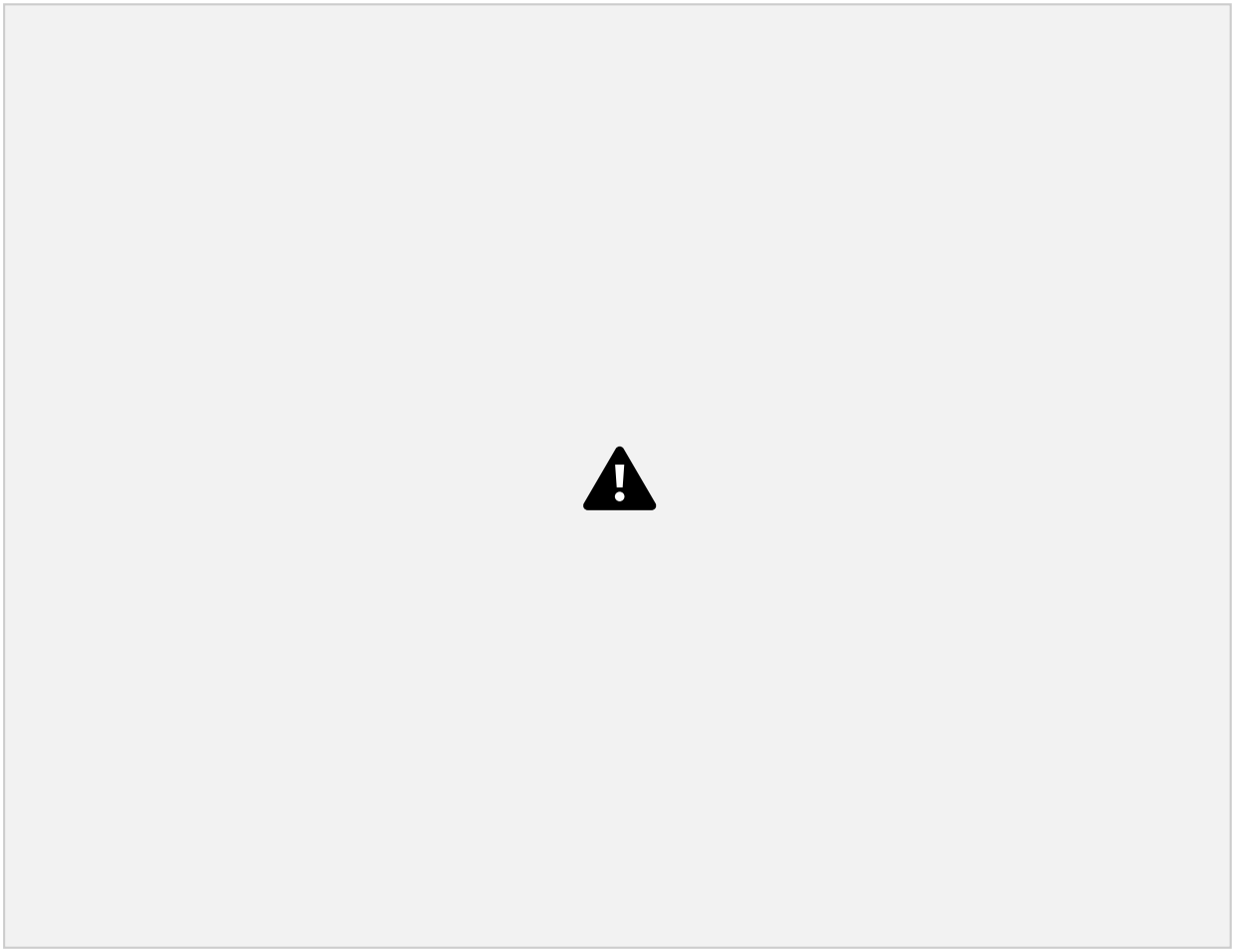
Each card within the Cohort Builder can be used to apply the corresponding filters on the current cohort. As filters are applied, they will be displayed on the Query Expressions section.

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

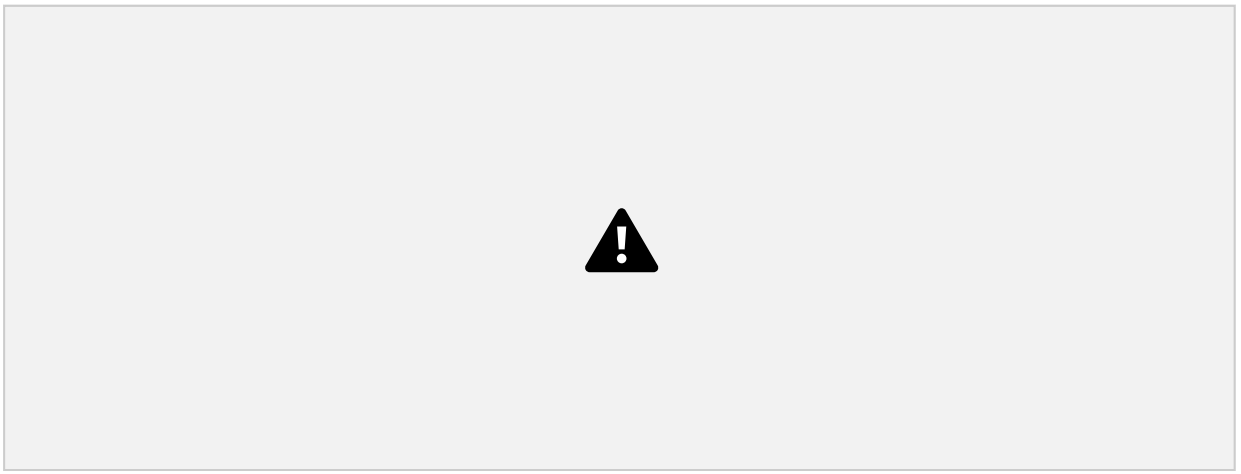
Search: the search icon can be selected to reveal or hide a search field for entering text to search within the values of the current card. This feature is only available when the values are enums.

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.

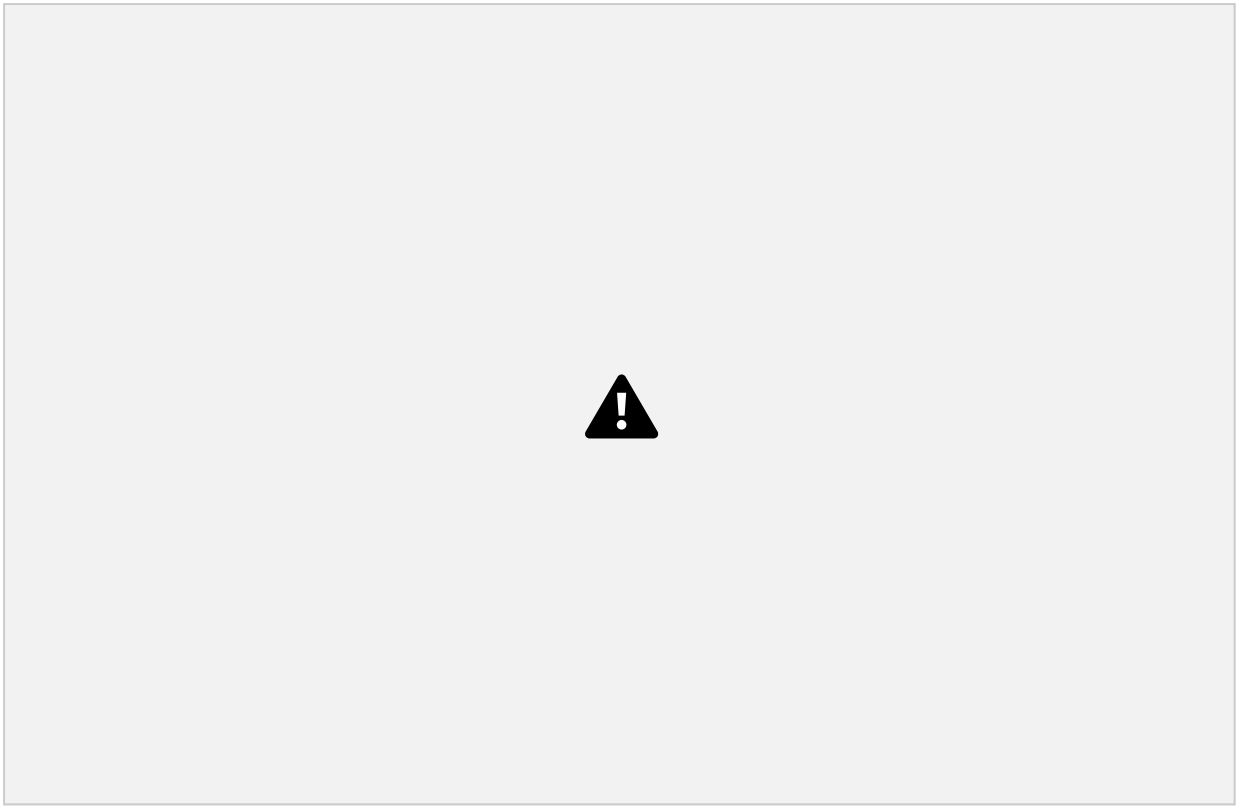
Reset Card: this button will reset any filtering that has been applied within the card.



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

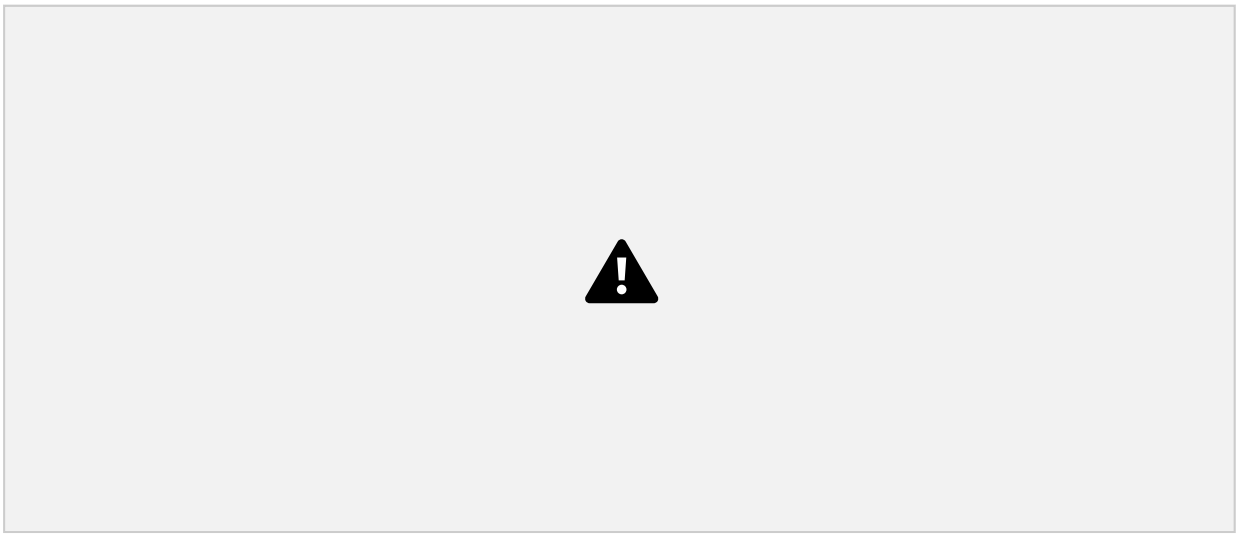


The first six (or fewer) filters are shown for each card, but can be expanded to show 20 filters at once by clicking the "+" button which also indicates the number of additional filters not in view. The expanded view can be toggled off by clicking the resulting "show less" button.

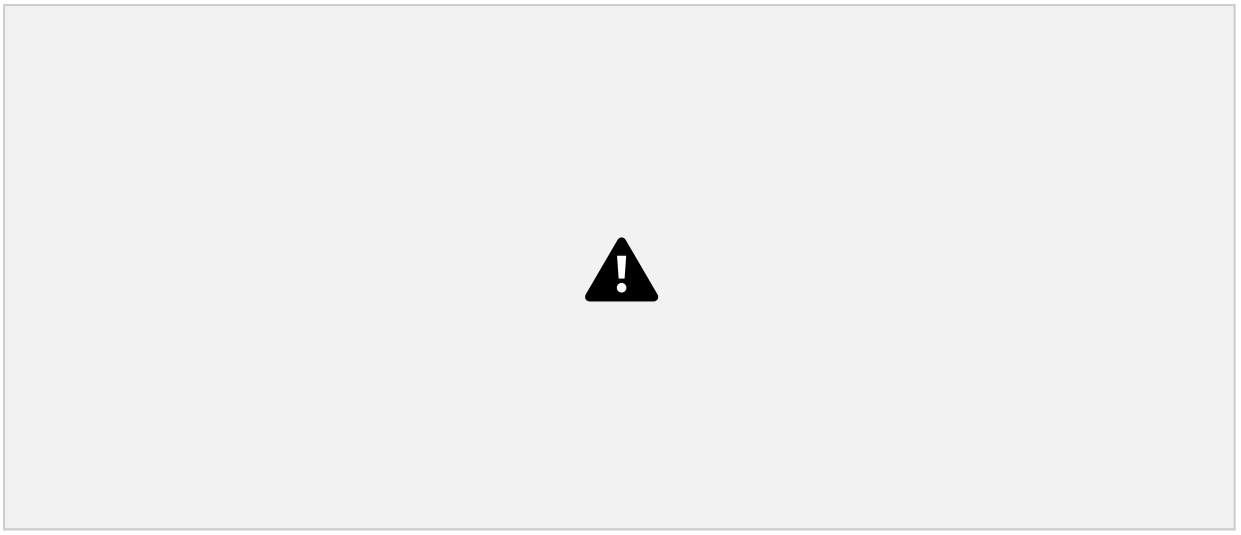


Custom Filters

If the necessary filter cannot be found within one of the categories, use the “Add a Custom Filter” button in the "Custom Filters" category to access additional filters. 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.



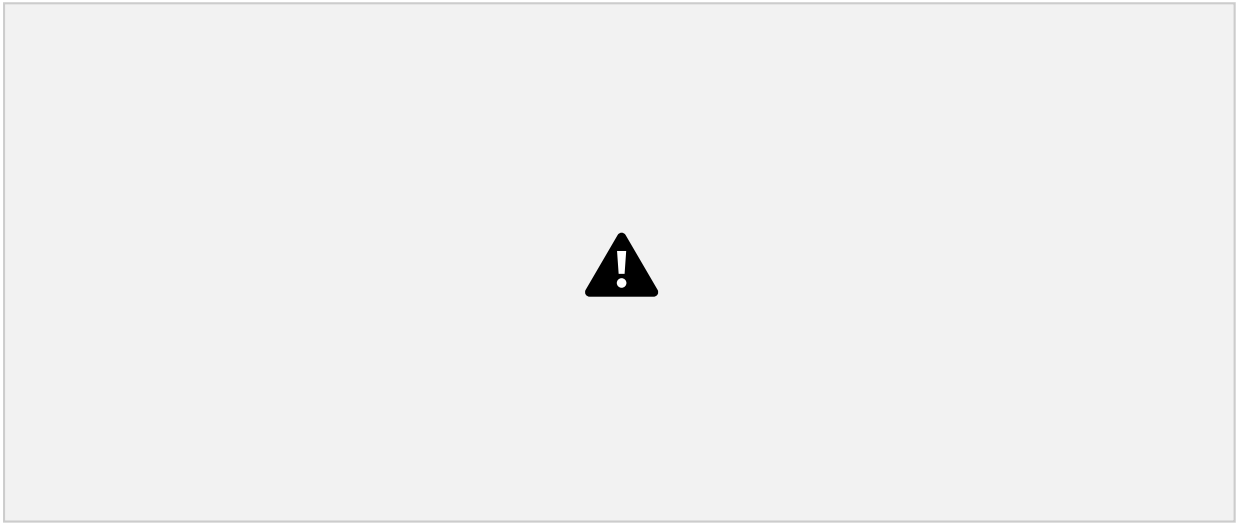
Filters that exist in the GDC but do not have any cases that have a value for the filter can also be removed from the "Custom Filters" list by selecting the "Only show properties with values" box.



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 the desired 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 all 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.

Repository

Introduction

The Repository tool is where the files associated with each case in the current cohort can be browsed and downloaded. It also offers users a variety of 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

When searching for files to download, many users will have in mind a specific cohort whose associated files they wish to download. The set of files displayed in the Repository at any time will reflect the files that are associated with the active cohort. The current active cohort can be seen in the Main Toolbar (namely on top of the page in the Analysis Center:

The active cohort can be changed by clicking the cohort name and choosing a new one from the dropdown menu. If the cohort of interest does not appear in the list, try creating a new cohort.

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

Users may only be interested in browsing through or downloading a subset of files associated with the current cohort. For this purpose, a set of commonly-used default facet cards is provided 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 follow:

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.

Experimental Strategy: Experimental strategies used for molecular characterization of the cancer.

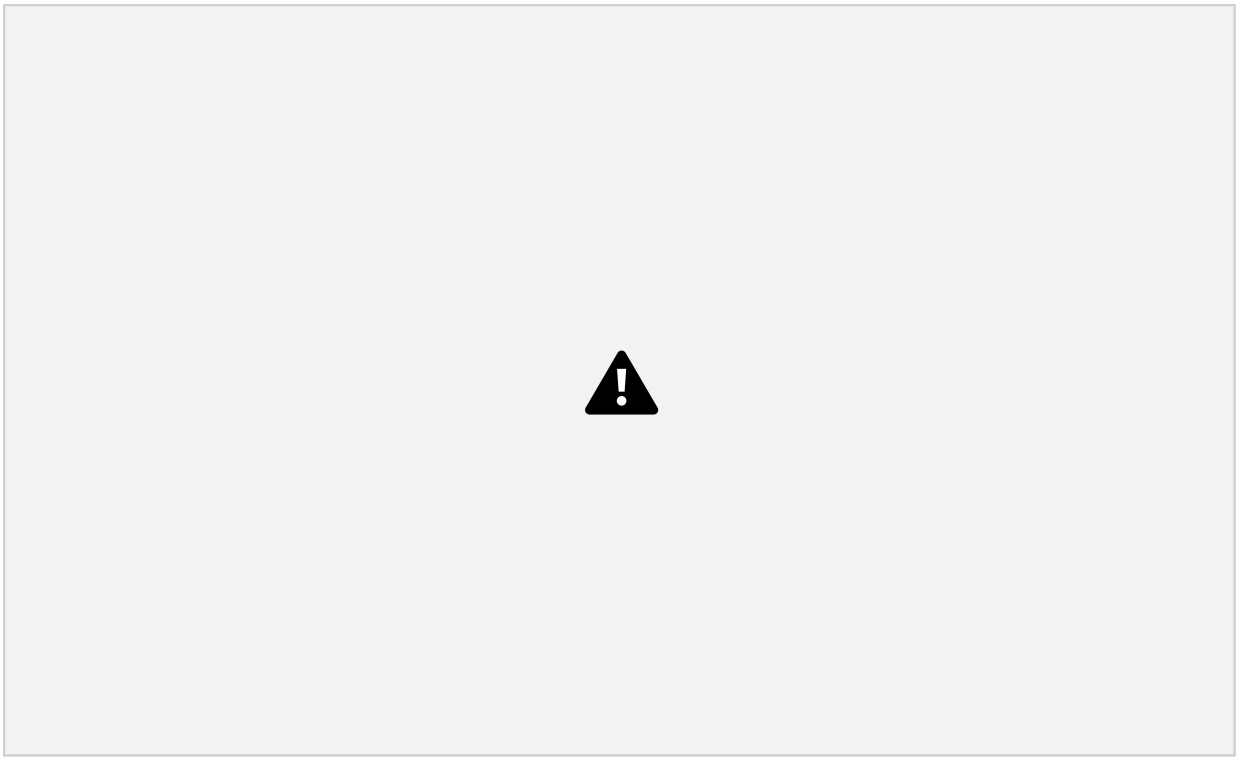
Workflow Type: Bioinformatics workflow used to generate or harmonize the data file.

Data Format: Format of the data file.

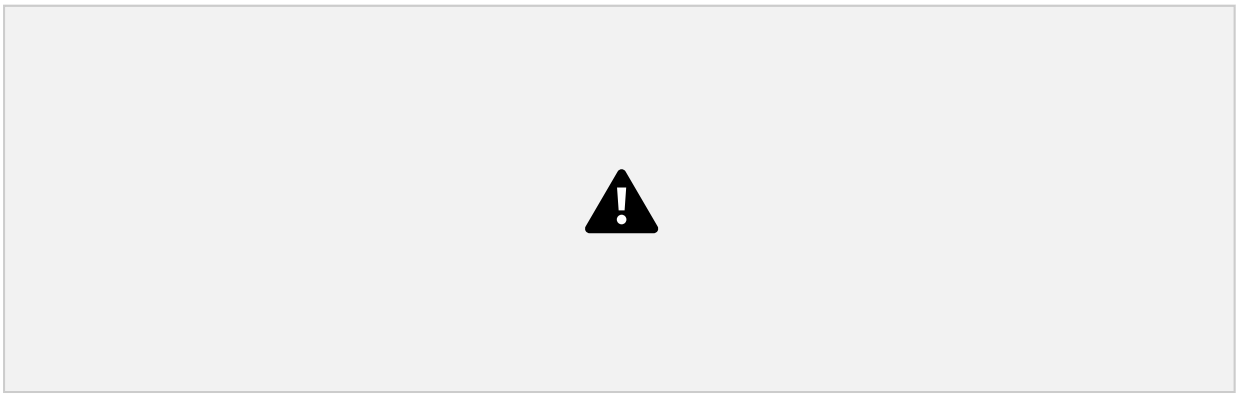
Platform: Technological platform on which experimental data was produced.
Access: Indicator of whether access to the data file is open or controlled.

Values within each facet can be sorted alphabetically by choosing the “AZ” button on the top left of each card. Alternatively, the frequency sort button next to the "Files" labeled 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 File 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.



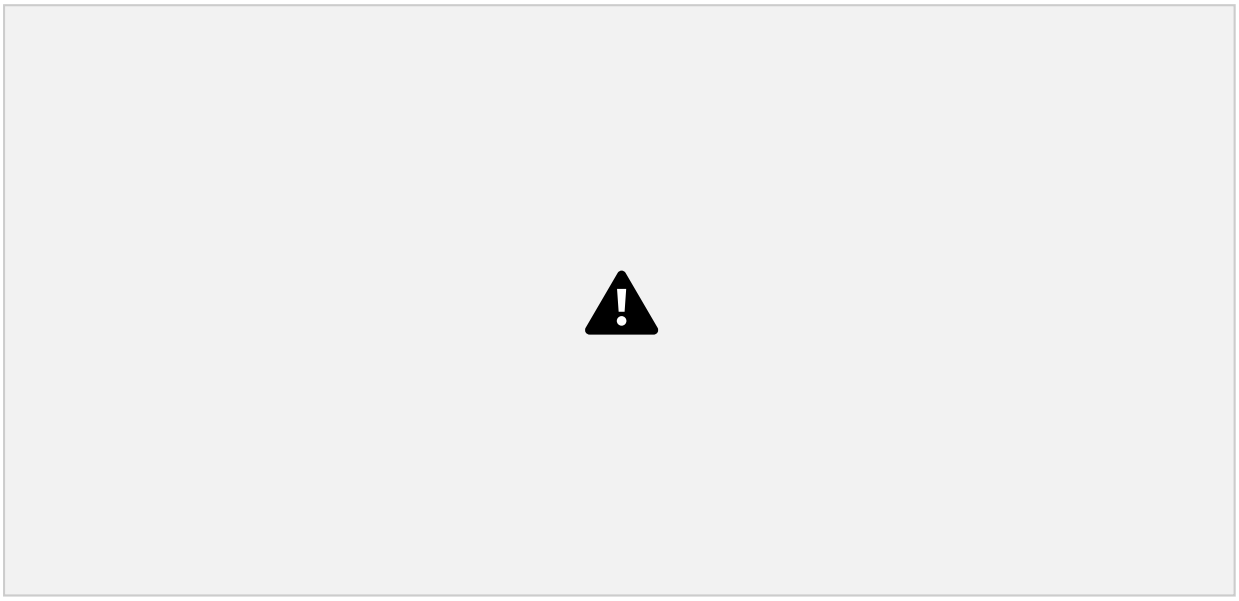
Downloading a Set of Files

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.

Adding/Removing Files to the Cart for Download

To perform a download, first, add a set of files to the Cart. This can be done using the following methods:

By clicking on the cart icon at the left to each file, it will toggle between adding to / removing the file from the cart. (Coming Soon!) 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.
JSON / TSV Buttons: These two buttons will download the files' details (file name, file size, data category, access type, etc) in JSON and TSV format, respectively
The Manifest button will generate a manifest file in text format that contains file details required for batch download (using Data Transfer Tools).
View Images: When image slides files selected, this button will create open the Slide Image Viewer, containing a collection of image files selected.



Cart

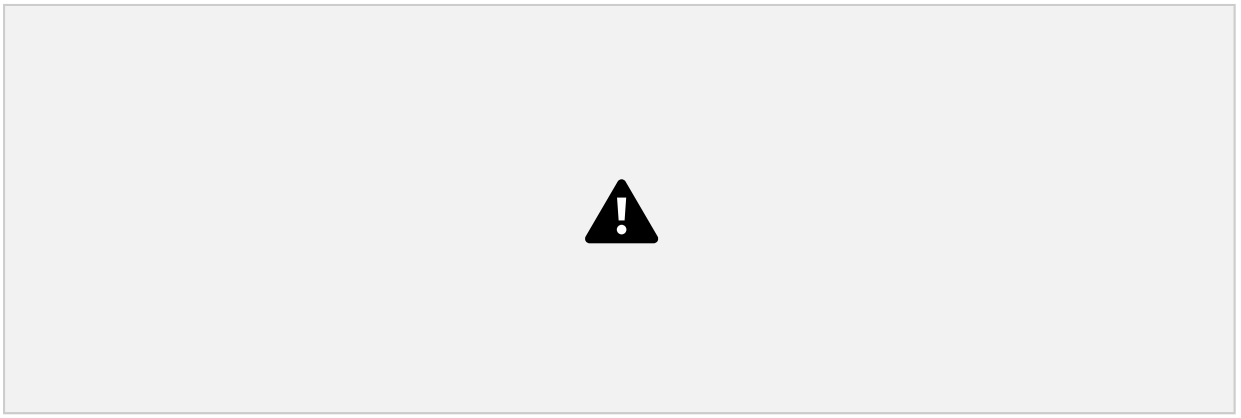
The Cart page can then be reached by clicking the Cart icon at the top right of the portal.

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.

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



Cart Items Table

The Cart Items table shows the list of all the files that were added to the Cart. The table gives the following information for each file in the cart:

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.

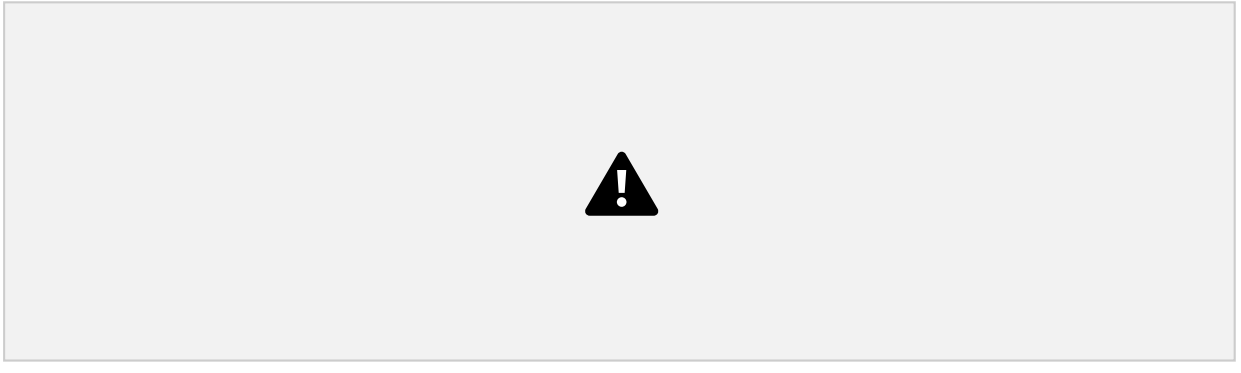
Downloading Files from the Cart

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 (Coming Soon!) - This includes all clinical information from the cases that are associated with the files (available as TSV or JSON)

Biospecimen: TSV / Biospecimen: JSON (Coming Soon!) - 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.

Projects

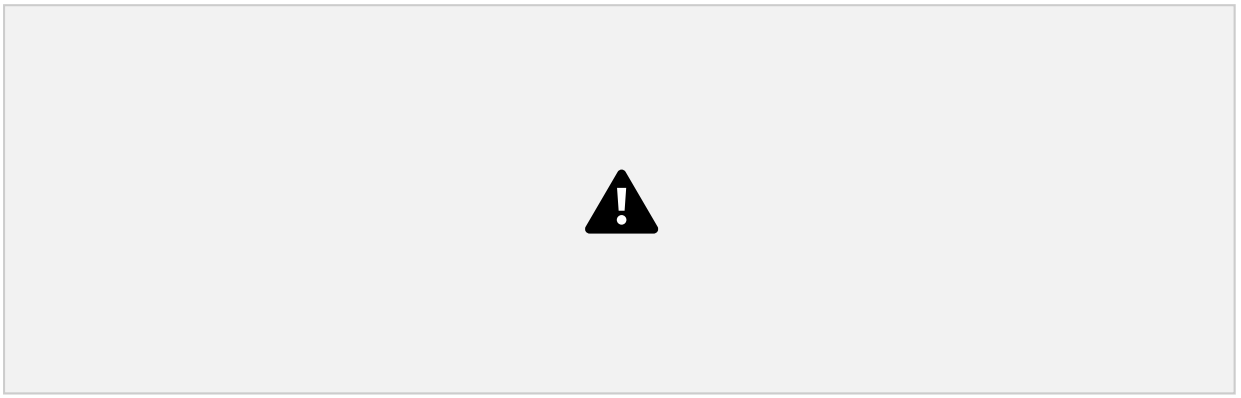
At a high level, data in the Genomic Data Commons is organized by project. Typically, a project is a specific effort to look at particular type(s) of cancer undertaken as part of a larger cancer research program. The GDC Data Portal allows users to access aggregate project-level information via the Projects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

Projects Tool

The Projects tool provides an overview of all harmonized data available in the Genomic Data Commons, 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.



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.

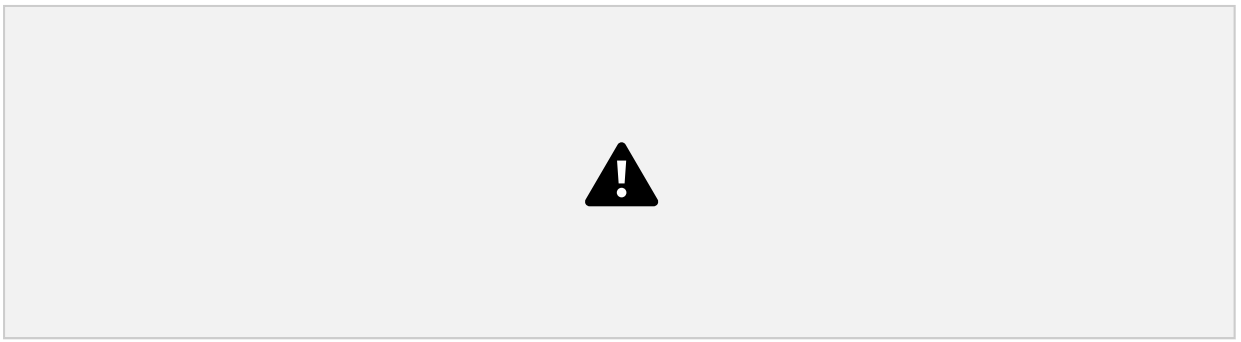
Data Category: Type of data available in the project.

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.

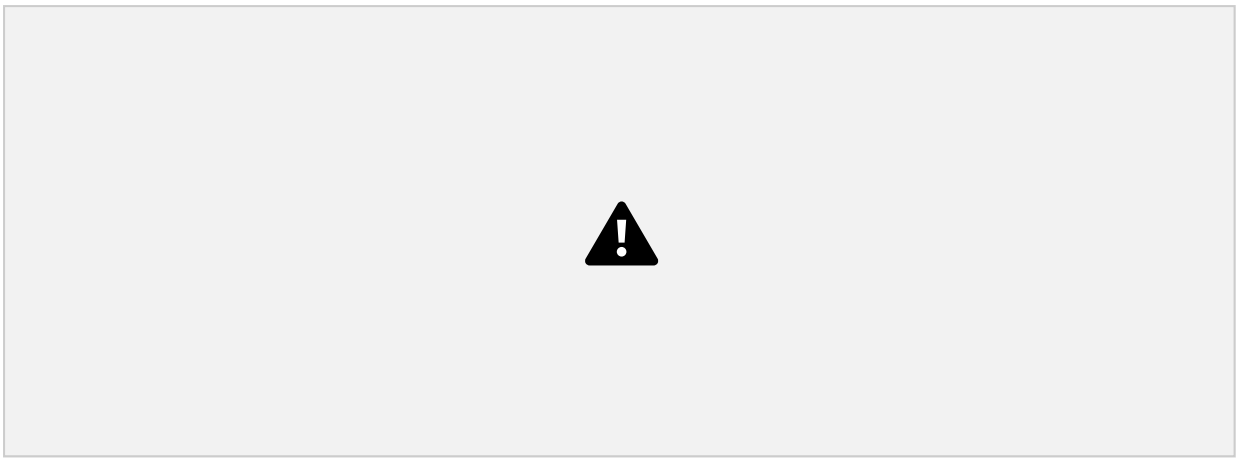
Creating Cohorts From Selected Projects

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 "Create 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.



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

Create New Cohort: Creates a new unsaved cohort consisting of all the cases in the project.

Biospecimen (Coming Soon!): Downloads biospecimen metadata associated with all cases in the project in either TSV or JSON format.

Clinical (Coming Soon!): 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.

Clinical Data Analysis

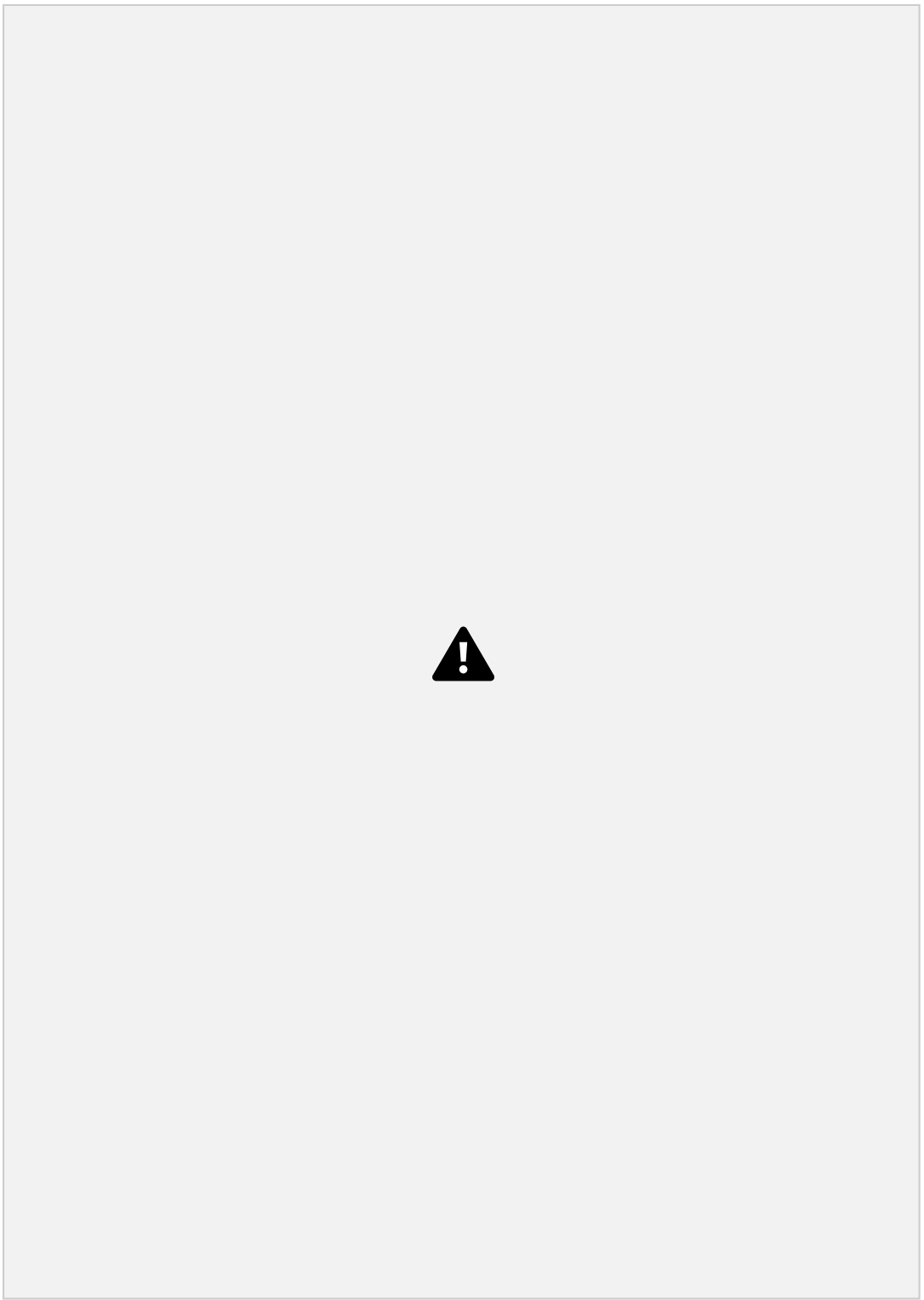
The Clinical Data Analysis tool allows for a set of customizable charts to be generated for a set of clinical attributes. Users can select which clinical fields they want to display and visualize the data using various supported plot types. The clinical analysis features include:

- Ability to select which clinical fields to display
- Examine the clinical data of each field using these visualizations:
  - Histogram
  - Survival Plot
- 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.

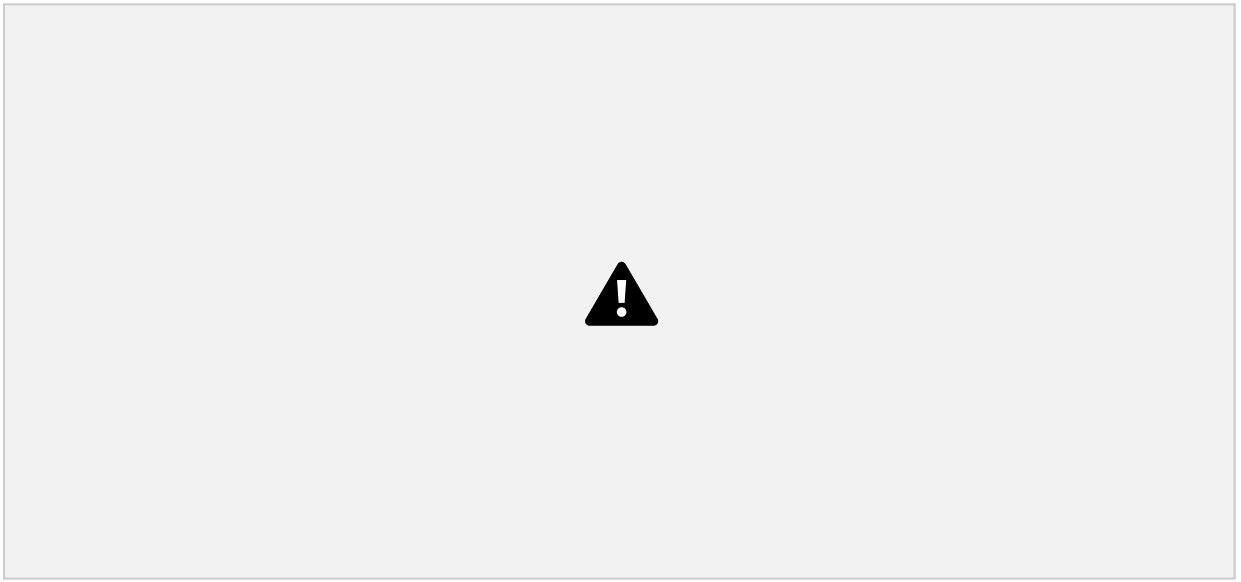
Exploring Clinical Card Visualizations

Users can explore different visualizations for each clinical field they have enabled for display. Each card supports histograms and survival 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

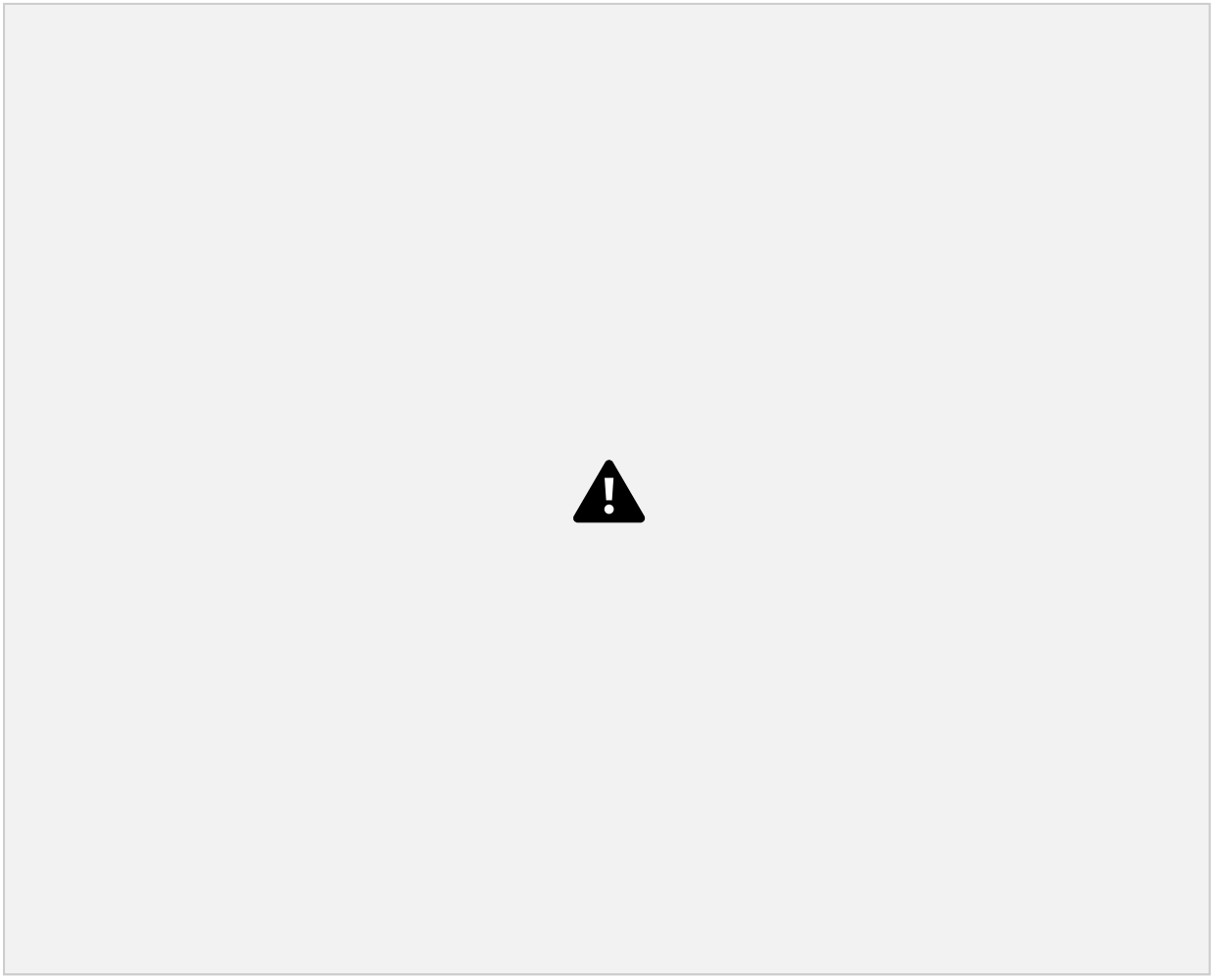


Note that the histogram plot applies to, and can be displayed for, both categorical and continuous variables.

Survival Plot

The survival 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.

Creating Custom Bins

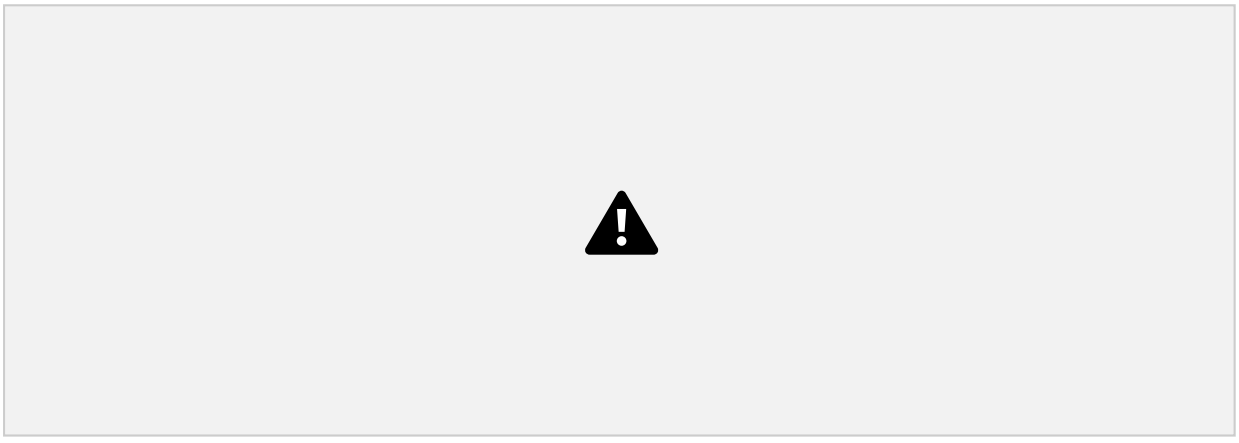
For each clinical variable, whether categorical or continuous, users can create custom bins to group the data in ways they find scientifically interesting or significant. Once saved, the bins are applied to these visualizations and they are then re-rendered:

- 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. Note that custom bins are **saved per analysis**.

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:



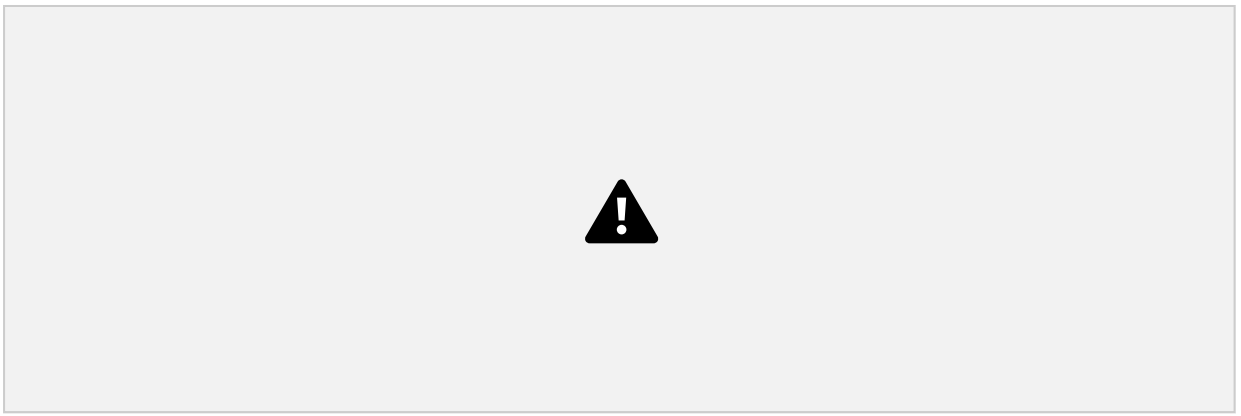
The user can:

- 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

Cohort Comparison

The "Cohort Comparison" tool displays a series of 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:

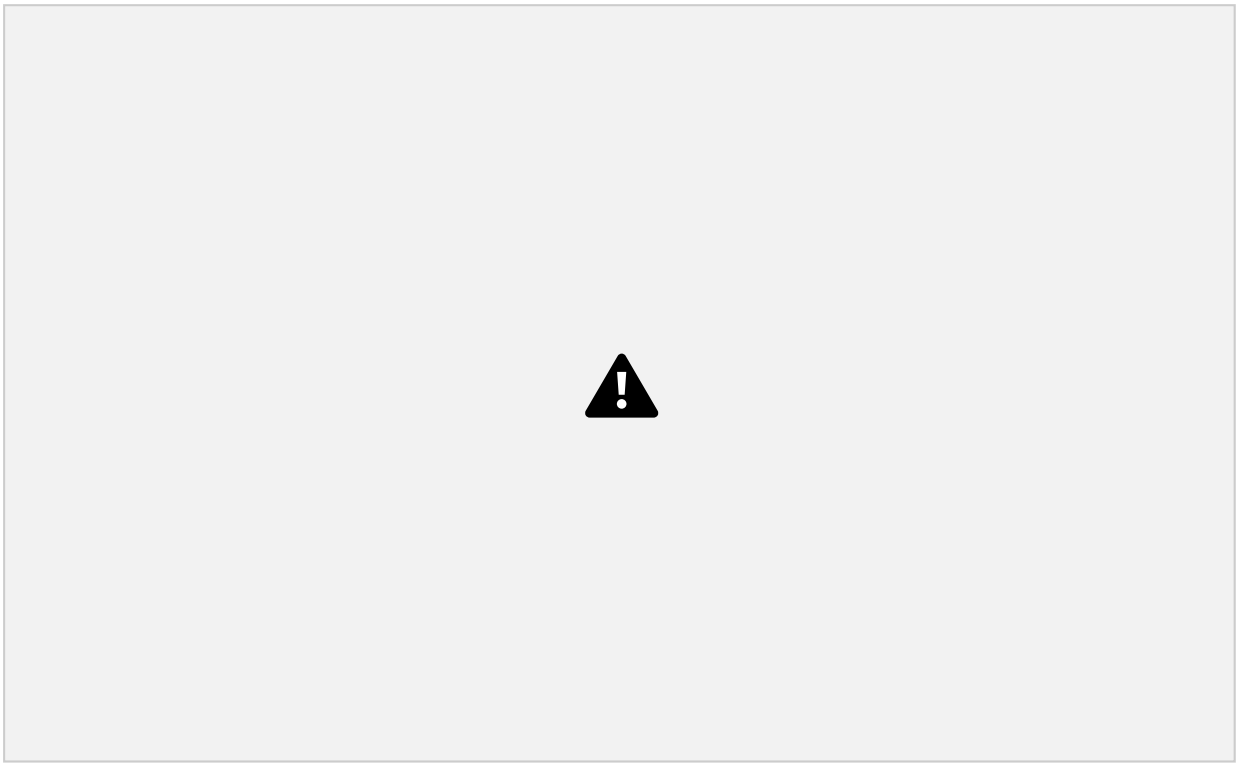
A key detailing the number of cases in each cohort and the color that represents each (blue/orange).

A Venn diagram, which shows the overlapping cases between the two cohorts.

A selectable survival plot that compares both sets with information about the percentage of represented cases.

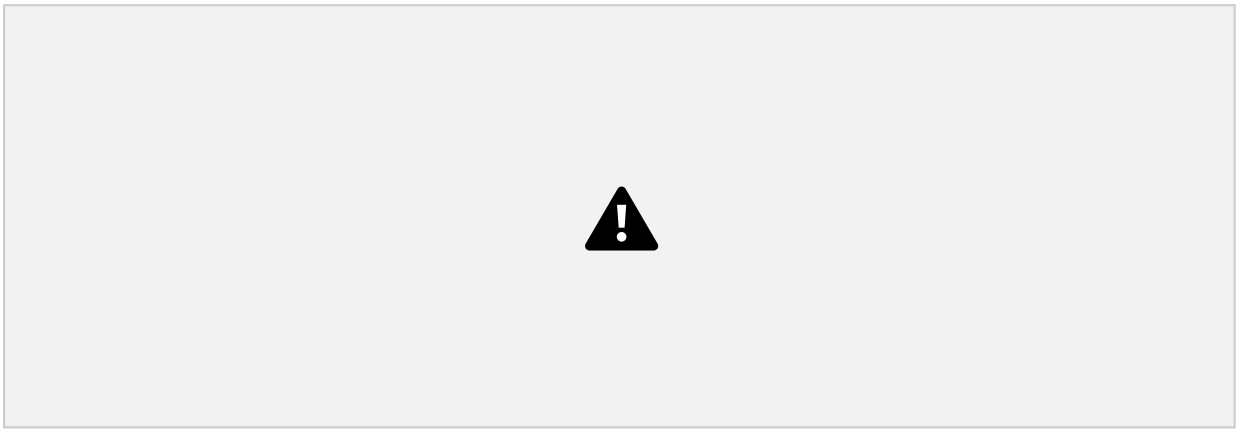
A breakdown of each cohort by selectable clinical facets with a bar graph and table. The facets included are `Vital_Status`, `Gender`, `Race`, `Ethnicity`, and `Age_at_Diagnosis`.

Additional cohorts can be created containing subsets of these two cohorts.



Mutation Frequency

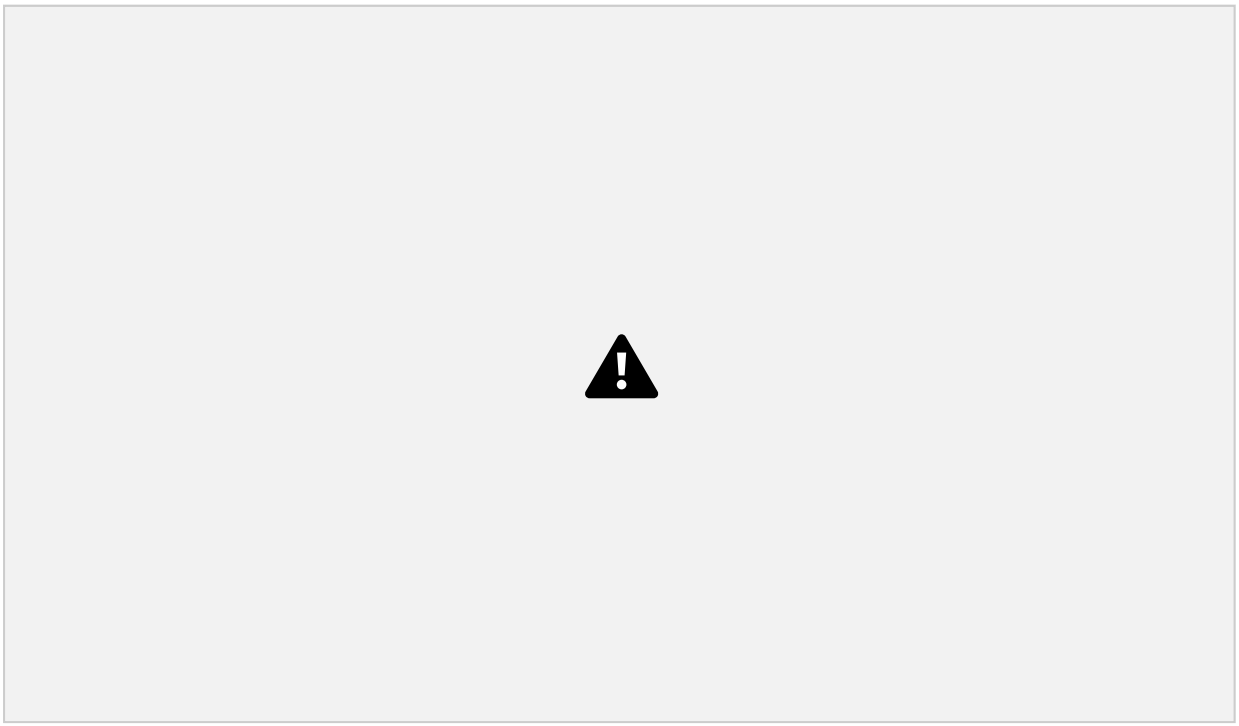
The Mutation Frequency tool visualizes the most frequently mutated genes and the most frequent somatic mutations for the active cohort. To launch the Mutation Frequency tool, click on its card from the Tools section of the Analysis Center.



This tool includes the following visualizations:

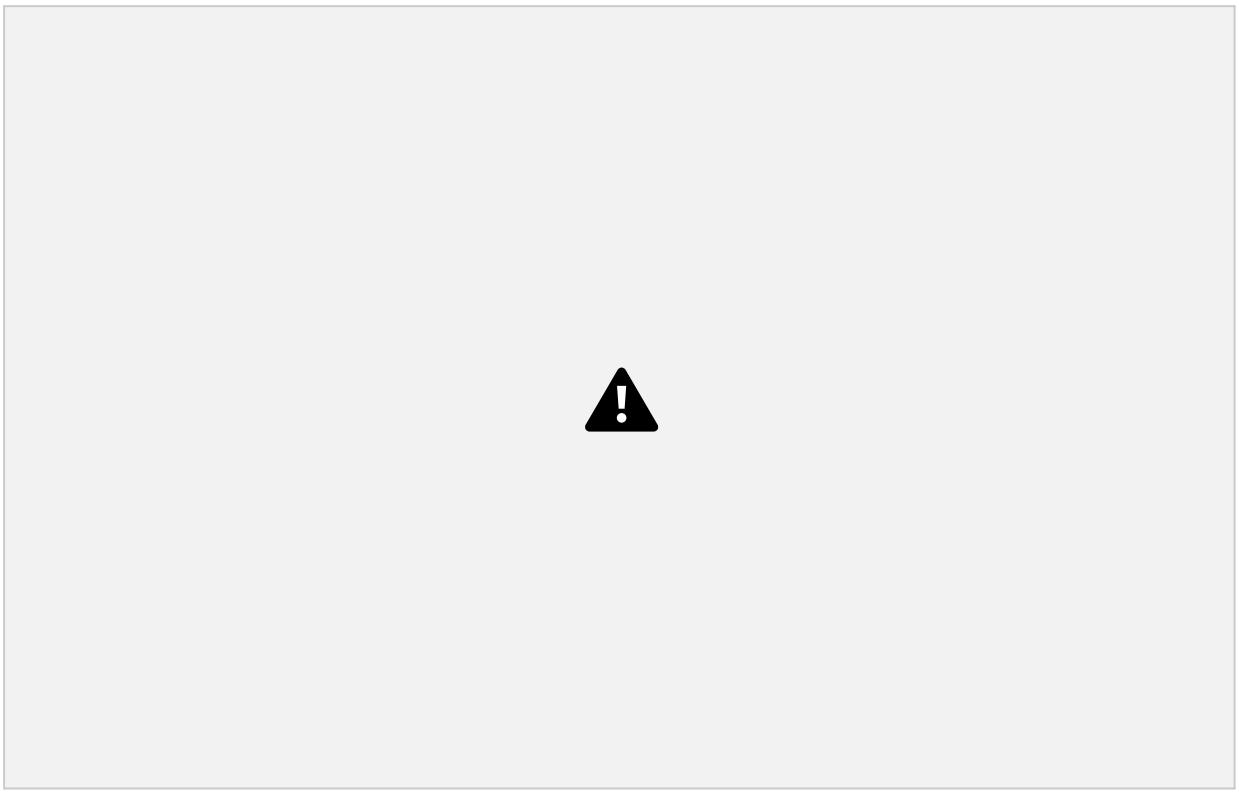
Mutated Genes Histogram

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.



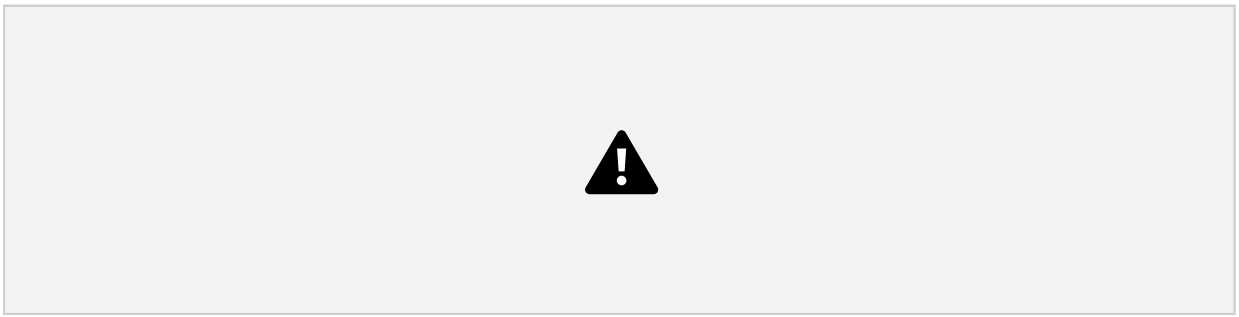
Survival Plot for Mutated Genes and Mutations

The mutation frequency survival plot is represented with a Kaplan-Meier curve for the active cohort. This can also be separated into two curves based on cases with and without a specific mutation or mutated gene. 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.



Gene/Mutation Table

The gene/mutation 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 JSON or TSV using the buttons at the top left of the table. Additional cohorts can be created using buttons located within the table.



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

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.

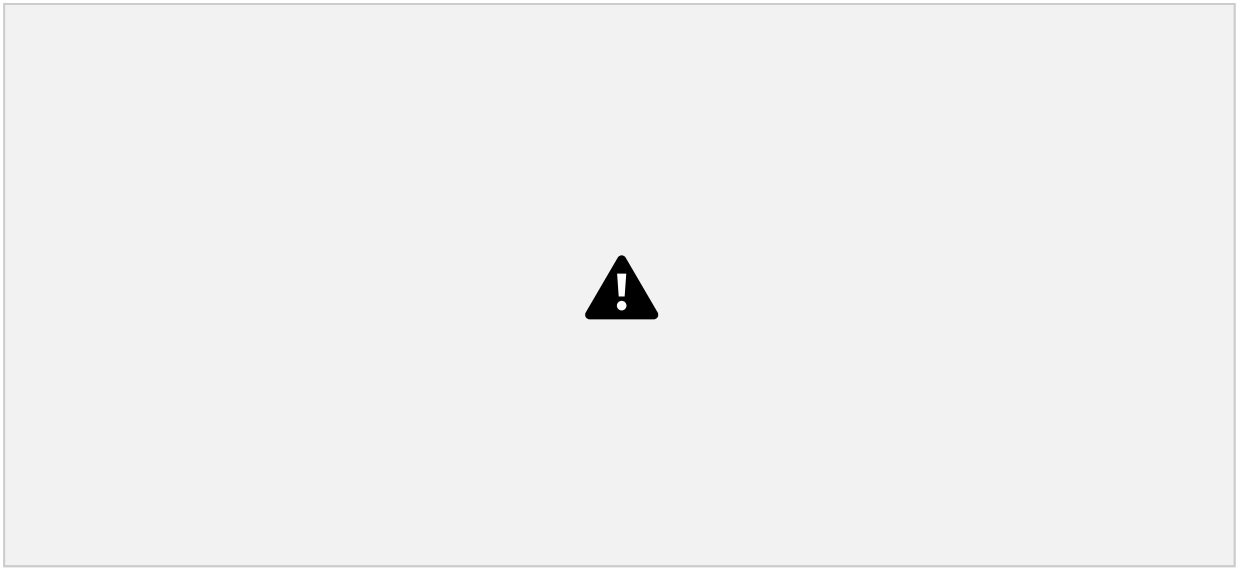
Type: A general classification of the mutation.

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.

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:



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.

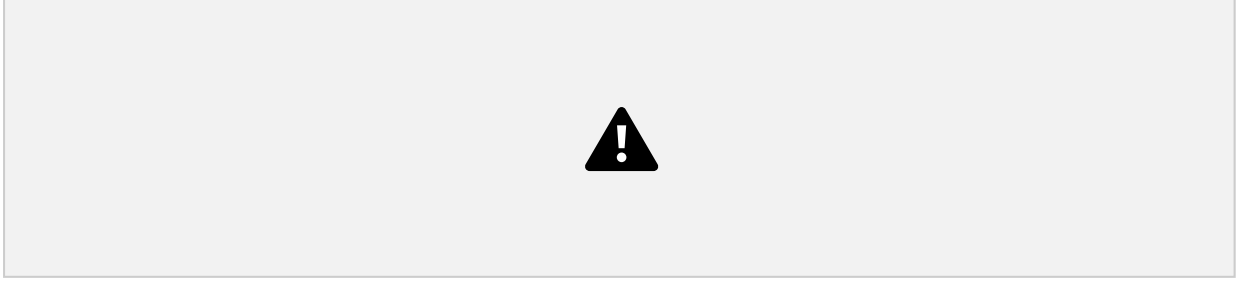
ProteinPaint Tool

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.

Accessing the Lollipop Chart

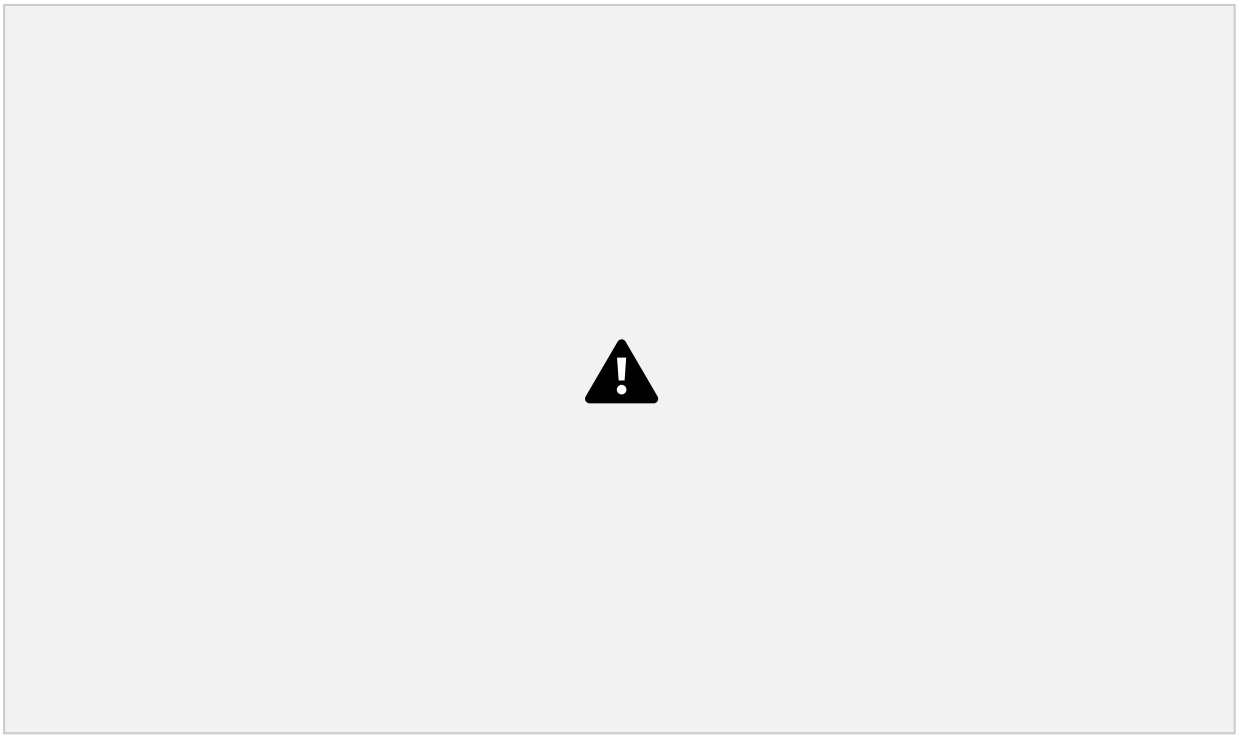
At the Analysis Center, click on the “ProteinPaint” card to launch the app.



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

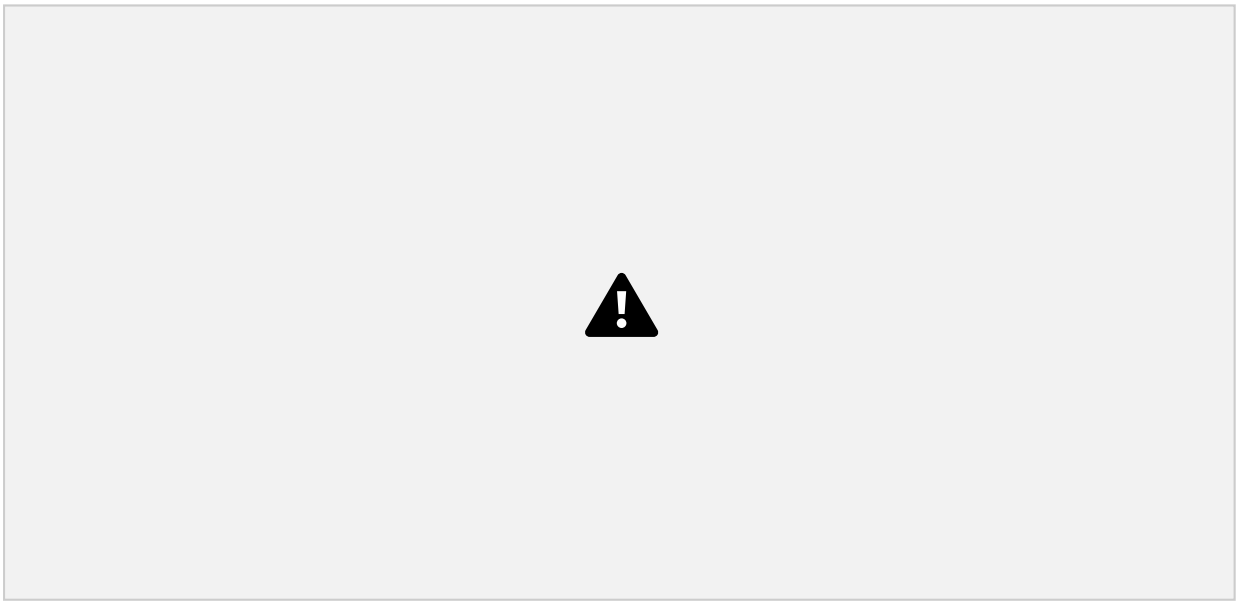
ProteinPaint Features

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.



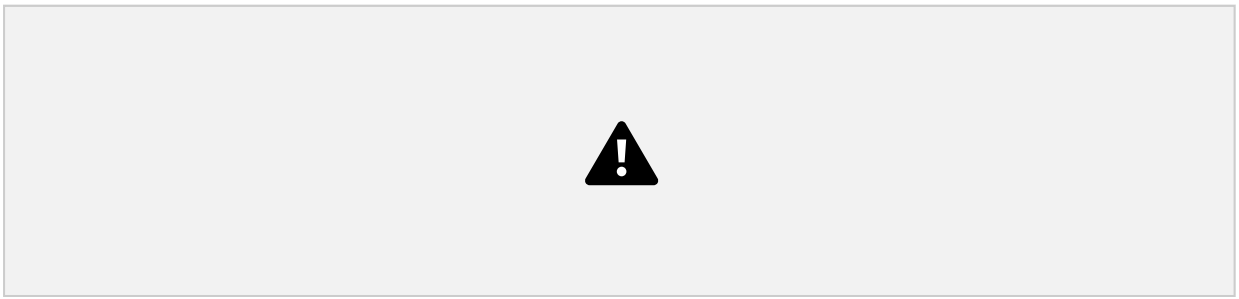
There are 3 main panels as outlined in the figure below:

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



Search Box

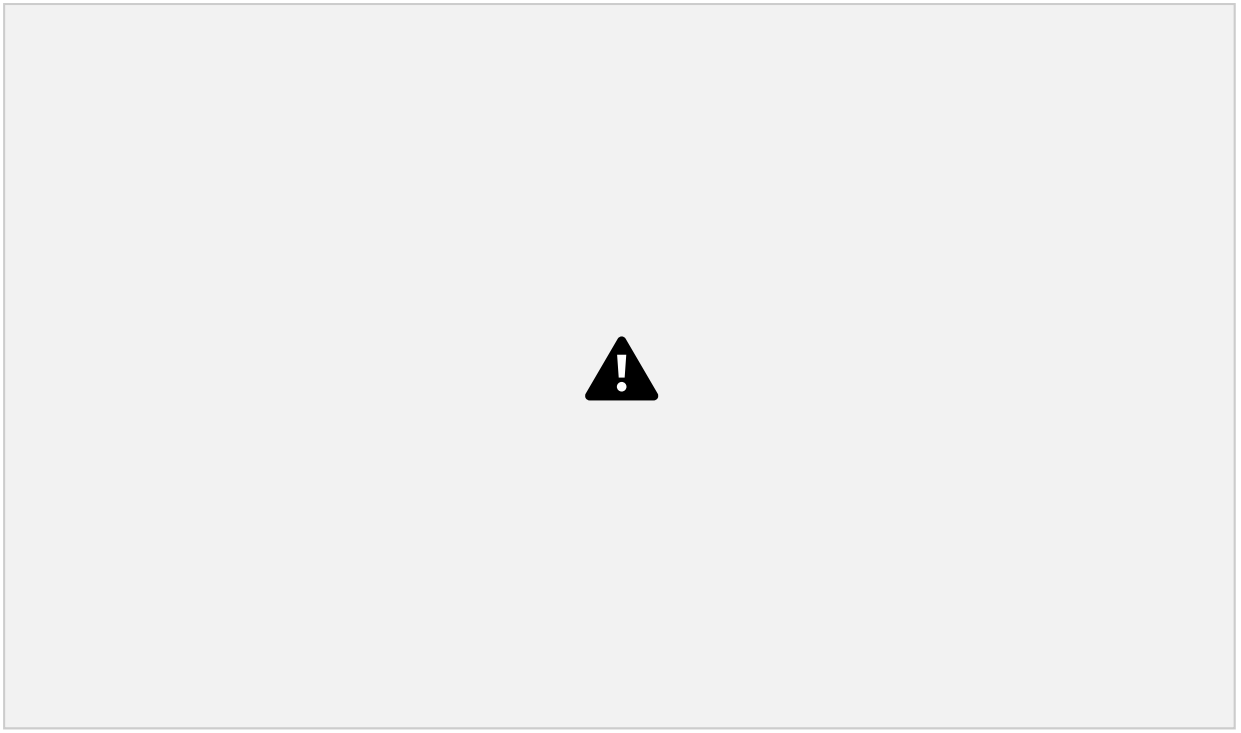
The example below uses the KRAS gene. The name of the gene (e.g., 'KRAS'), GENCODE accession no. (e.g., ENST00000311936, ENSP00000308495) or RefSeq accession (e.g., NM_004985) can be used as the search item. In case a wrong gene is entered, the search box will display an error. For gene searches only, typing a few letters reveals a menu of possible matches. Choose from either a menu option or hit enter.



Lollipop Chart Panel

Protein View

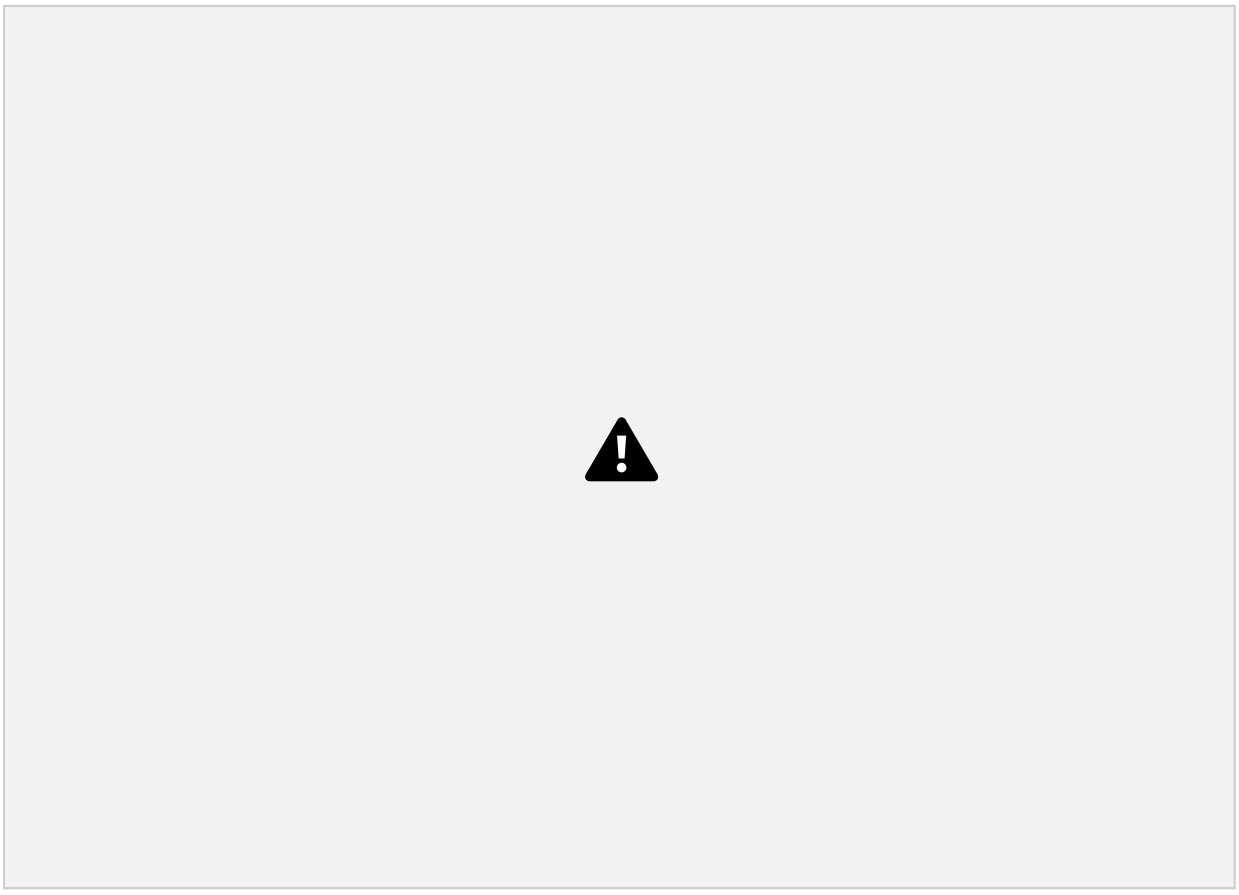
After searching for KRAS, the Protein View for the default isoform appears in a new frame. The Protein View displays the nucleotides, codons in the exon region, introns, and protein domains as shown below.



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.



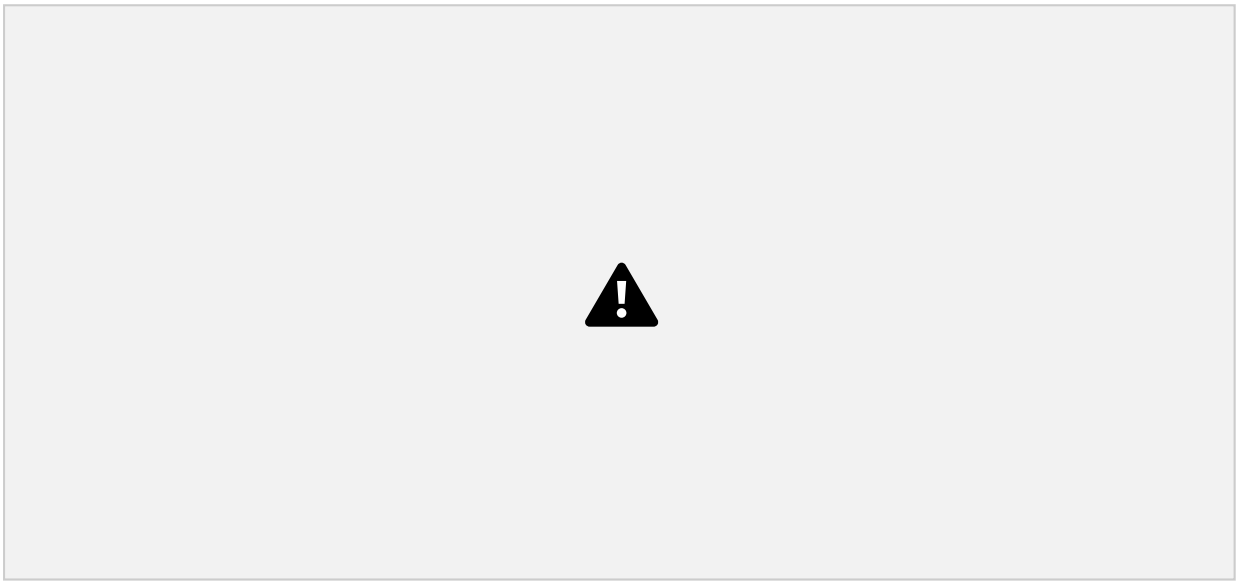
From **Switch Display**, a user can update to one of the following:

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

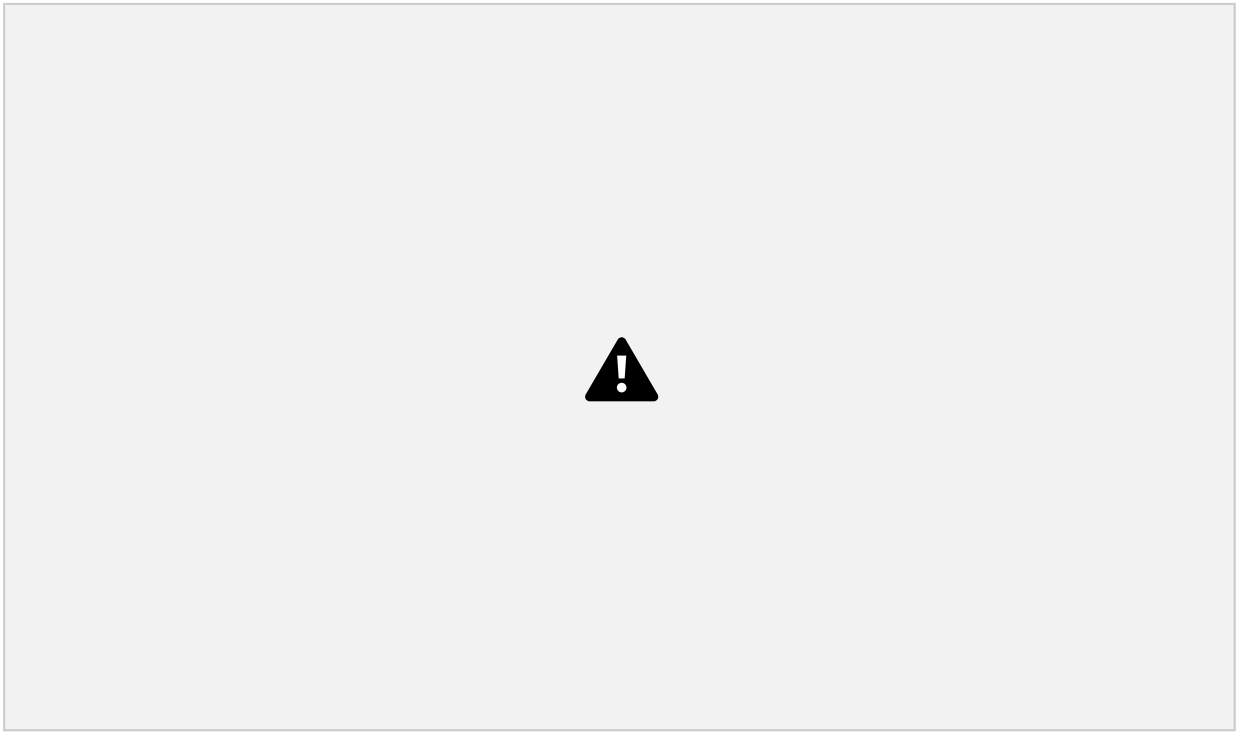
Under **Switch Isoform**, the available RefSeq and Ensembl isoform builds are listed. A condensed display and the protein length is shown for each isoform. The current selection appears in red text. The default KRAS isoform for example, is NM_004985 with 189 amino acids. To change the isoform, click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



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

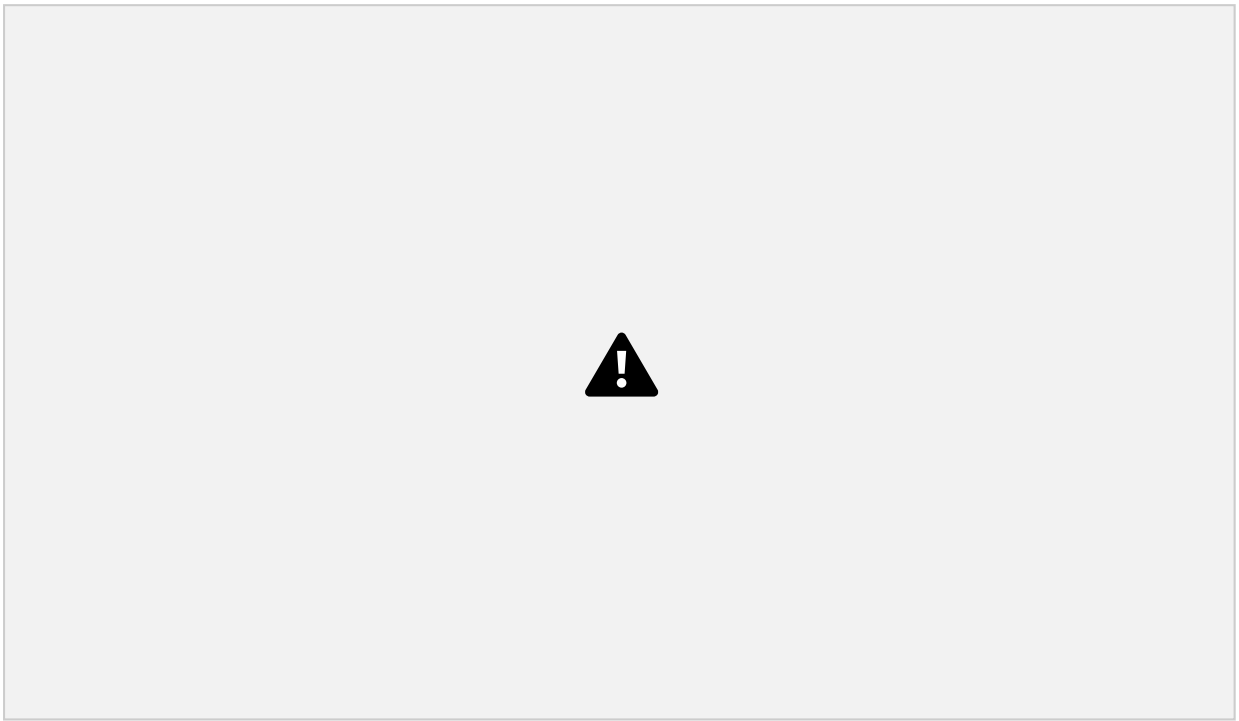
The lollipop chart for the GDC variants appears above the Protein View. The circular disc for each variant is proportional to the

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).



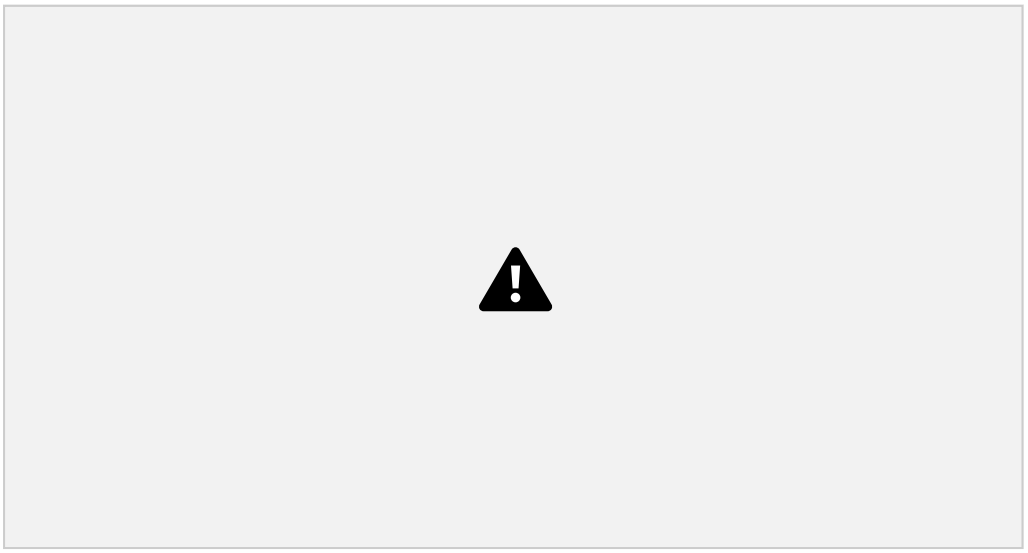
Exon variants report the amino acid change at the referenced codon. For example, G12D is a G > D substitution at the 12th codon of the protein.

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).

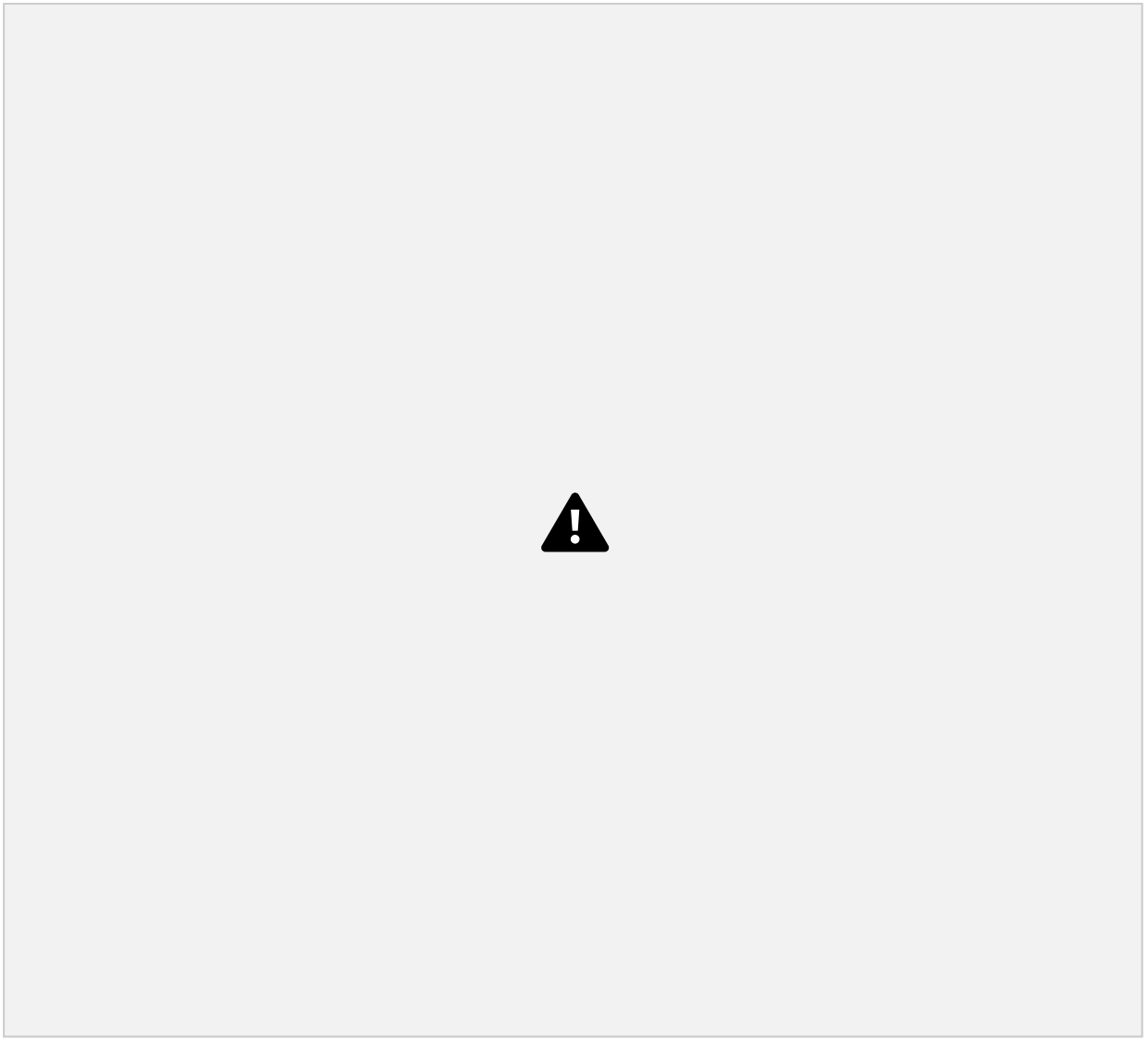


Viewing Variants and Case Samples
Variant Annotations and Chart Manipulation

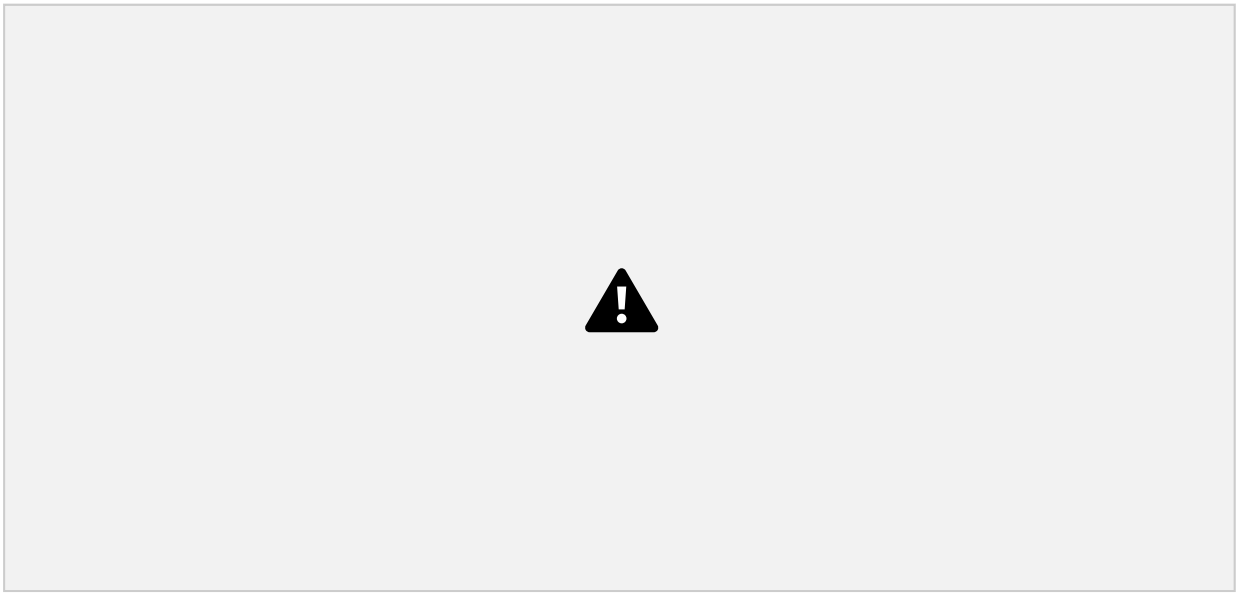
Click on the number of variants linked to the left of the lollipop for viewing annotations and manipulating the lollipop. For variant annotation, click on ‘List’.



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



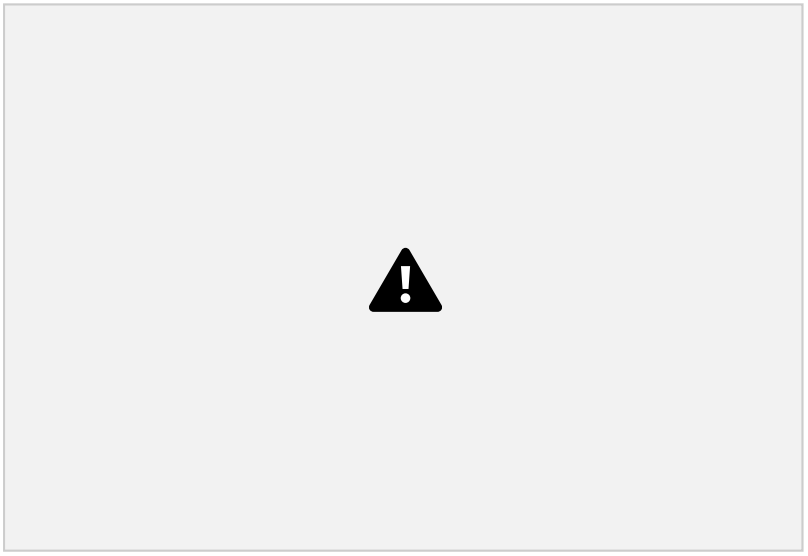
Click on the variant of interest and a new annotation table appears. From the table, view various associated features per sample such as: Disease type, Primary site, Project id, Gender, Race, Ethnicity, and Tumor DNA Mutant Allele Frequency(MAF). In the figure below, 333 occurrences are shown for the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



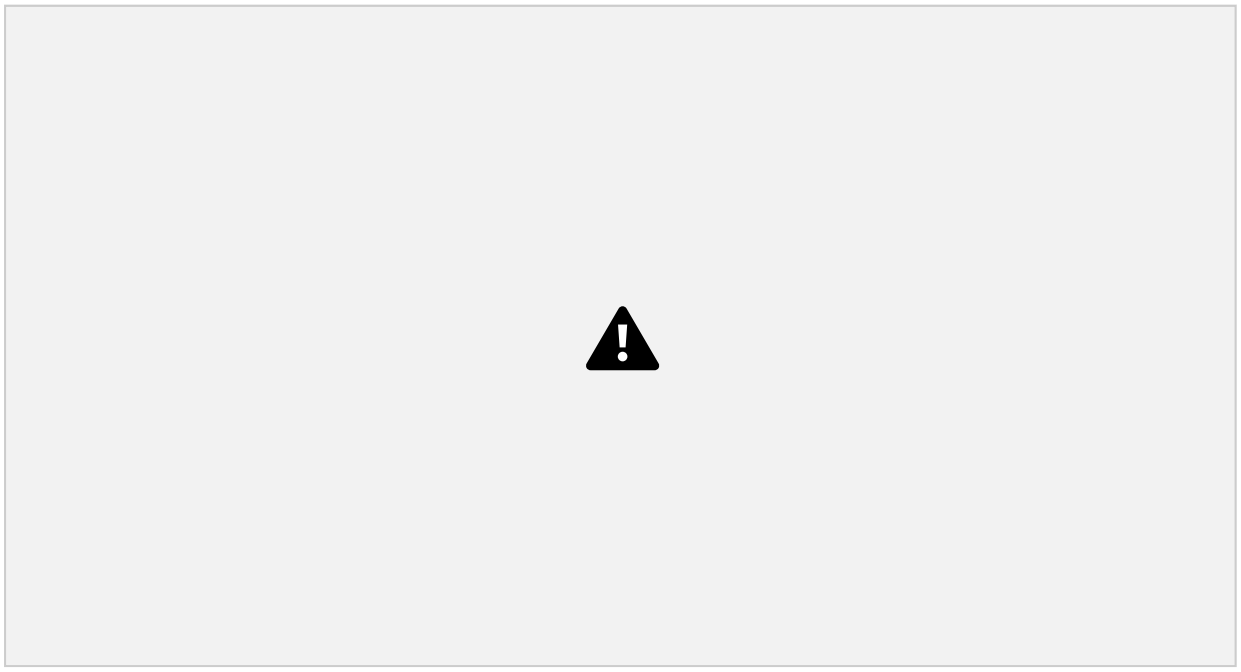
The first sample that is highlighted in yellow is a male with ductal and lobular neoplasms with a tumor DNA MAF of 31/125. This indicates 31 mutant alleles were found out of 125 total alleles.

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.

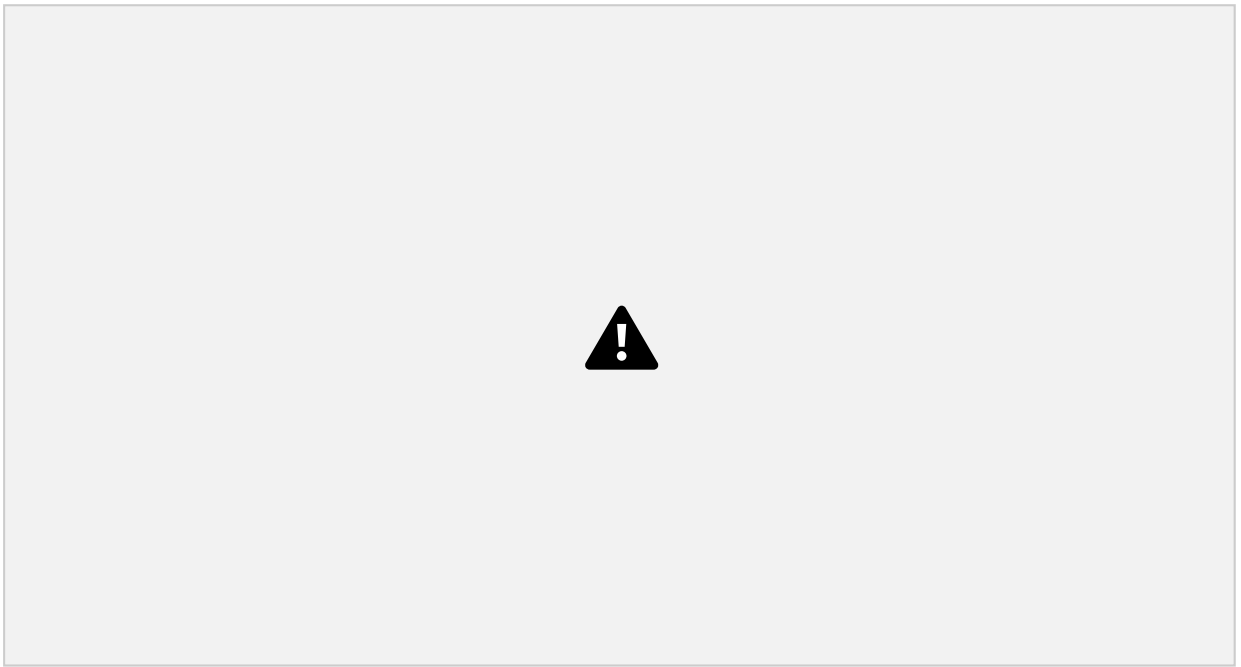


After clicking on the variant menu again, select the ‘Collapse’ option to collapse all skewers in the lollipop.



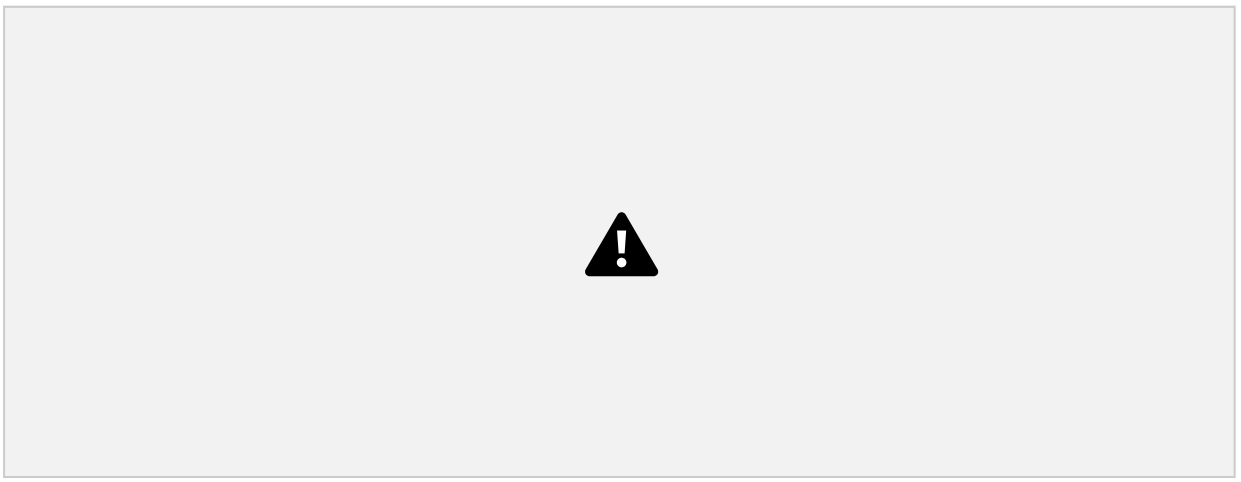
To expand any previously collapsed skewers, open the variant menu, and click on ‘Expand’.

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’.

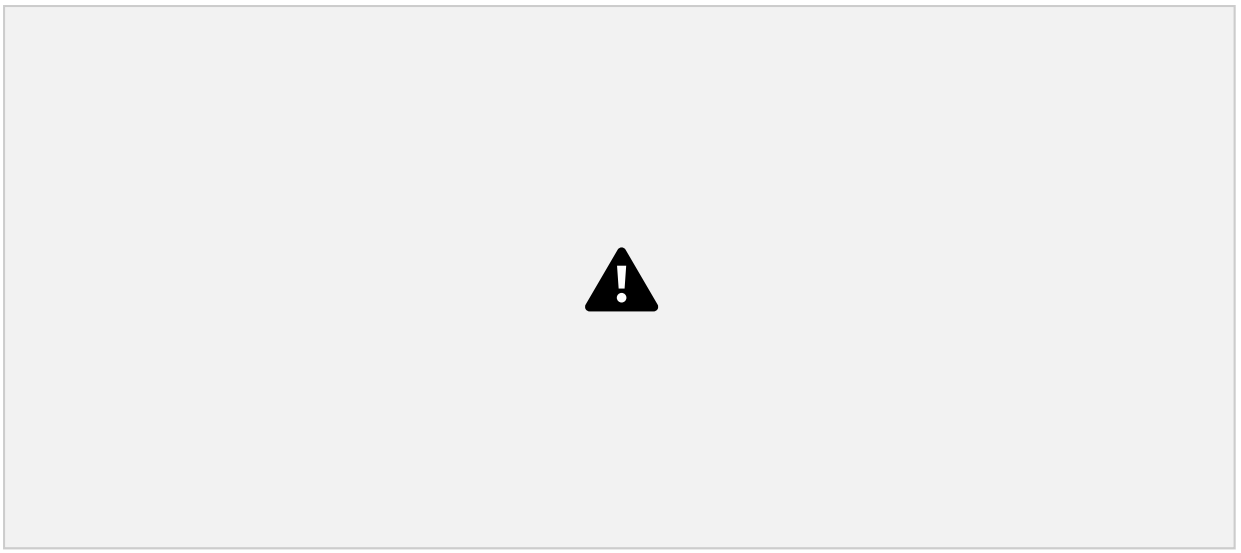


The lollipop re-renders with the variants sorted on the y-axis from lowest and highest occurrence. Hover over a variant to display the number of occurrences. In the example below, a user is hovering over G12D to display 333 occurrences of this

variant.

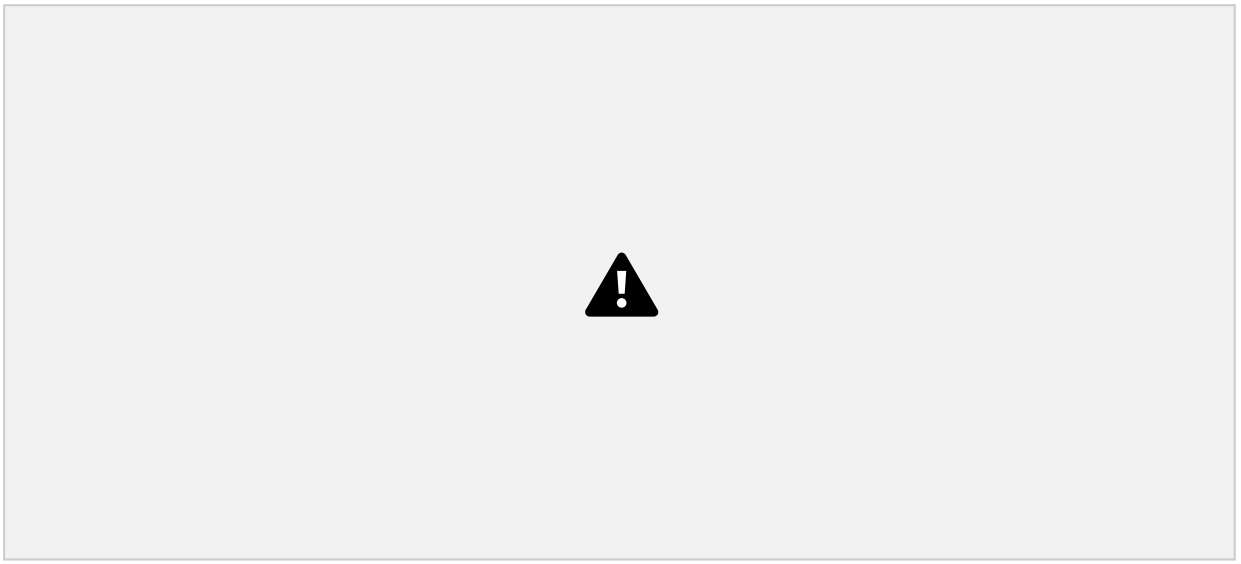


Clicking on the variant loads the sample table again as shown below.

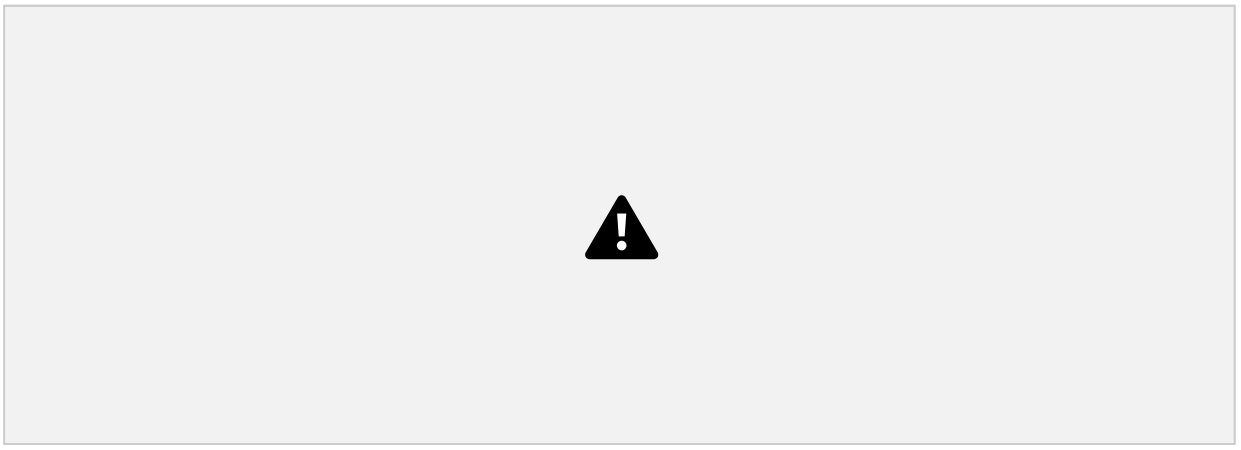


Case Filtering

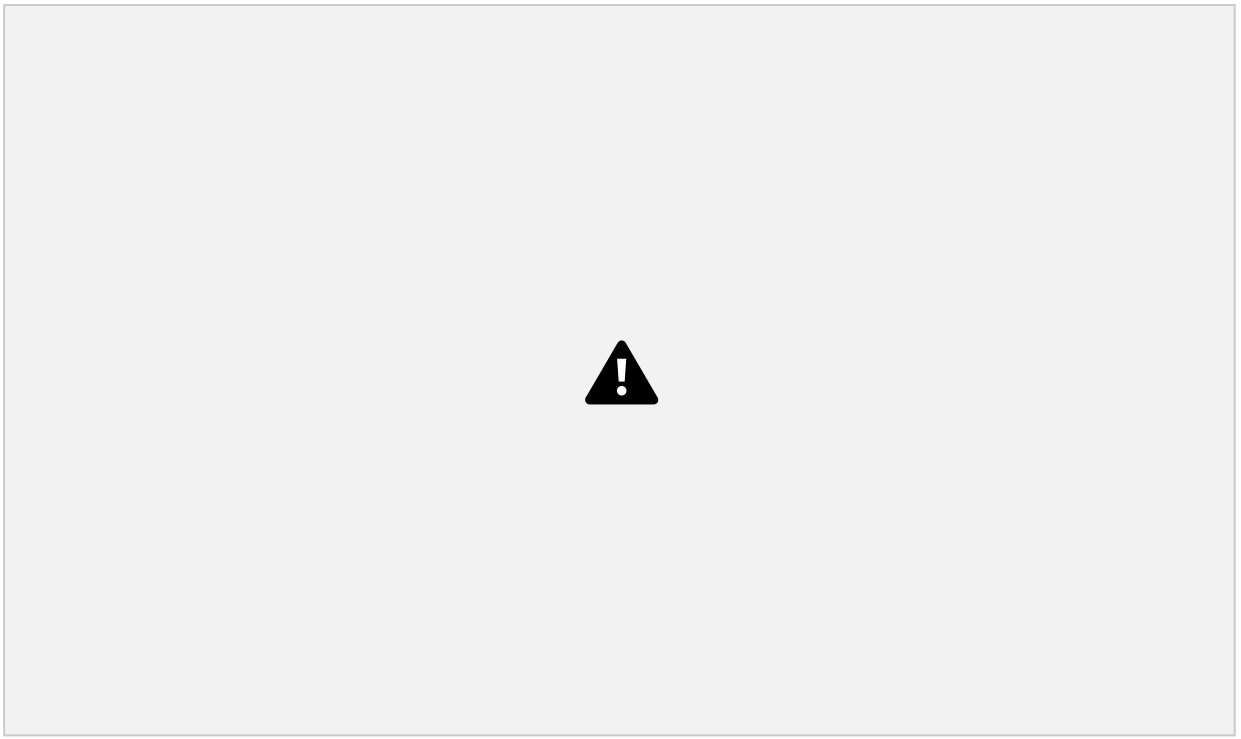
Clicking on the sample hyperlink on the left of the lollipop (e.g. 1315 samples) opens a menu to list all samples. Aggregate data for all samples by attribute appears in a series of tabs. The ability for advanced filtering and creating subtracks is available from this new display.



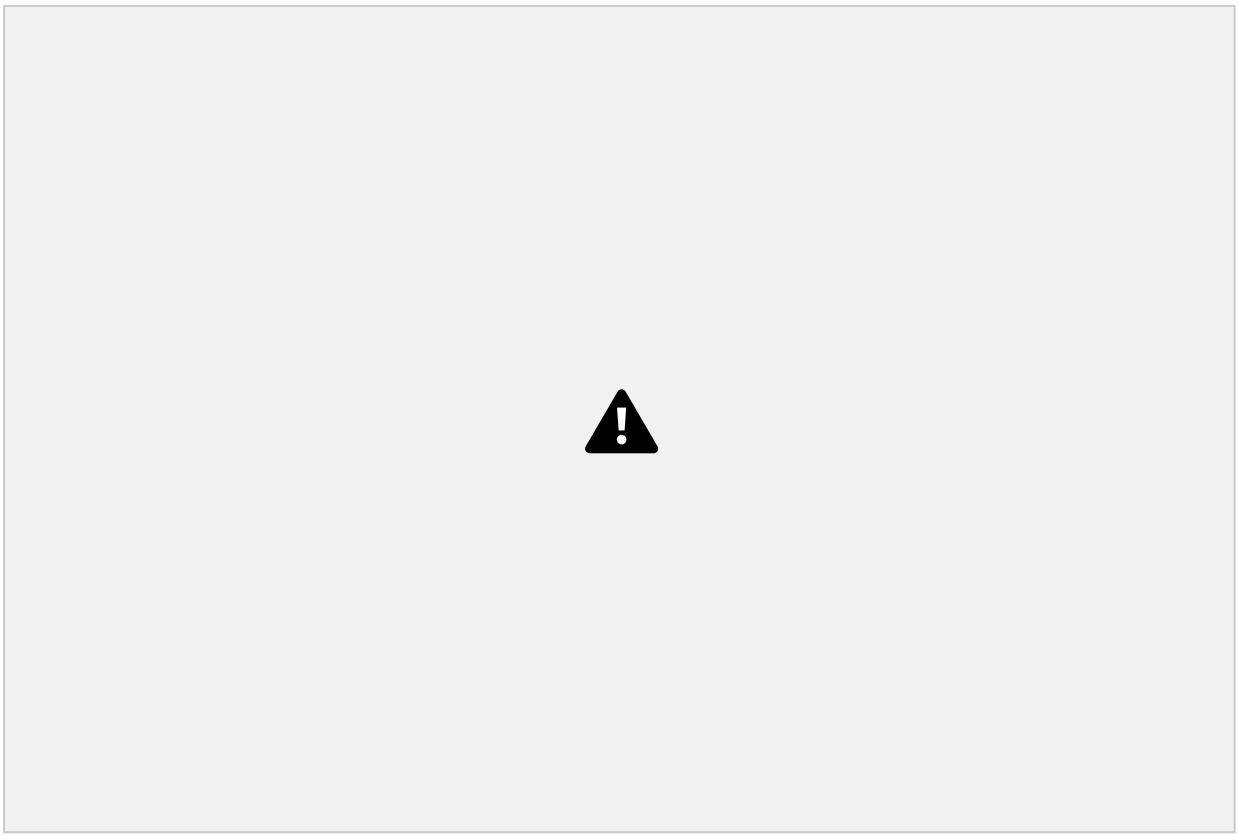
Click on 1315 samples to view annotations grouped by attributes such as: Disease type, Primary site, Project id, Gender, Race, Ethnicity, etc.. For each attribute, the number of values is represented by ‘n’ to the right of the group label. In the figure below, 21 values for Disease type are reported.



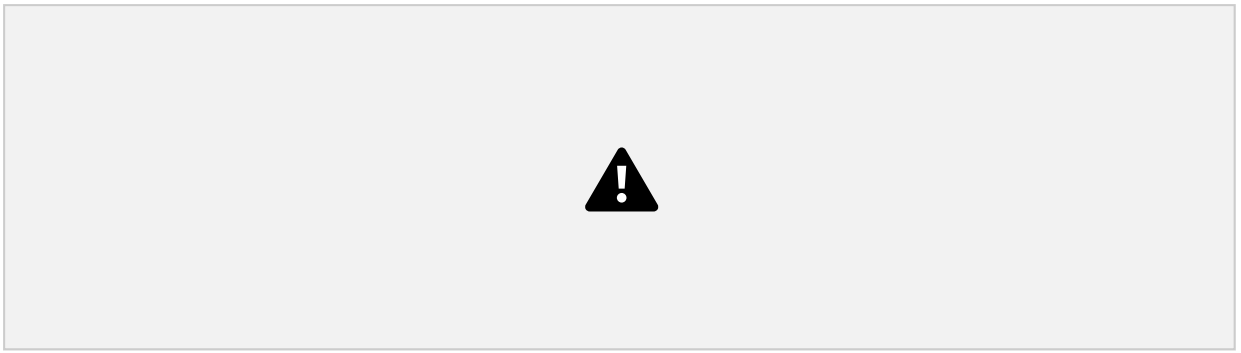
To start filtering, click on the value label or the value's sample fraction. Clicking on 'Adenomas and Adenocarcinomas' or '675/4866' for example, loads a new lollipop subtrack underneath the main GDC lollipop track.



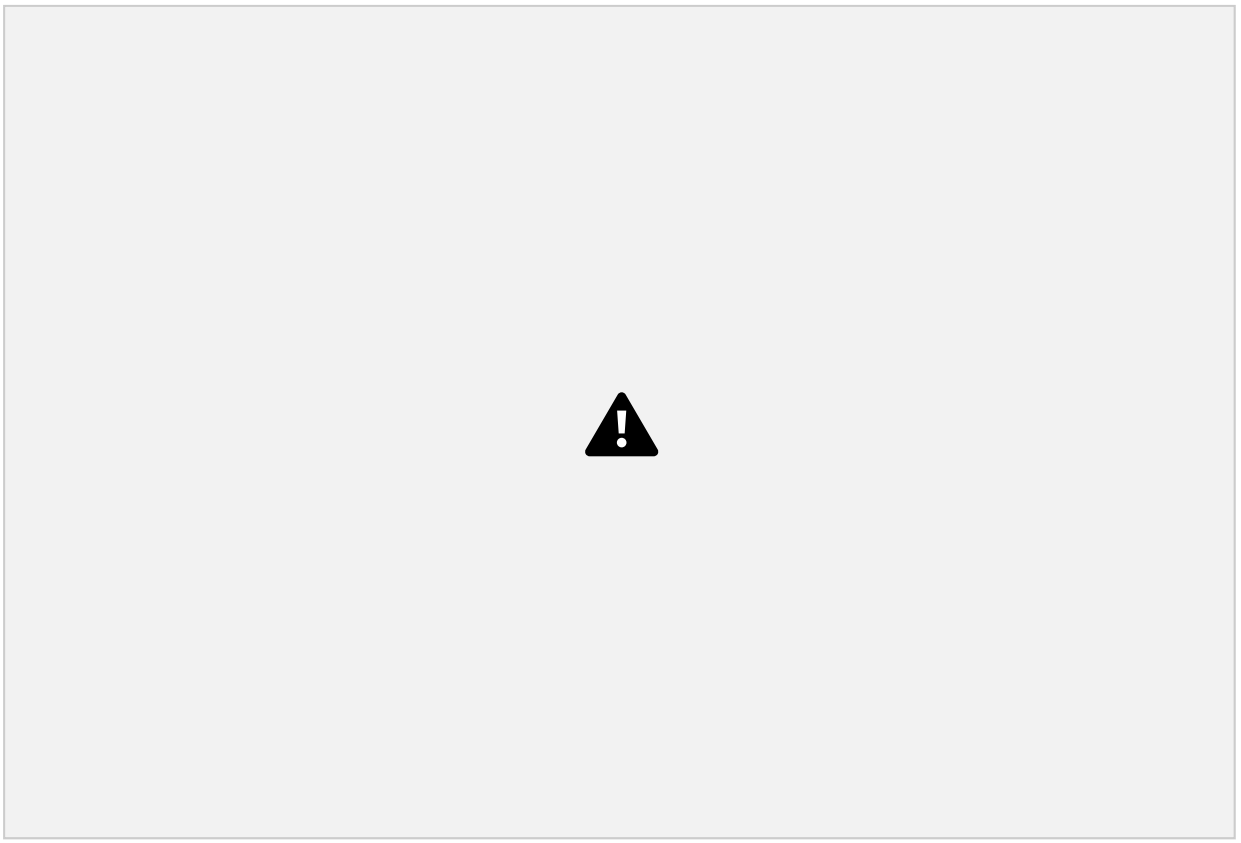
This new subtrack only shows the 675 Adenomas and Adenocarcinomas samples. This side-by-side view allows for a comparison between the mutations in the main track vs the subtrack.



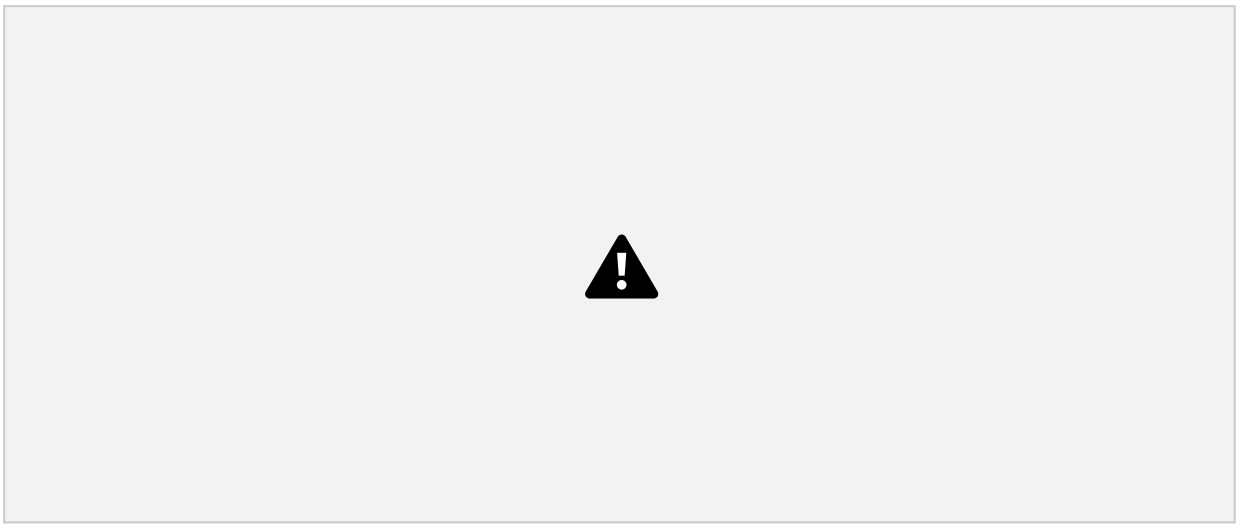
Each subtrack offers advanced filtering, shown below, for users to narrow down particular features.



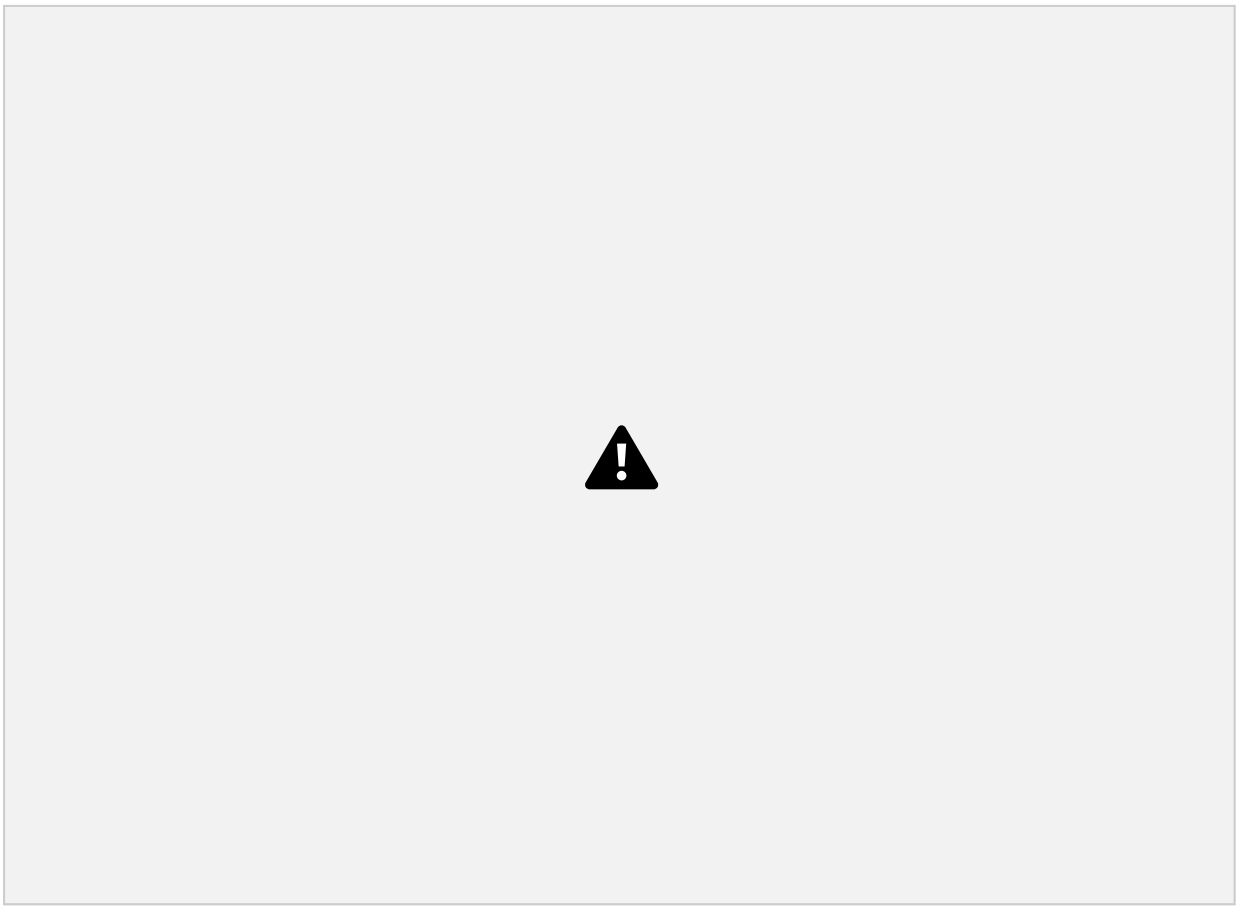
Clicking on ‘Filter’ displays a pop-up window with the feature the user selected previously from the sample annotation menu (e.g. Disease type: Adenomas and Adenocarcinomas). Clicking on either +AND or +OR displays a new pop-up with a search bar. Search for the desired term and click on the term’s button.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



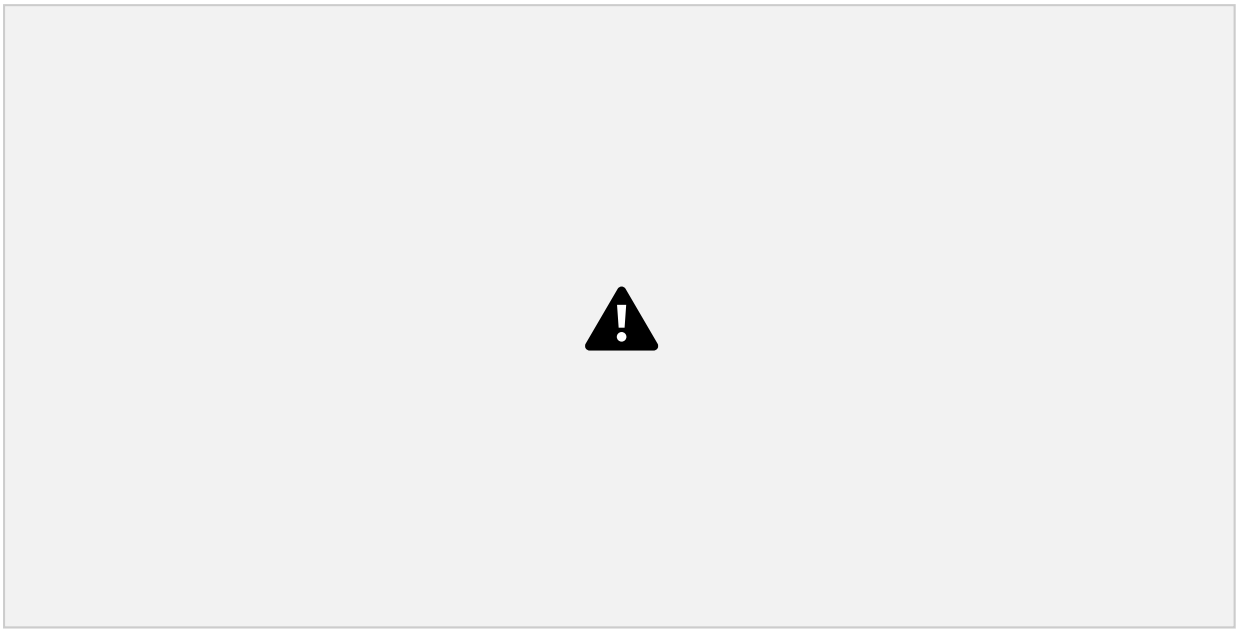
By clicking on ‘Gender’, all available values appear with checkboxes (i.e. male and female) as shown below. In this example, male with 293 data points is selected.



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.

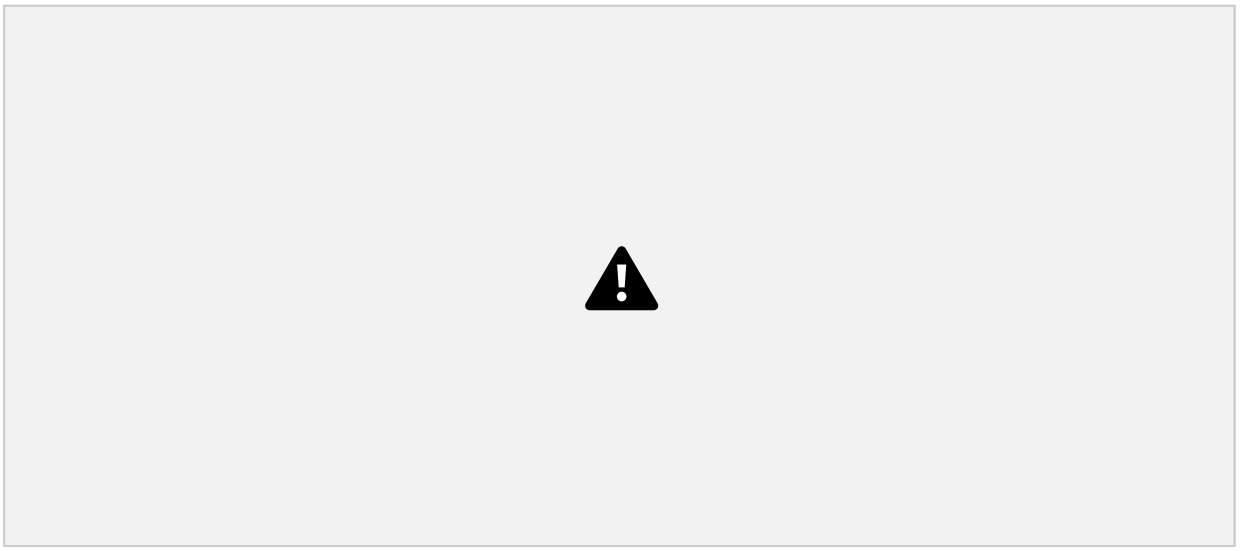


Click on the ‘Close’ option to remove the subtrack from the page.



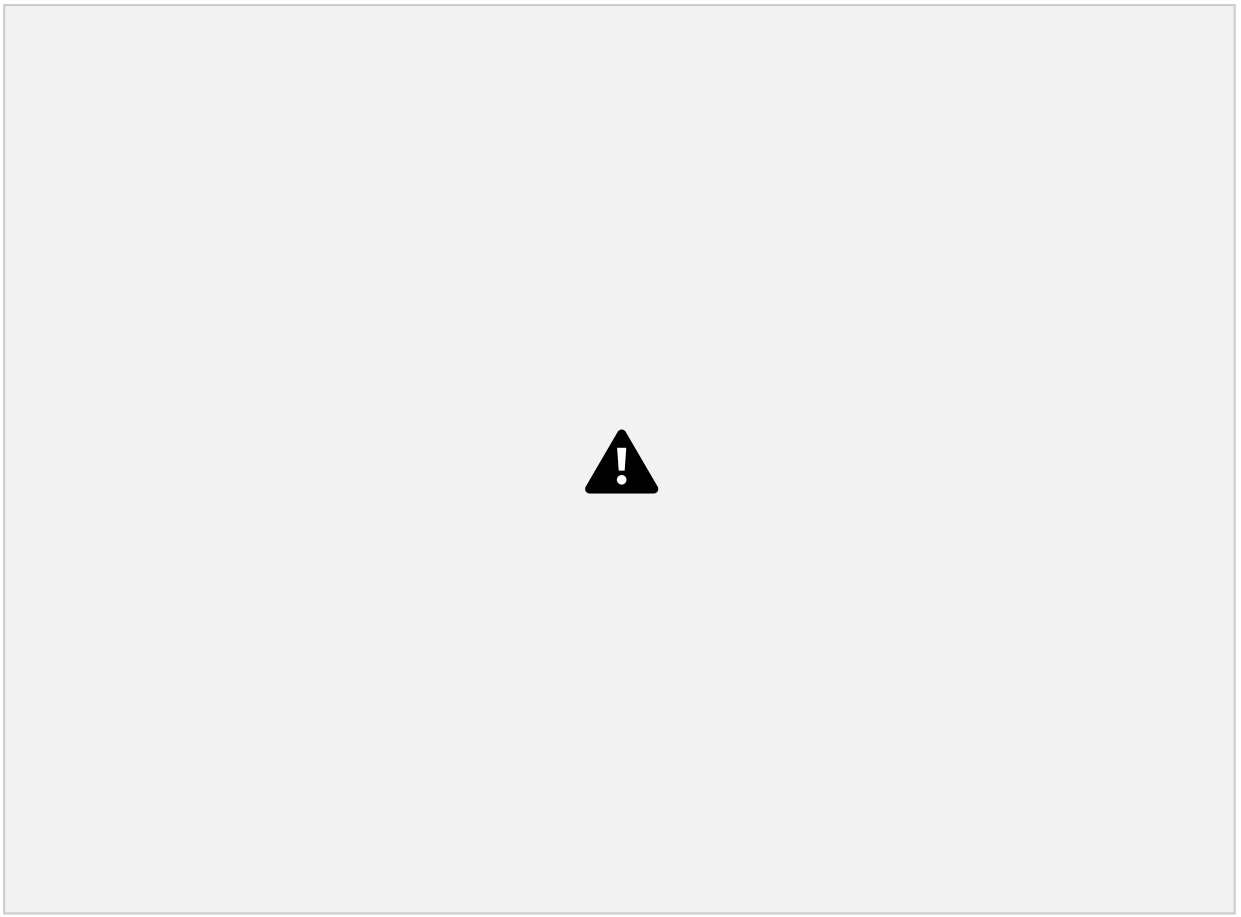
Viewing in the Lollipop Display

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.



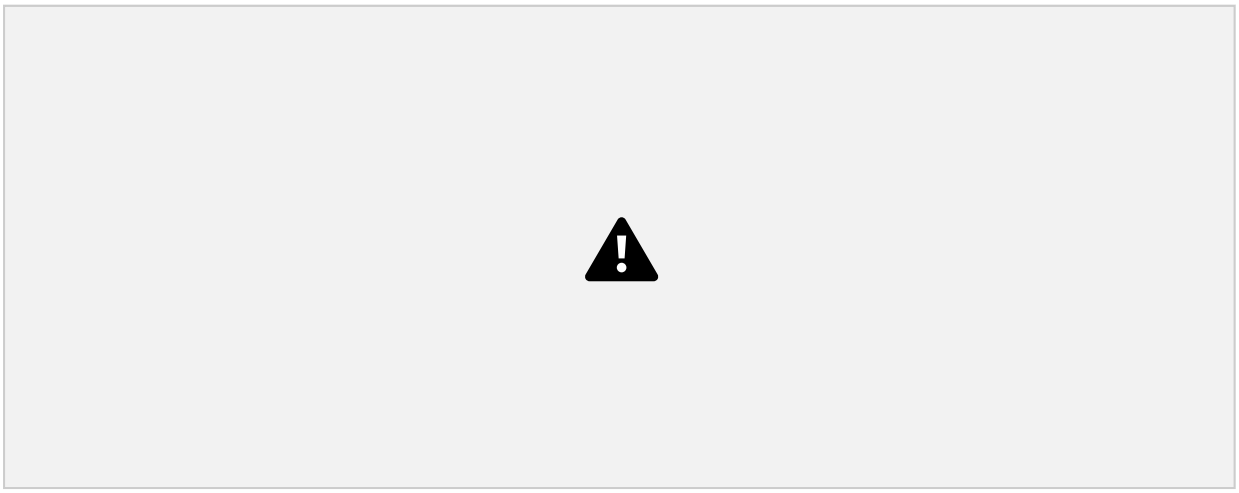
Detailed variant annotation is viewable by clicking on the variant disc next to the label. For G12D highlighted in a red outline in the image above, click on the ‘333’ disc. A sunburst chart will appear, shown below.

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

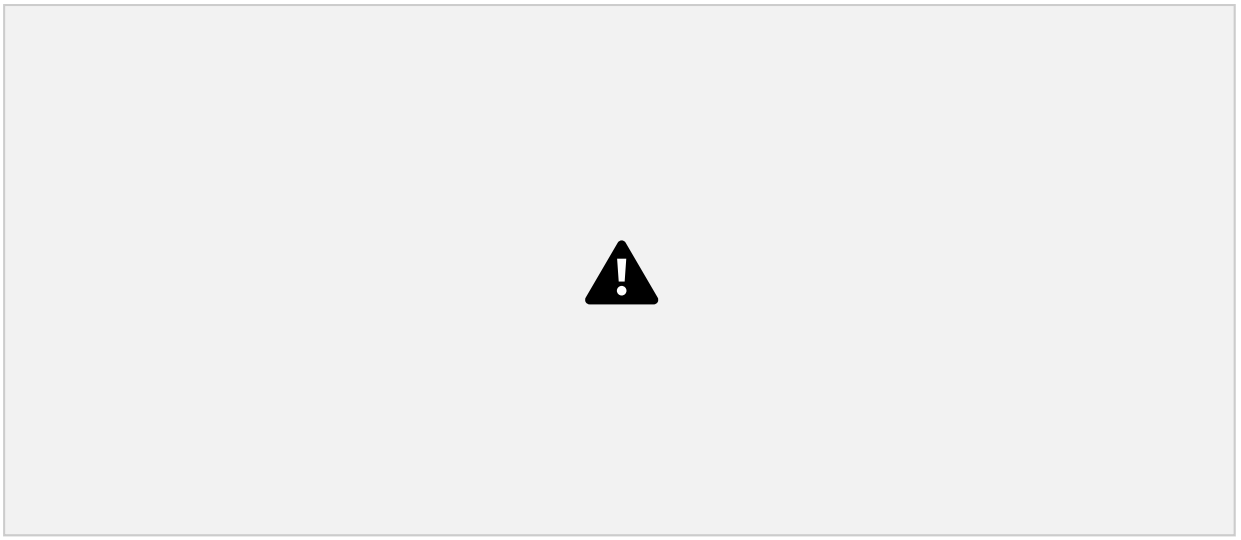


hierarchy is arranged by disease types then broken down by primary sites. Hovering over the inner ring displays the disease type, number of samples, and cohort size. In this example, the inner green ring displays ‘Plasma Cell Tumors’ with 28 samples out of a total 949 samples.

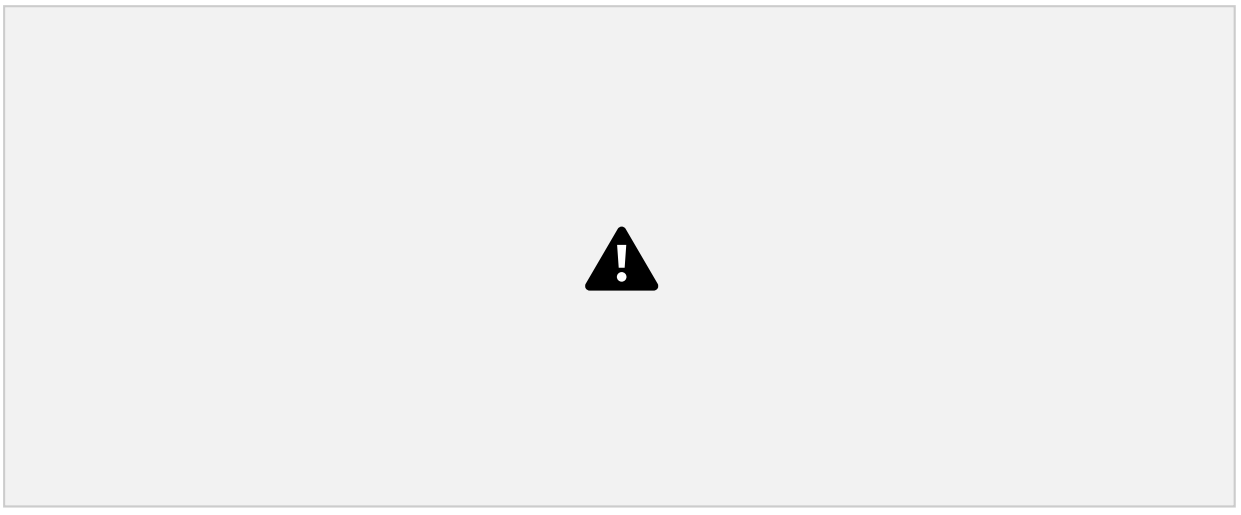
The outer ring represents the primary sites. Hovering over the primary site displays the number of samples relative to the disease type. In the figure below, for Ductal and lobular



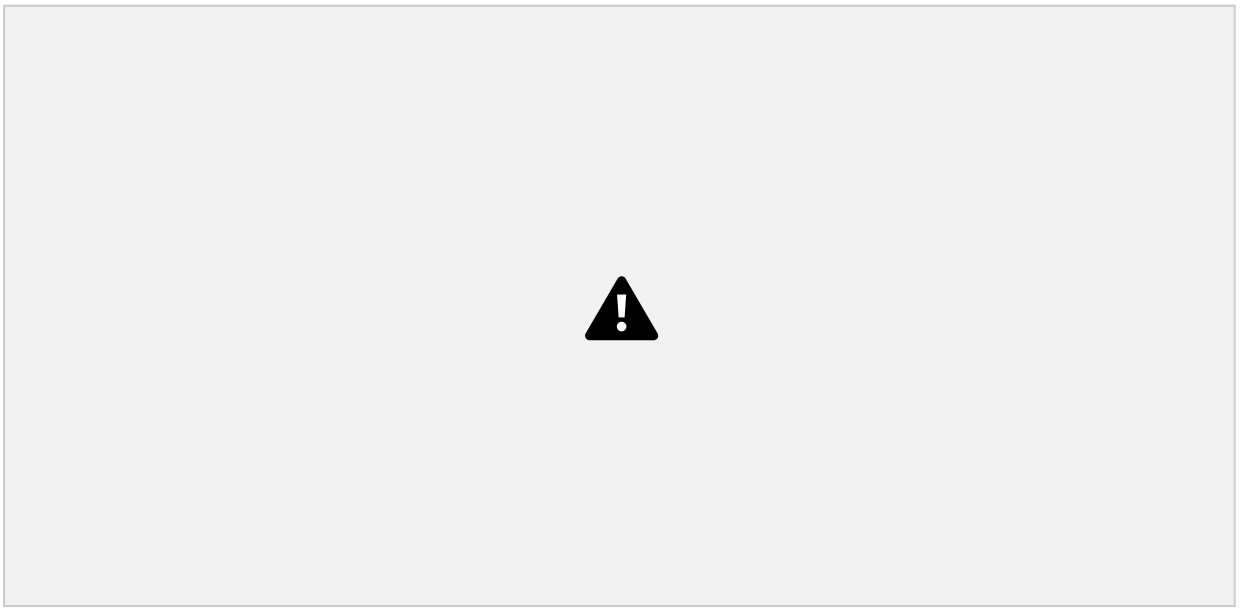
neoplasms, there are 105 samples with the primary site as pancreas out of 316 total samples.



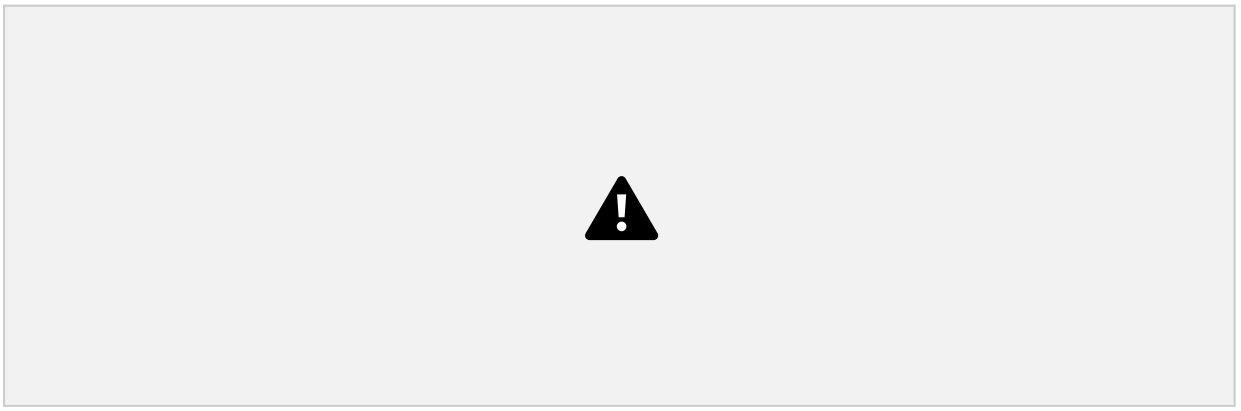
Clicking on a node displays a sample table for the disease type or primary site. In the figure below, the user selected ‘Plasma Cell Tumors’. The sample annotation table appears for all Plasma Cell Tumors.



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.

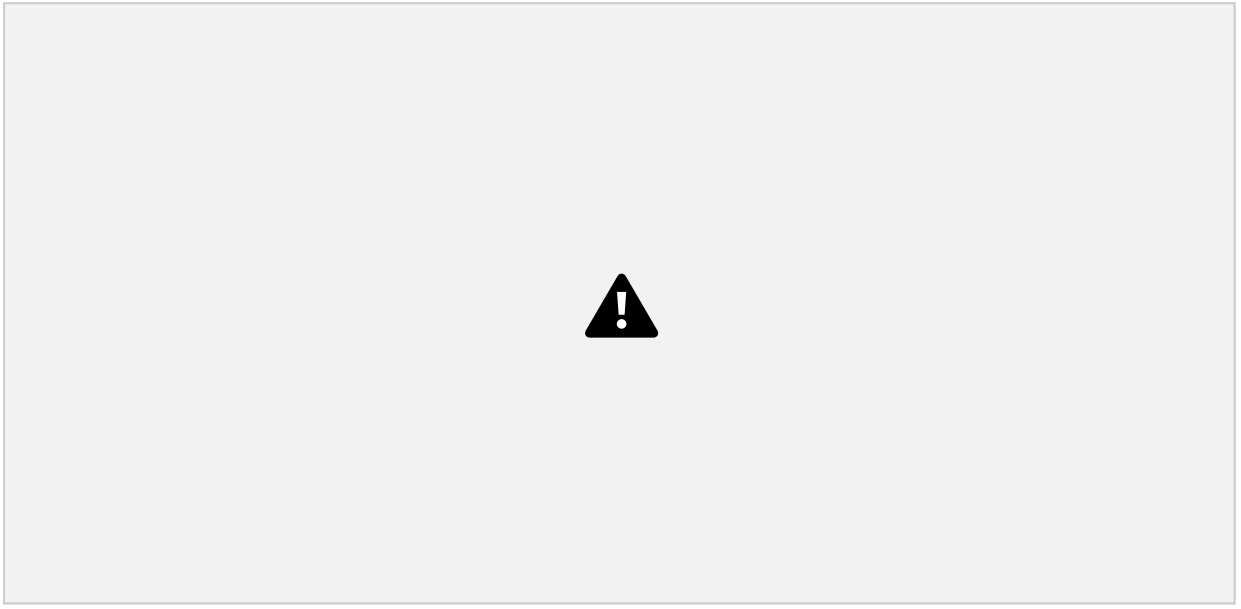


Clicking on the sample name hyperlink opens a new tab to the sample's GDC Case Summary page. Clicking on the variant label in the center removes the sunburst chart.



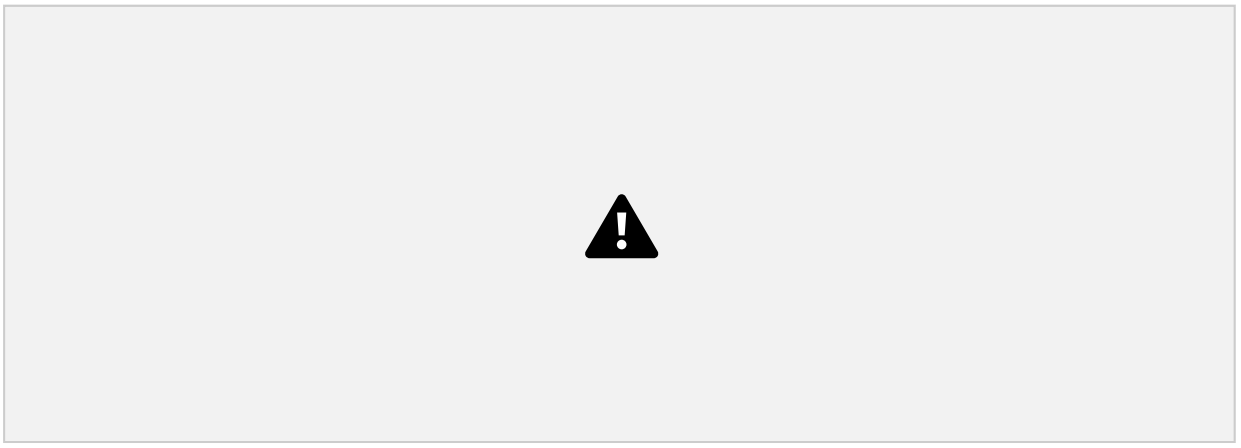
Working With the Protein Track

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.

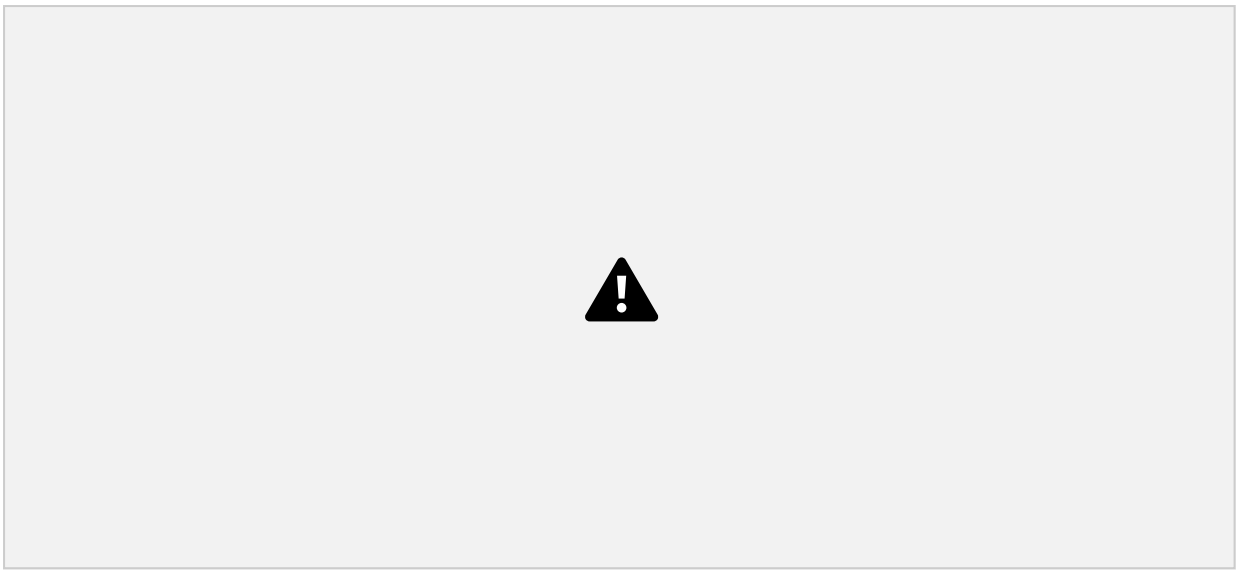


Once the mouse is released, the lollipop re-renders as the selected region.

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.

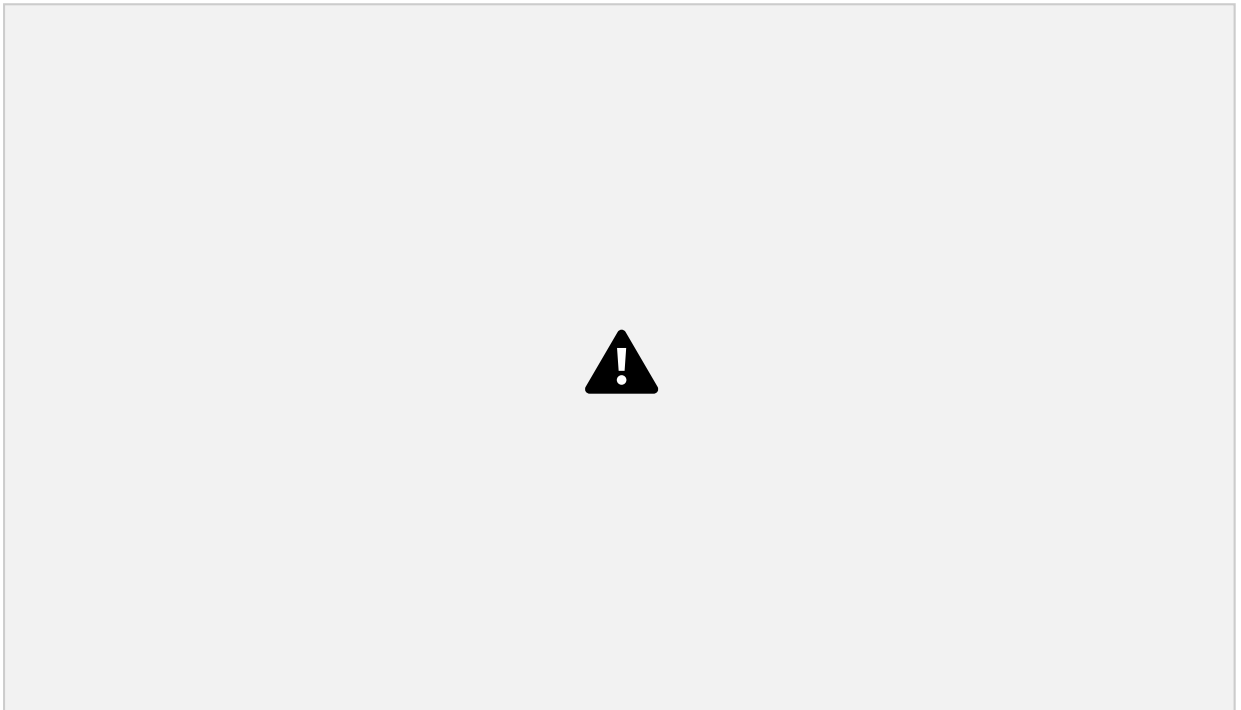


Zooming in causes the protein track to display the codons and the nucleotides as shown below. Hovering over the nucleotide position displays a tooltip with the exon, amino acids position, RNA position, and protein domain. As shown in the image below, at codon 12, the second exon of the transcript, RNA 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.



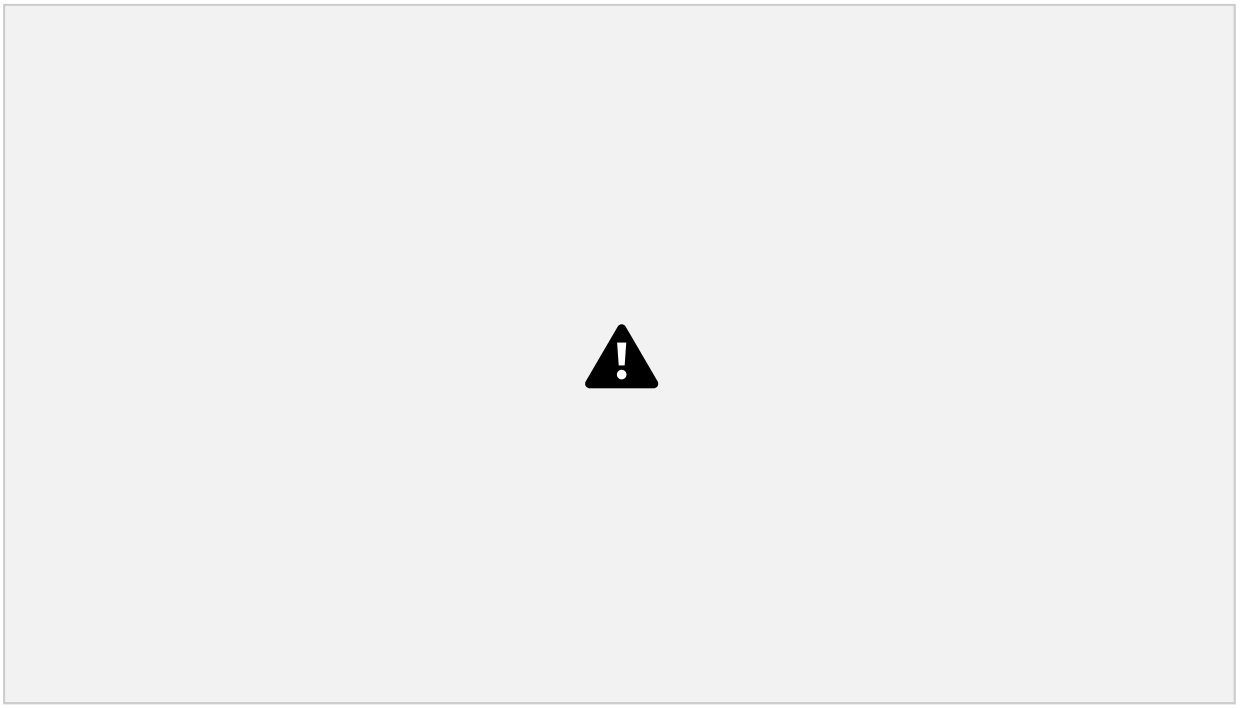
Legend Panel

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.

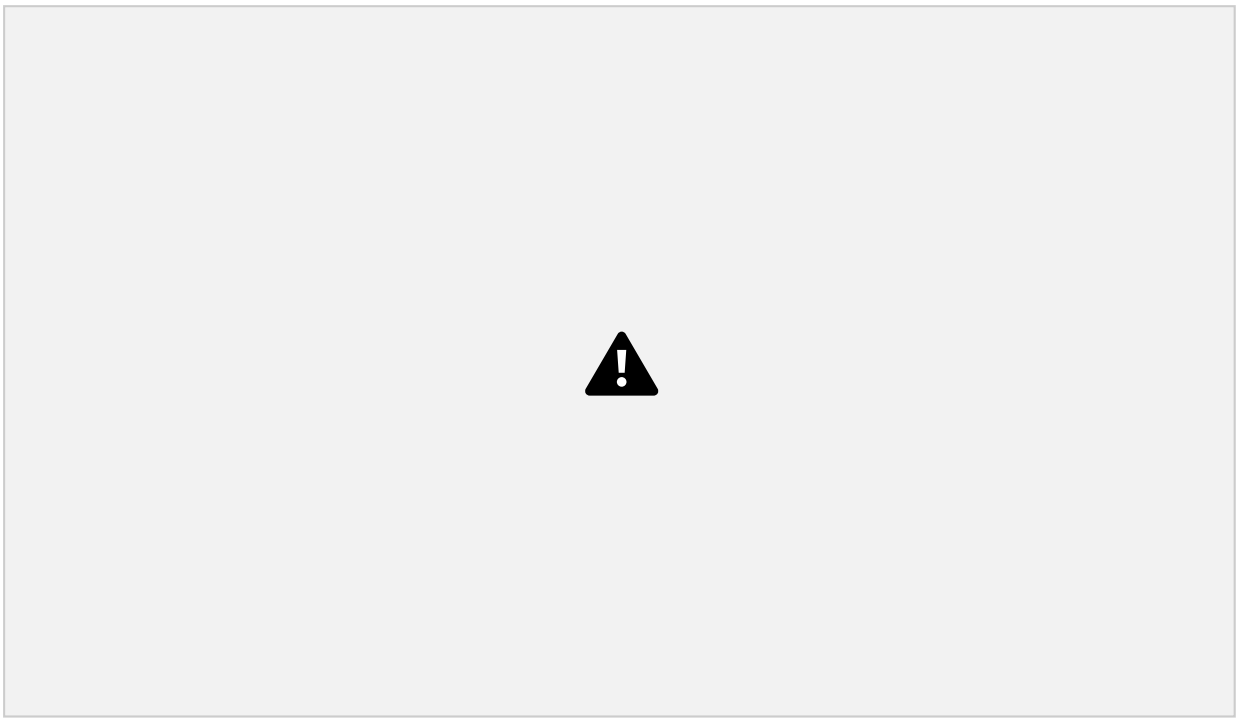


Custom protein domains are added by clicking on the '+add protein domain' button at the bottom of the list. An input box appears requiring the following information:

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.

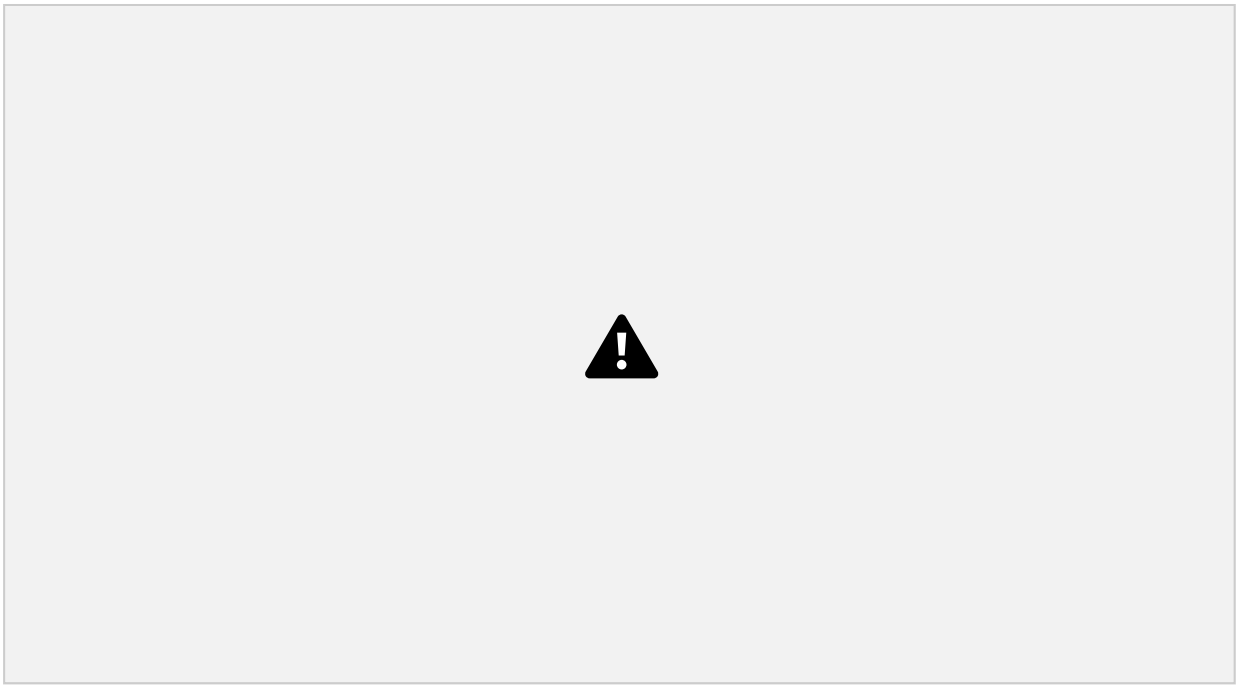
GDC Mutations

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.

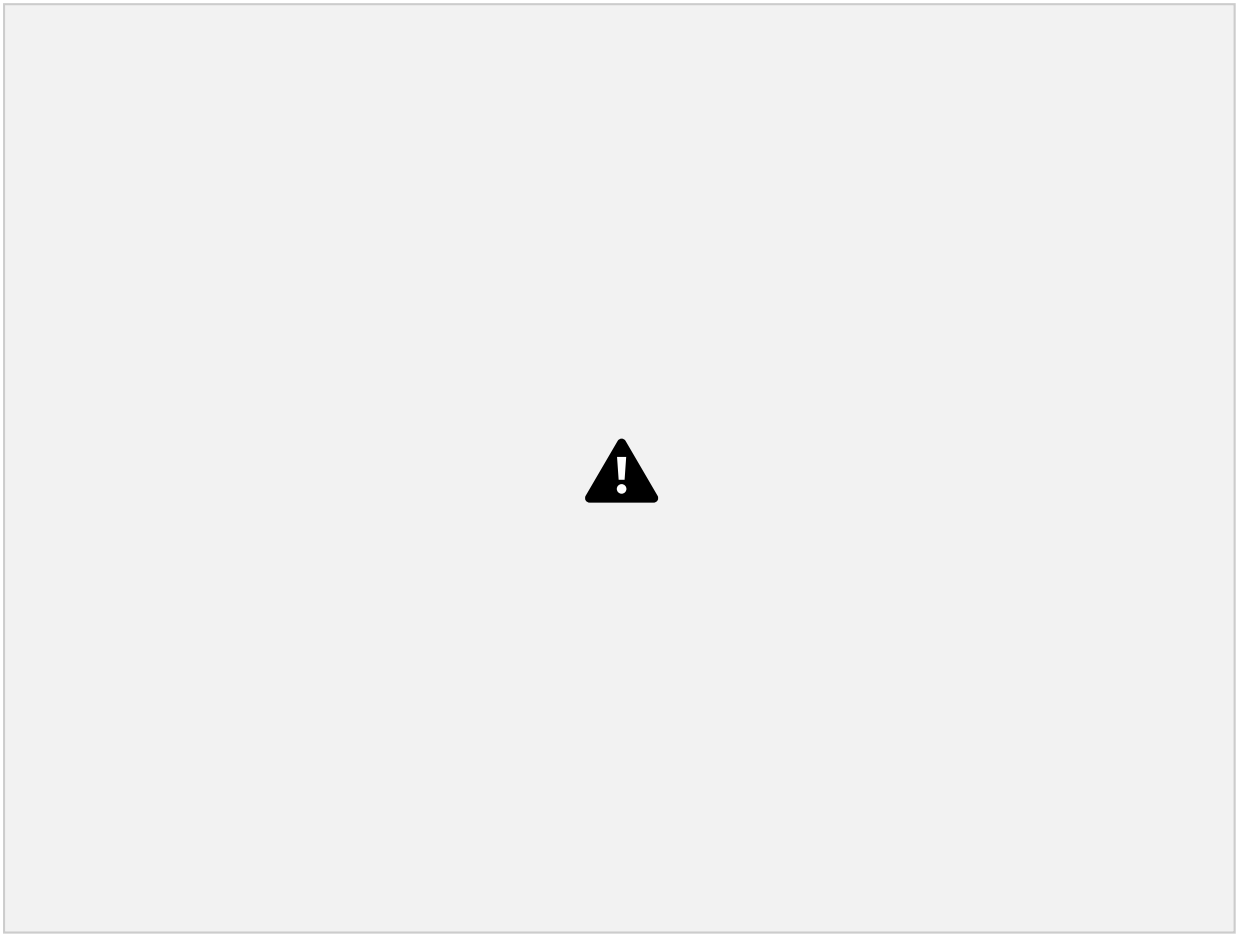


The classification for the type of variant is color coded as follows:

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.

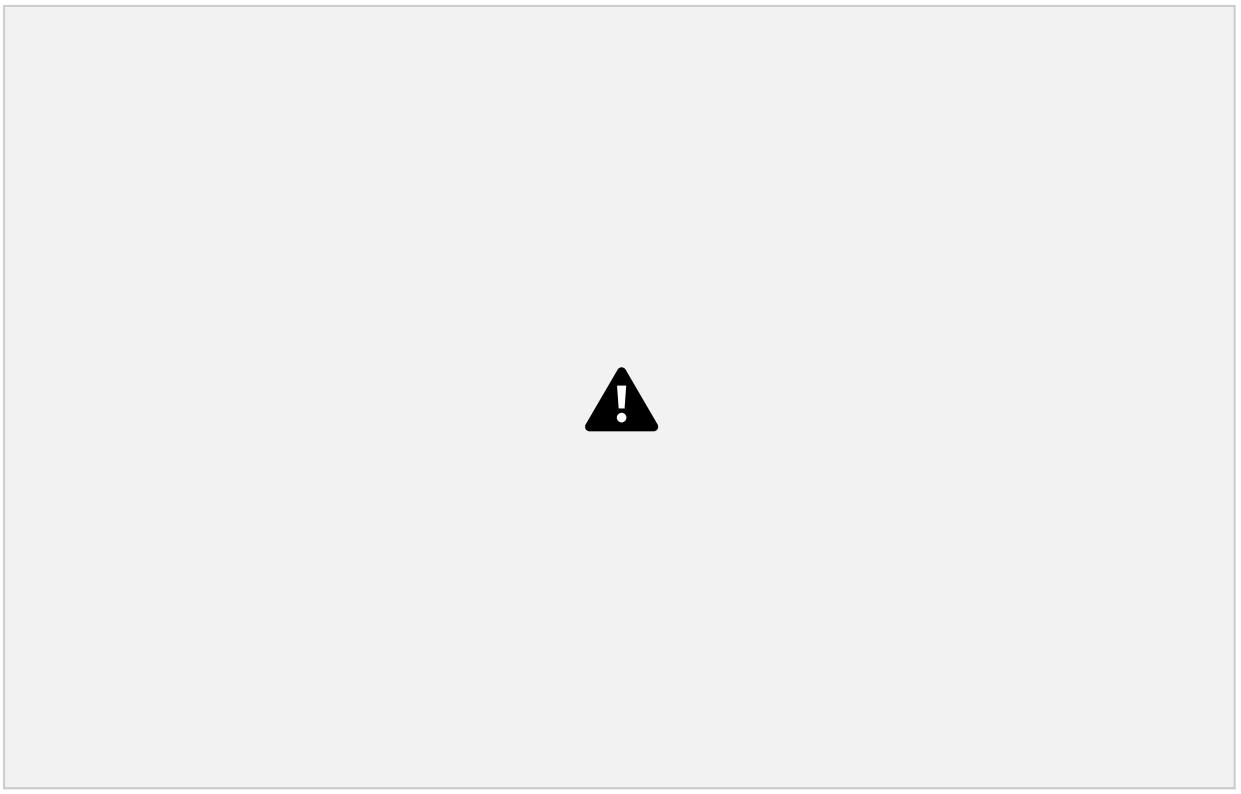


Clicking ‘Hide’ removes all of the mutation discs from the lollipop. The mutation is reordered to the end of the list and the font is striked through and grayed out. The discs reappear when the mutation label is clicked again.



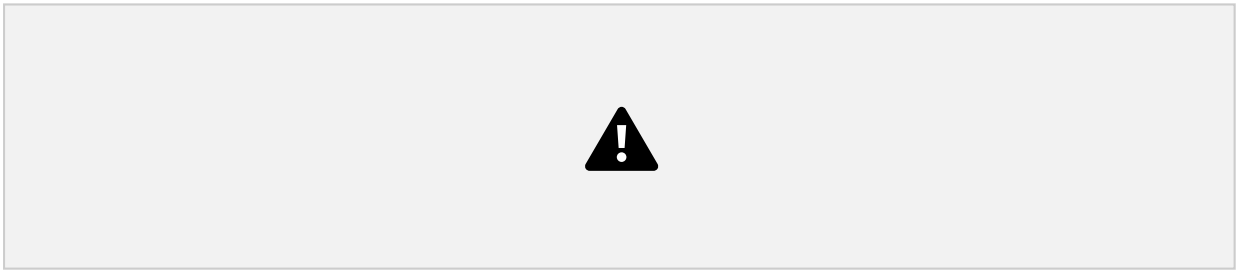
More Options

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.

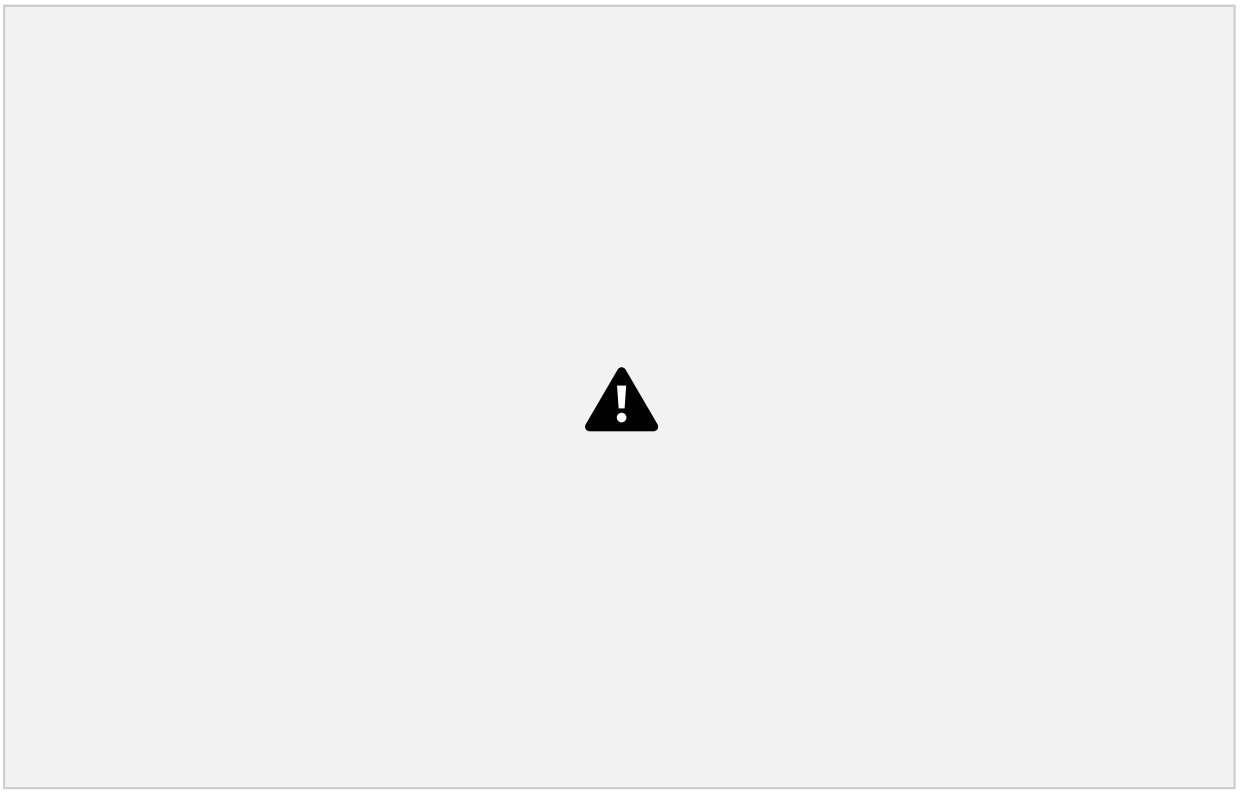


The exported figure will contain following contents, reflecting a user’s customization:

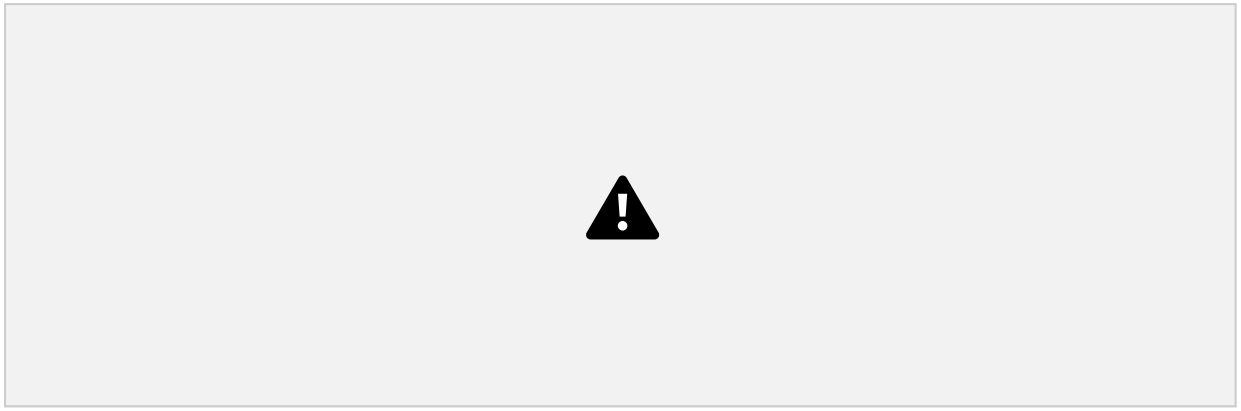
- Displayed datasets, including custom data
- Expand/fold states of all mutations
- Sequences in the protein if at zoom-in level
- Show/hide state of exon boundaries
- Sunburst charts
- Protein domains without the hidden ones
- All mutations without the hidden classes or origins
- Legend for protein domain, mutation class and origin

Copying the DNA Sequence

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

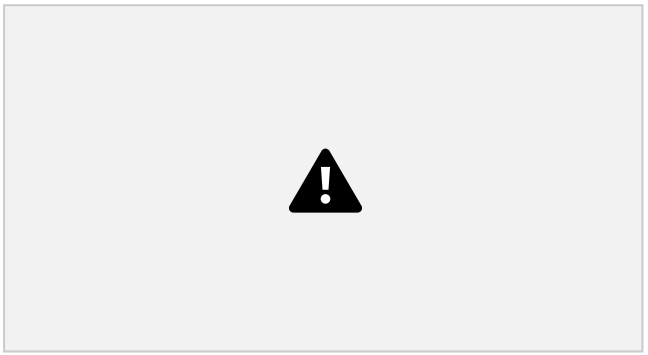


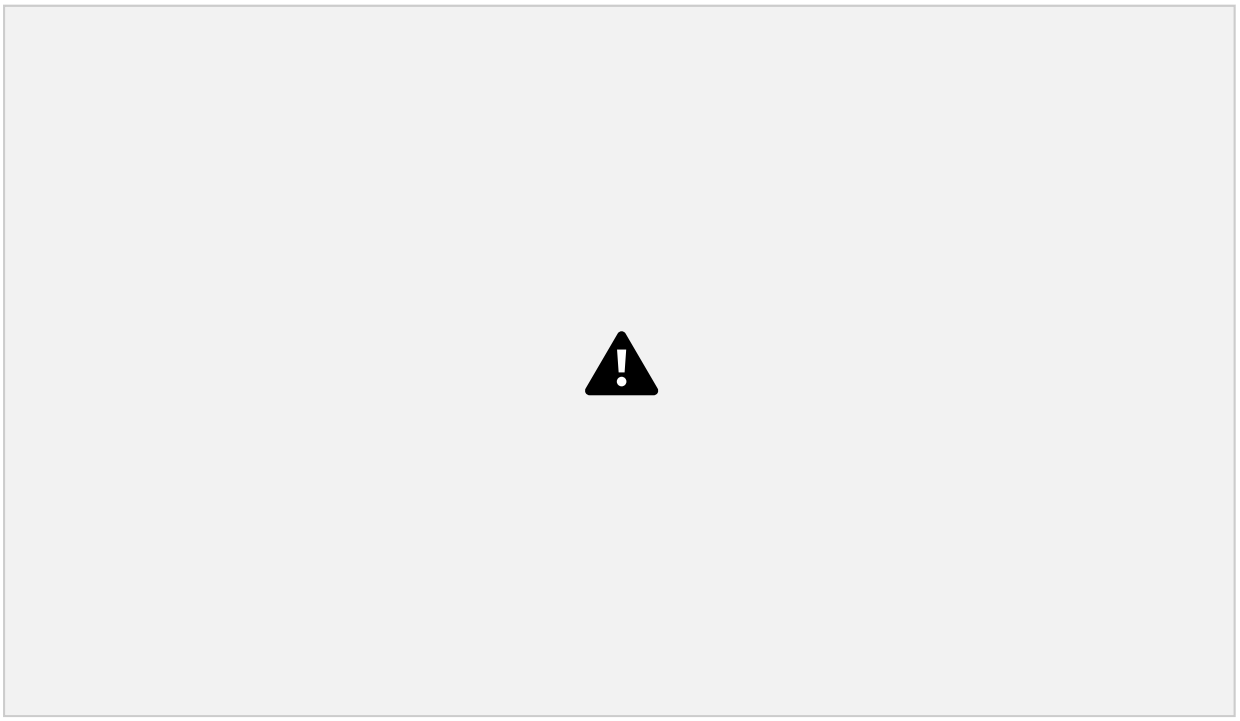
Clicking on ‘DNA sequence’ displays the DNA sequence as plain text for easy copying and pasting.



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.





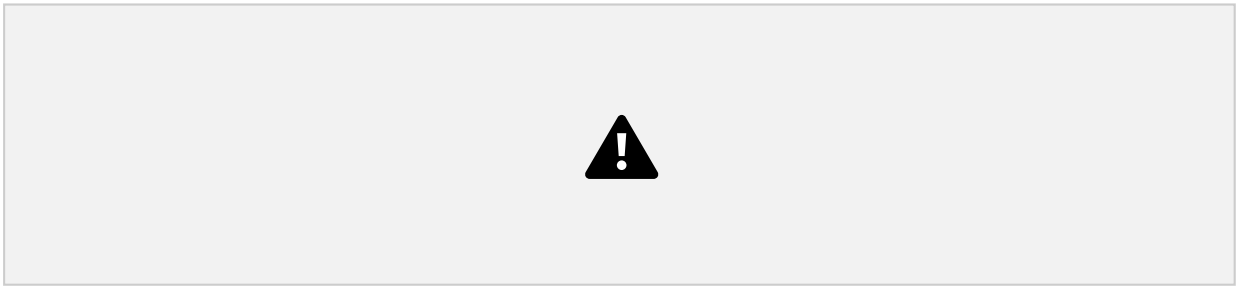
ProteinPaint Sequence Reads Tool

Sequence Reads Tool Introduction

The Sequence Reads Tool is a web-based tool that uses the ProteinPaint BAM track (ppBAM) and 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) it can classify reads supporting the reference and alternative allele into separate groups.

Launch the Sequence Reads Tool

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

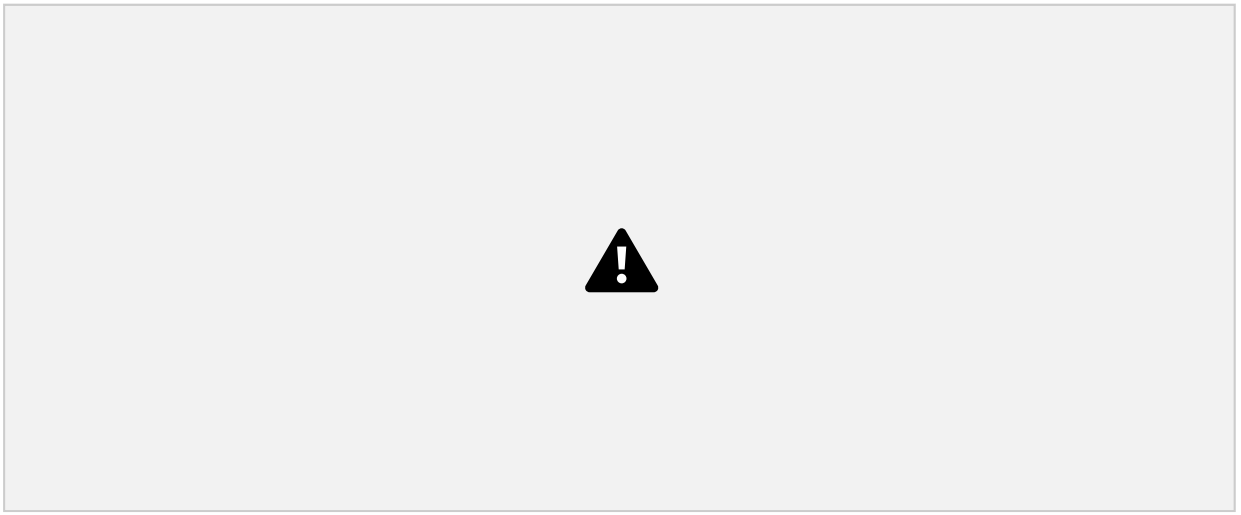


A user needs to be logged in to use this feature. If not, the user will be prompted to log in.
Once the user logs in, a search bar and submit button will appear as below.

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.

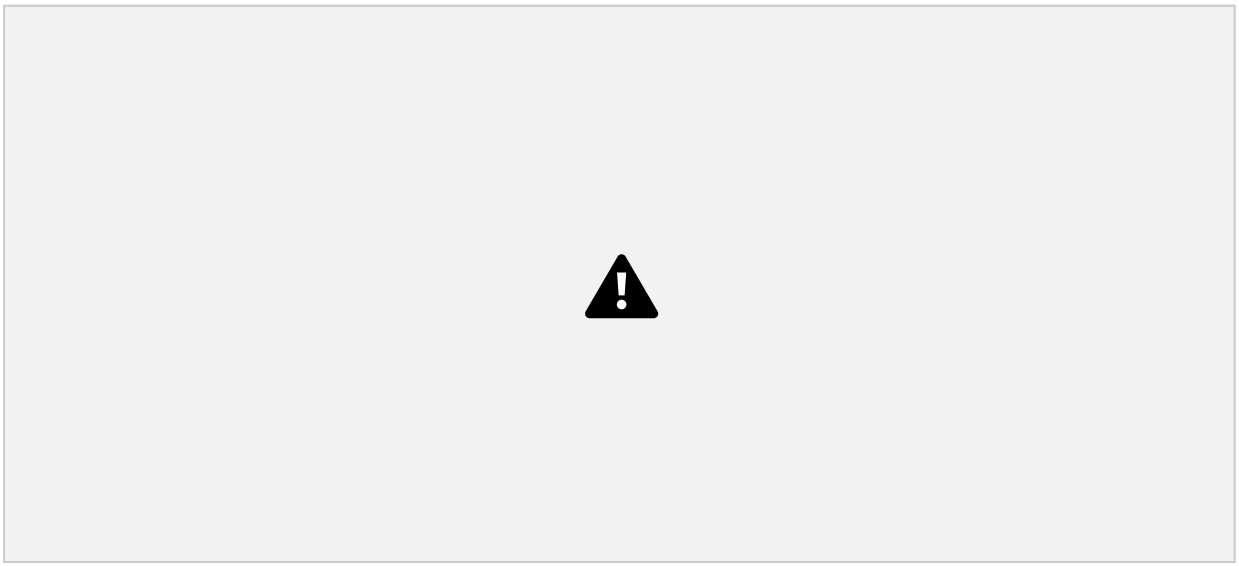
As an example, using case ID "TCGA-06-0211" will return 9 BAM files available from this case displayed in a table. One or multiple BAM files from this table must be selected to proceed.



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

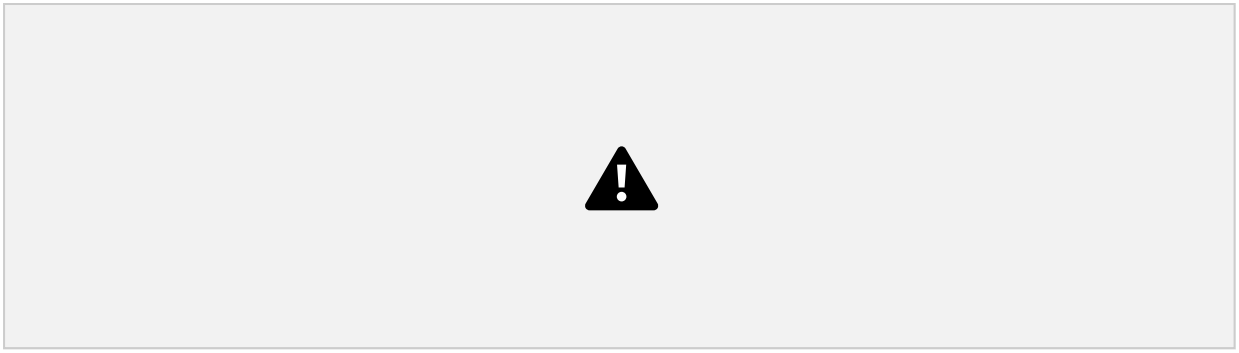
The Current Position in Genome text box displays the coordinates of the region currently displayed on the screen. It initially shows the coordinates specified in the URL. On pan/zoom by the user, this region displays the updated coordinates of the view region.

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 panel](#Read information panel)the 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

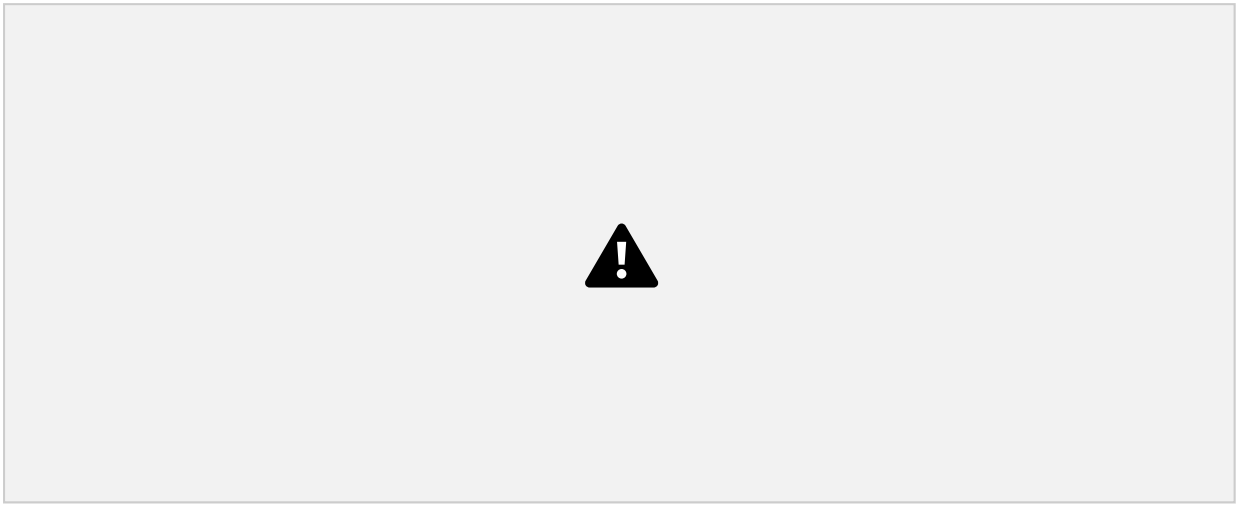
The rendering of the reads depends upon the zoom level (horizontal zoom) chosen by the user and the number of reads mapped at the display region (vertical zoom).

Horizontal Zoom

The BAM track has three levels of horizontal zoom:

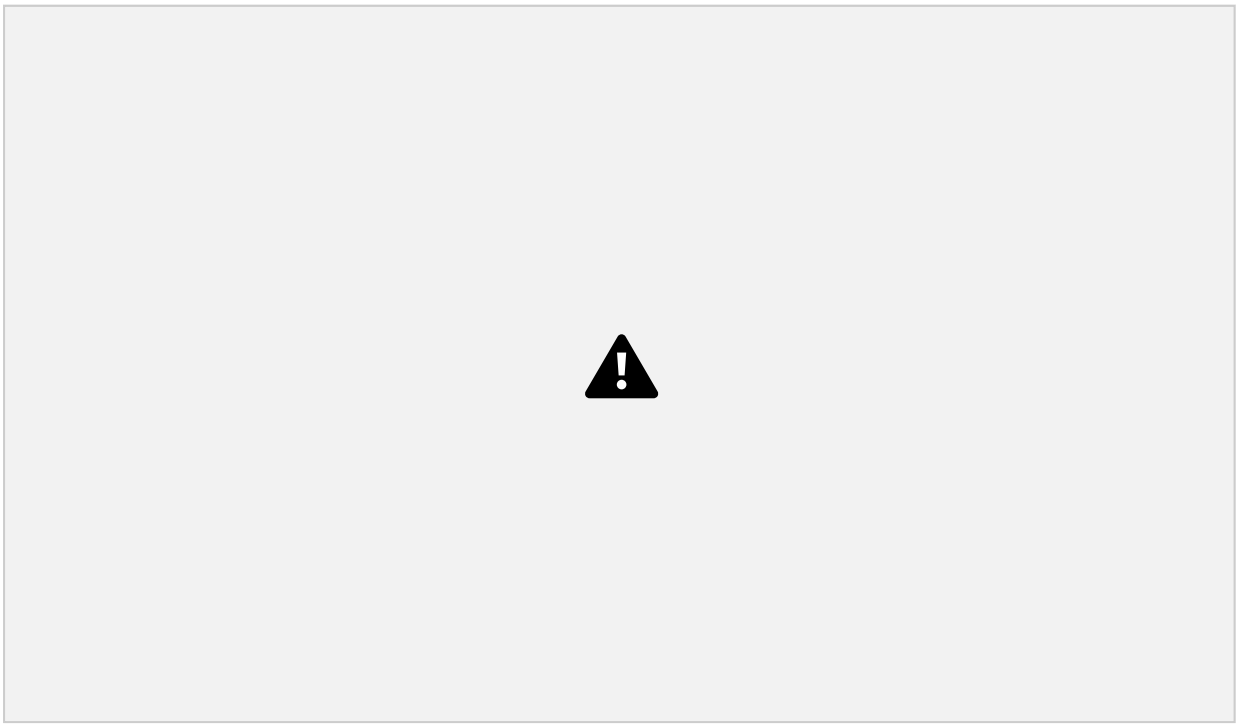
Overview Level

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](#Reference genome 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)](#Color coding of reads).



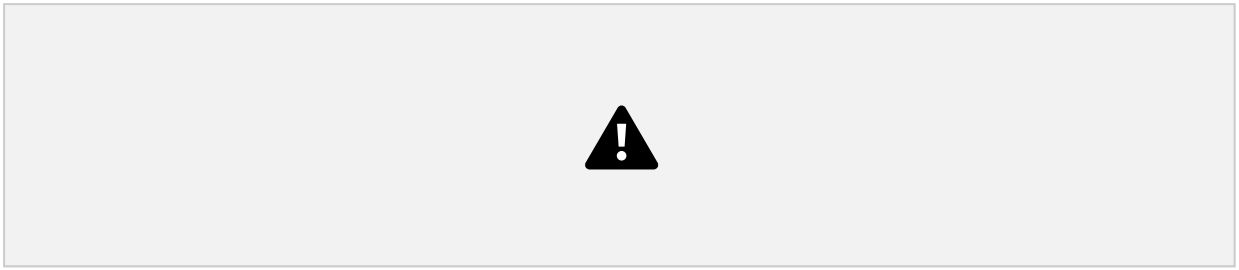
Base-pair Quality Level

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.



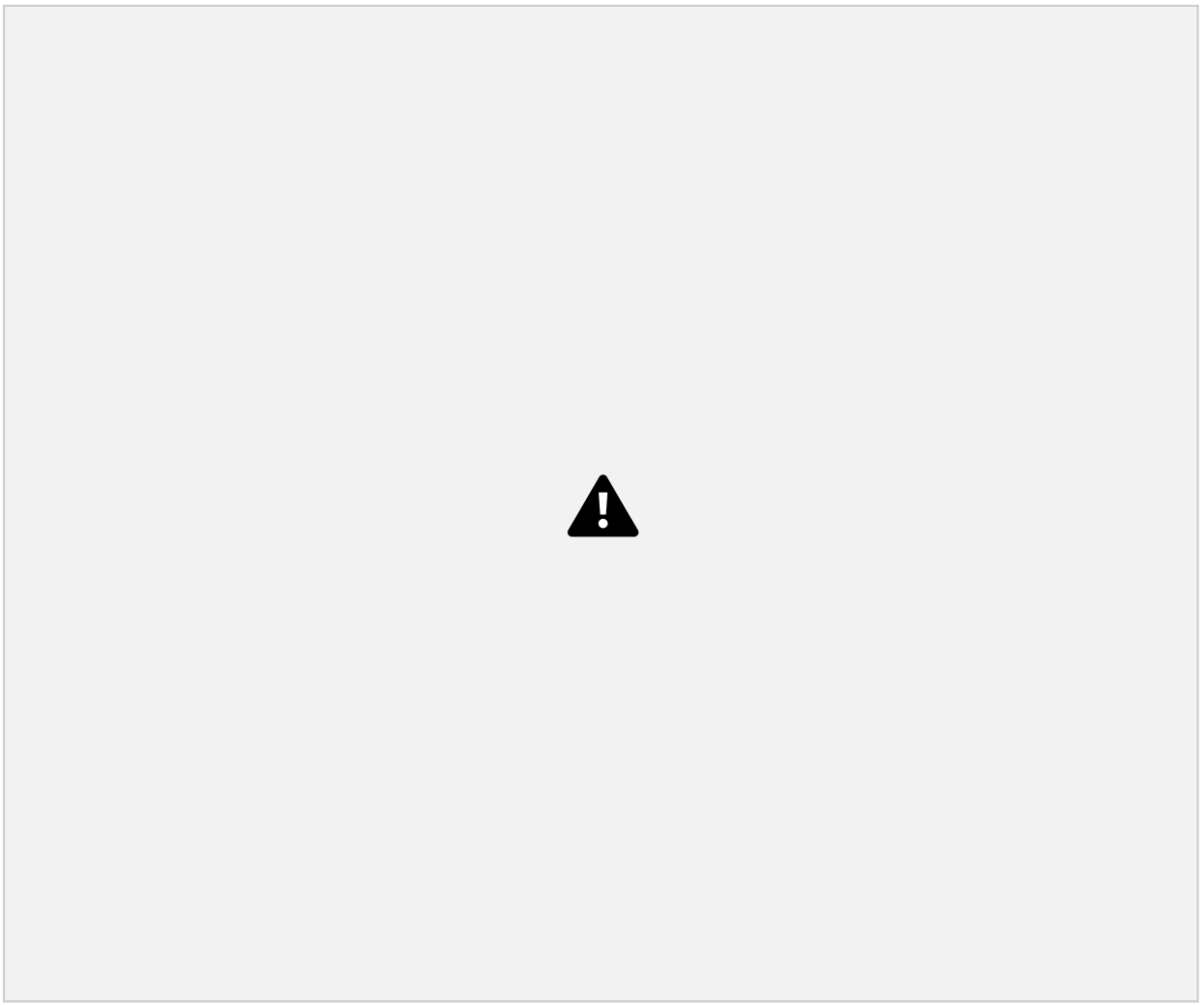
Base-pair Resolution Level

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](#Vertical zoom: examining subset of reads)), only a few reads are shown in the figure below.



Vertical Zoom: Examining Subset of Reads

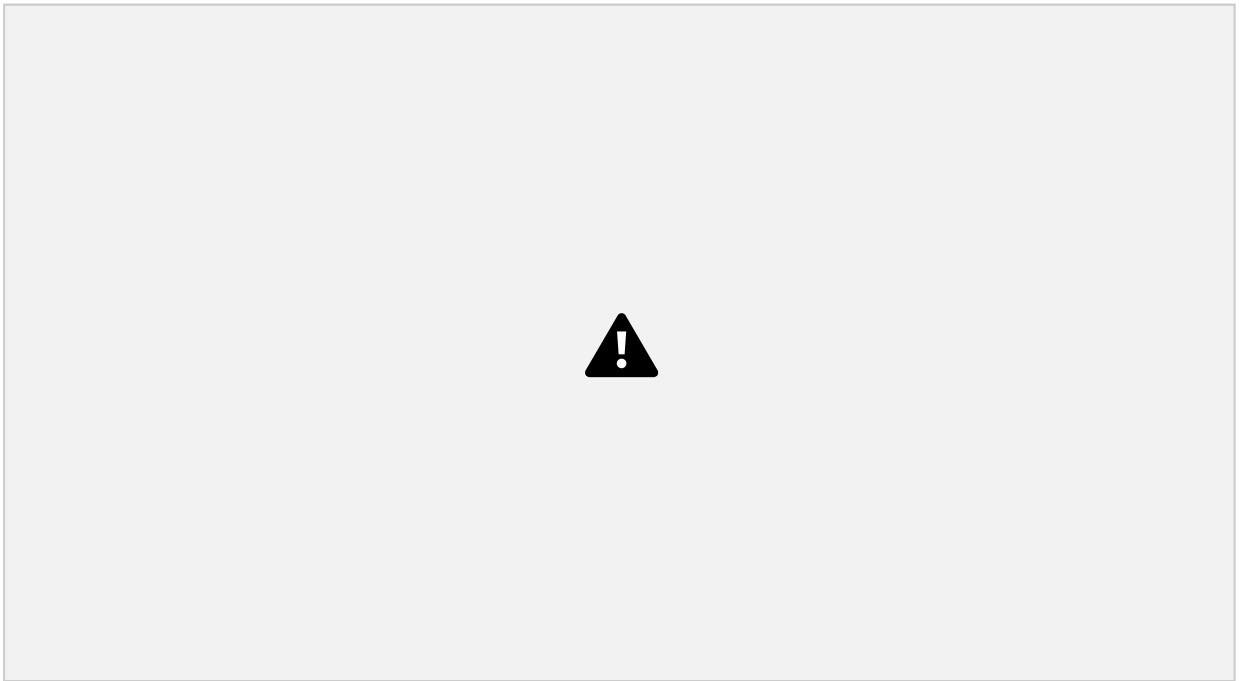
ppBAM can display up to 7000 reads, and will downsample if the number of reads in a region is over 7000. This is especially helpful for displaying high-depth sequencing data. However, displaying nucleotides from each read for such a large number of reads is not feasible. Therefore, the pixel width of 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.



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.

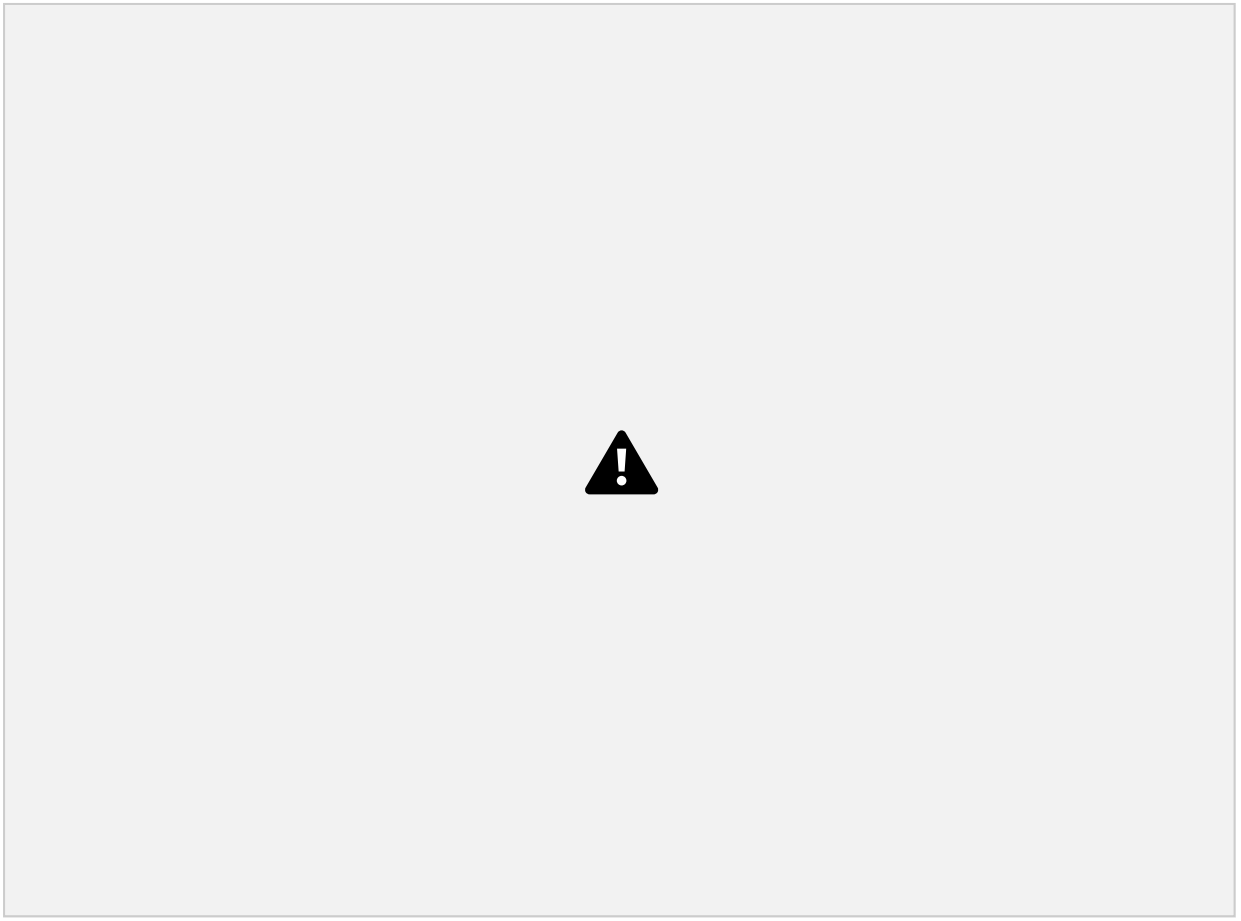
BAM track configuration panel figure



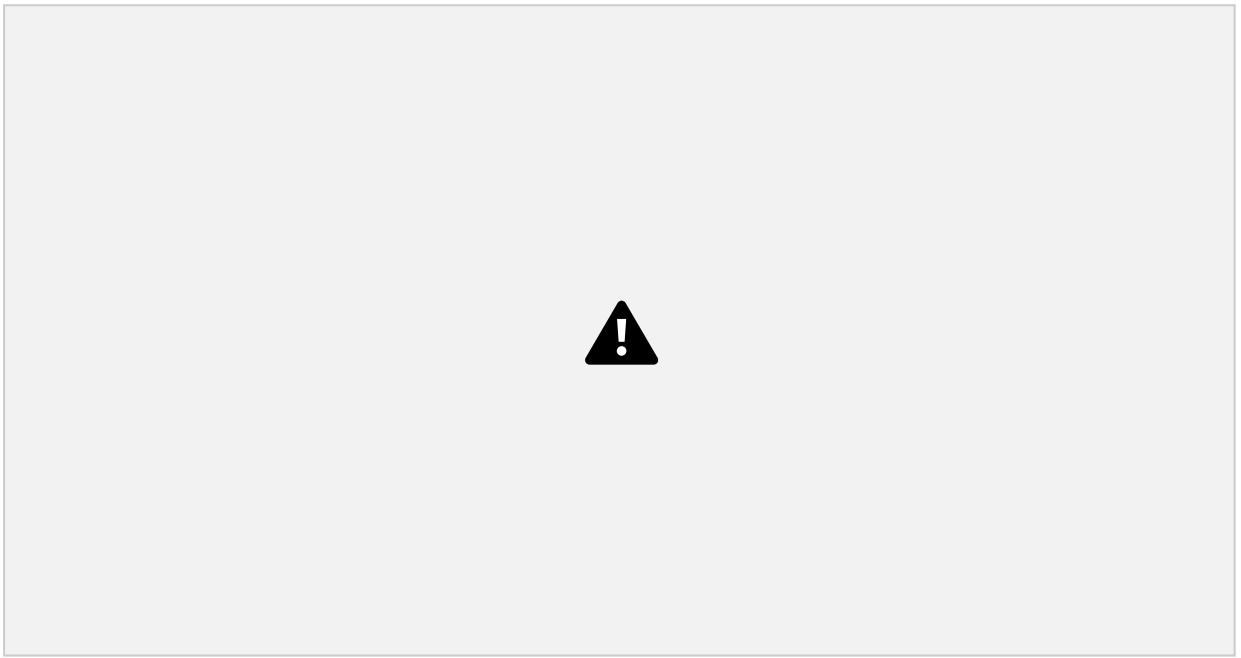
Single and Paired-end Read

[The configuration panel](#BAM track 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.



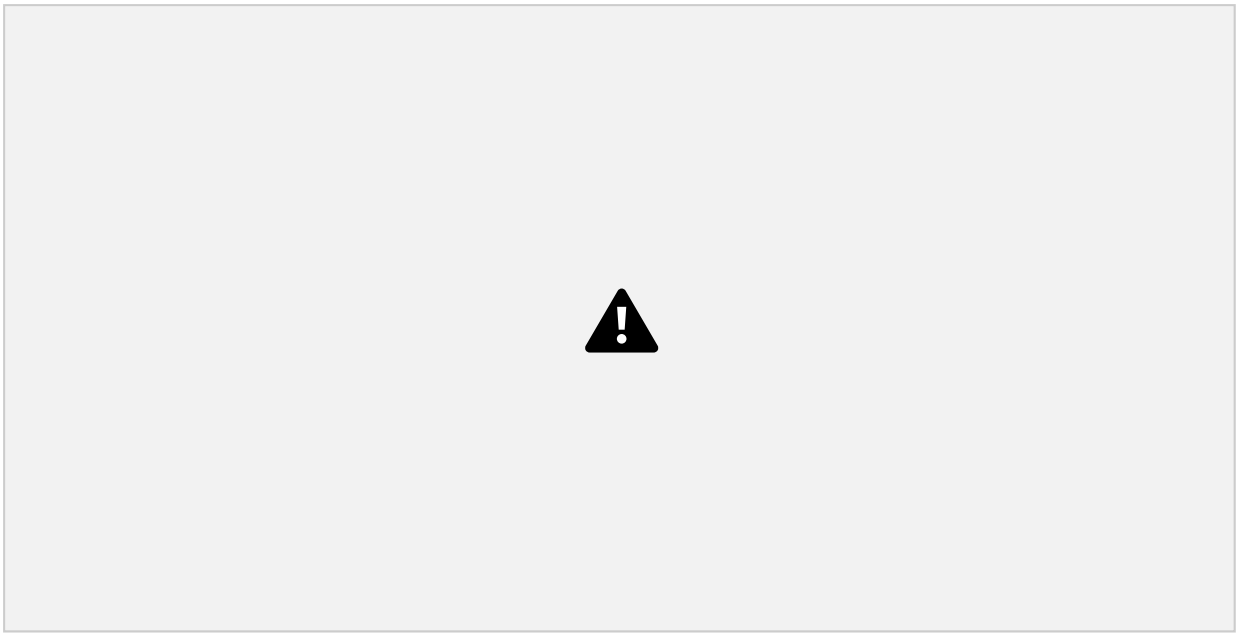
The same track above shows in paired mode.



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](#Base-pair resolution level) and is in nucleotide view ([vertical zoom](#Vertical zoom: examining subset of reads) in case of high-depth sequencing data).

Displaying PCR and Optical Duplicated Reads



The [checkbox](#BAM track configuration panel) 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](#Strictness in on-the-fly genotyping) 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](#BAM track 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](#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](#Using 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.

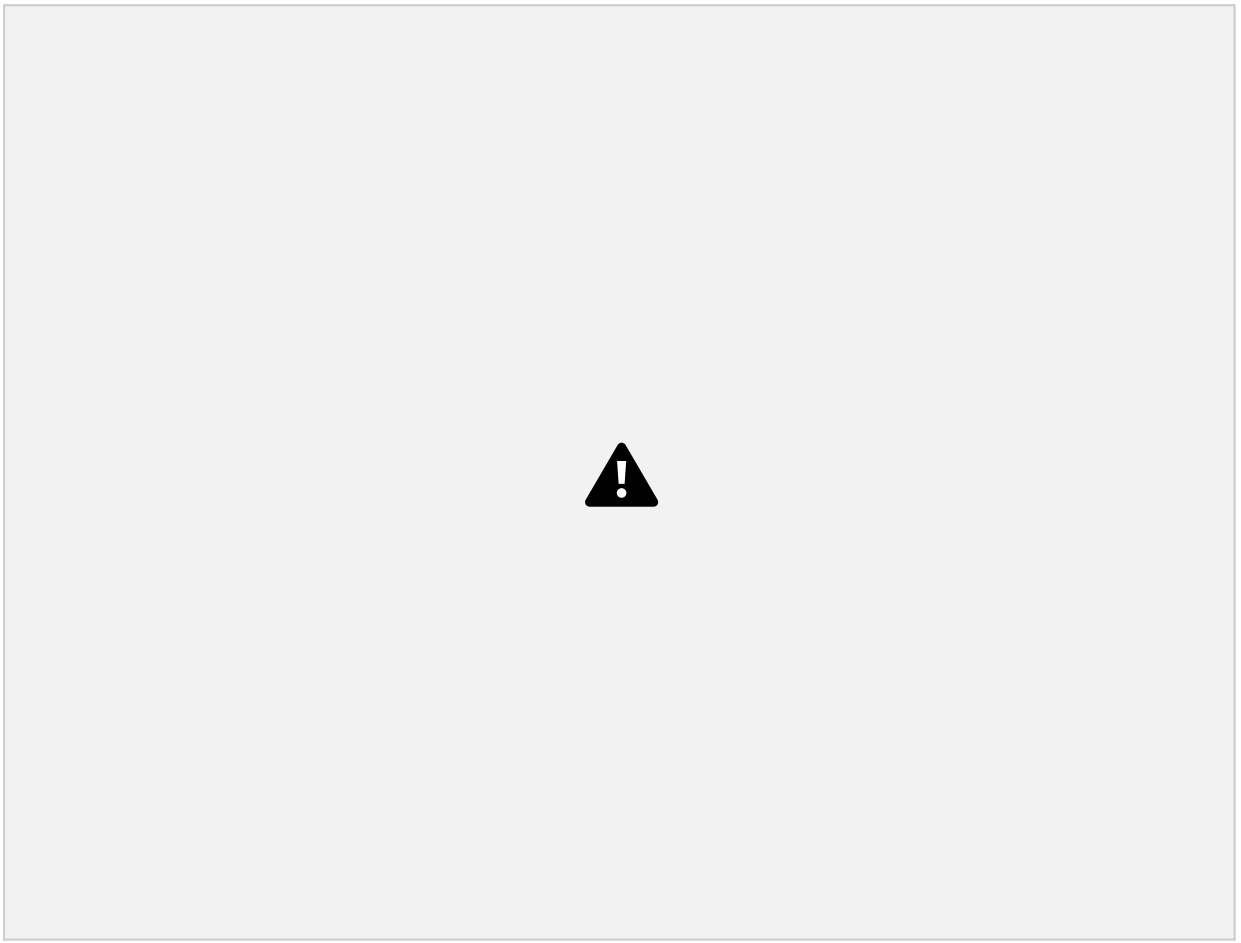
CIGAR -Contains CIGAR sequence of the read.

FLAG -Contains the flag number (from BAM file) of the read.

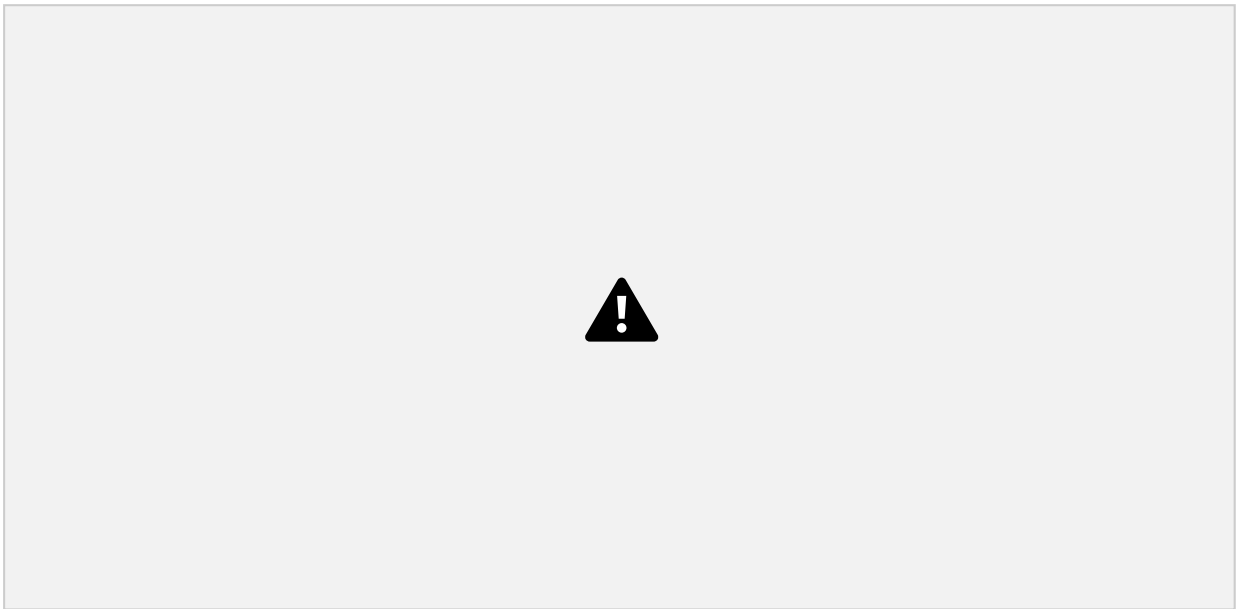
NAME -Contains the name of the read.

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

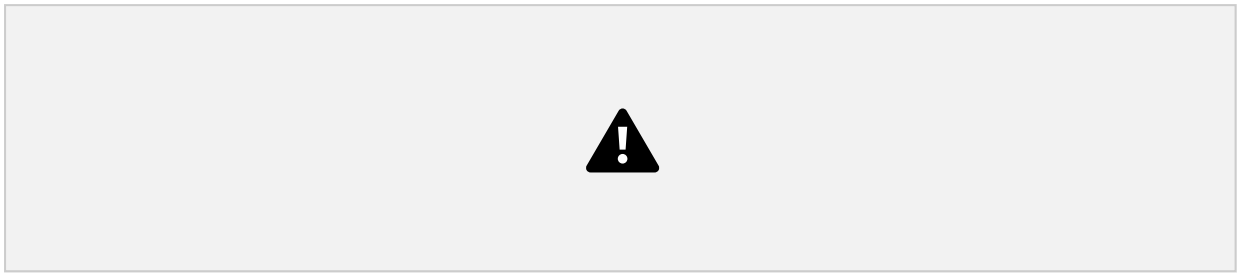
Presence of gray background nucleotides in a read suggests that both the read (at least part of it) and its mate are properly aligned and the insert size is within expected range (as shown below).

Blue

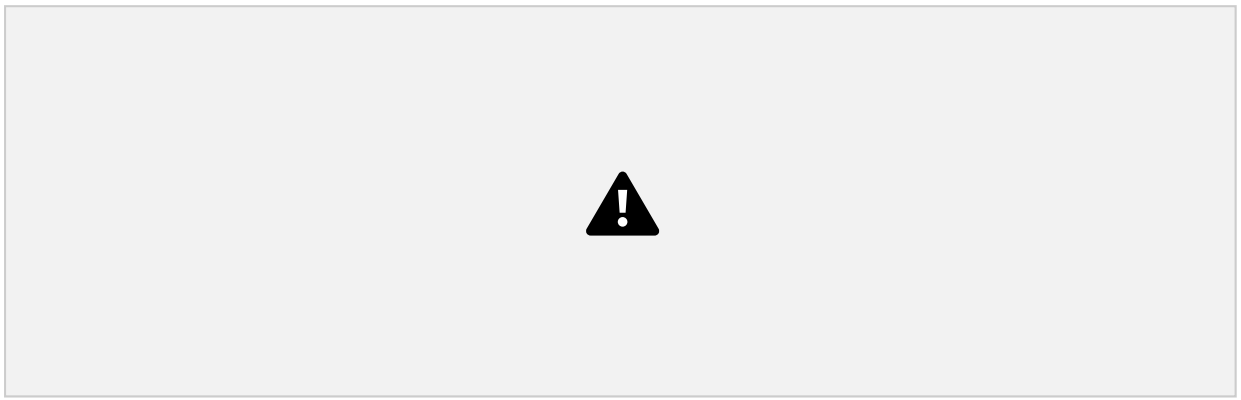
Presence of blue-background nucleotides in a read indicates that part of the read is soft clipped (as shown below). The last 94 nucleotides in the read below are softclipped based on CIGAR sequence (57M94S).

Brown

A brown colored background (in the main [read alignment plot](#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

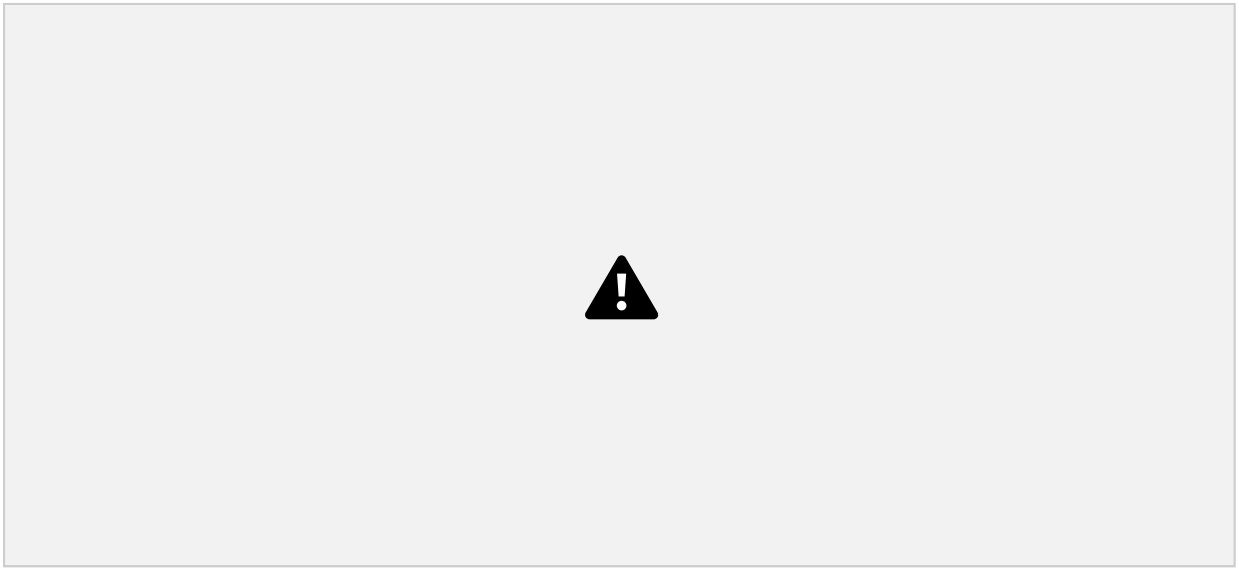
A green background (shown below) indicates that the template has the wrong insert size. As shown in the Read Info figure below, the reads labeled green have a higher insert size than normal (gray) reads because of the structural deletion. In paired end view, generally such read-pairs have a much longer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

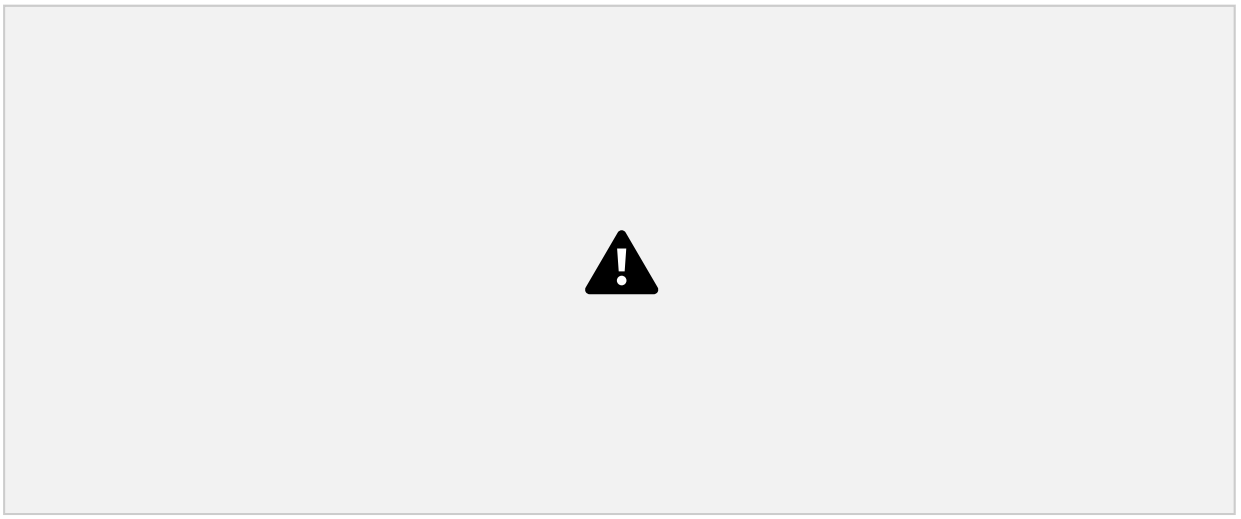
Orange background color indicates the read and its mate are mapped in different chromosomes (as shown below). The displayed read is mapped in chr7:16363-16512 whereas its mate is mapped in chr16.

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

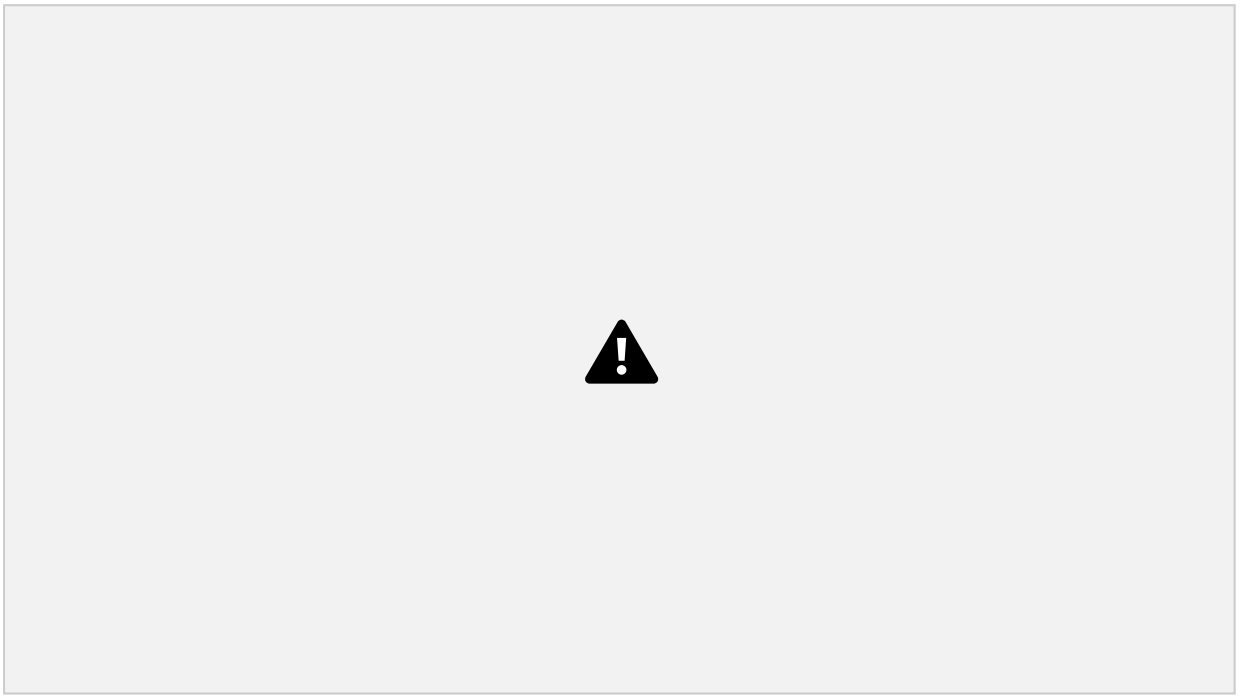
For a given variant (SNV or indel), reads mapping to the variant region are classified into Reference, Alternative, None (neither reference nor alternative allele) and Ambiguous (unclassified reads) groups by using the Smith-Waterman alignment (as shown in figure above). The difference (Diff score) between 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](#Strictness in on-the-fly genotyping) level = 'Strict') and those that have equal similarity (Diff score = 0) to both reference and alternative allele are classified into the [ambiguous](#Ambiguous reads) 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.



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](#Strictness in on-the-fly genotyping) = 'Strict' and into alternative group when [strictness](#Strictness in on-the-fly genotyping) = 'Lenient'. Read4 has equal similarity (Diff score = 0) towards both alternative as well as reference sequences and is classified into the [ambiguous](#Ambiguous reads) group.

Ambiguous Reads

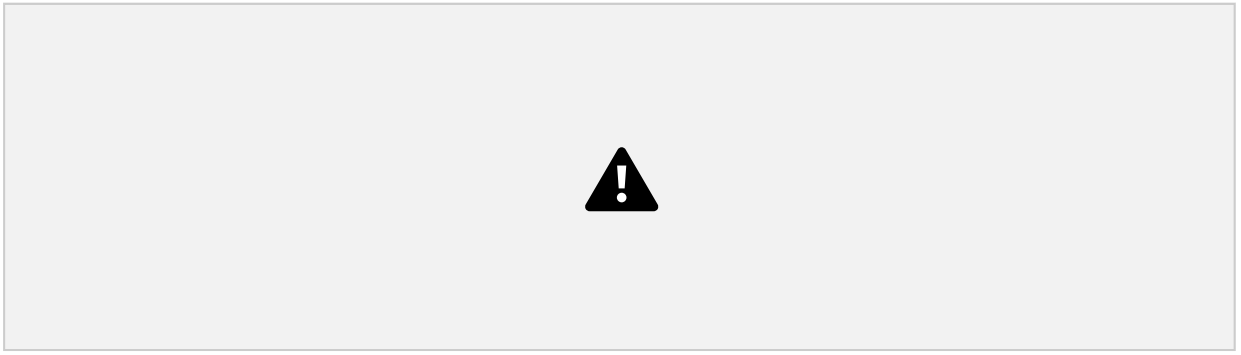
In certain indels such as in the TP53 example, certain reads in the variant region can have equal similarity towards both the reference and alternative allele. A large number of ambiguous reads are on the left-side of the indel (Figure a below) because the deletion starts with the sequence GCAGCGC which 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](#BAM track 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.

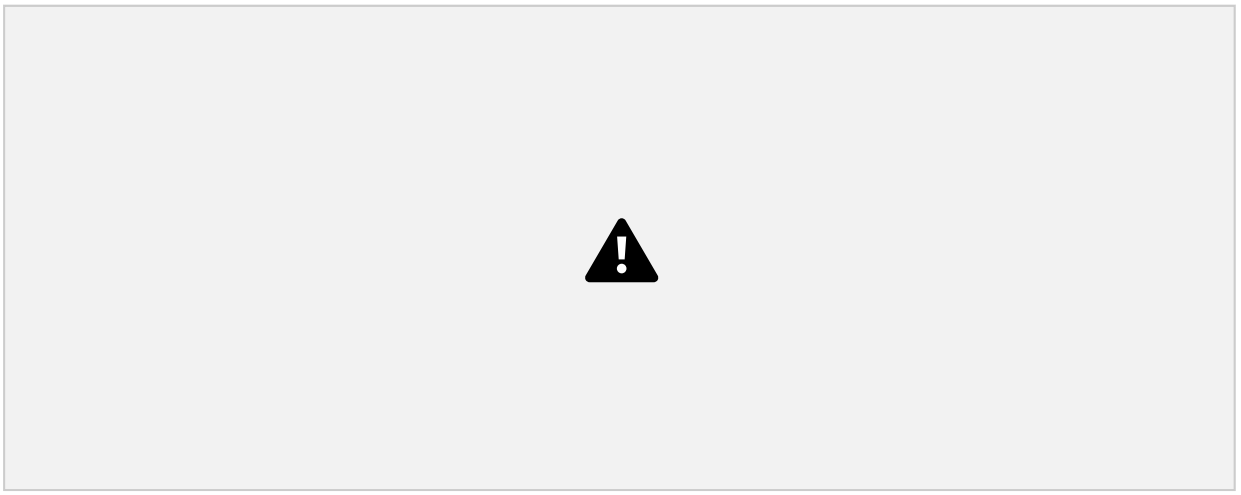
The lenient strictness level can be helpful, when the user wants a lenient estimate of the number of reads supporting the particular indel of interest or when the user is confident that only one alternative allele exists. This can also be helpful when there are reads with low base-pair quality calls 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.

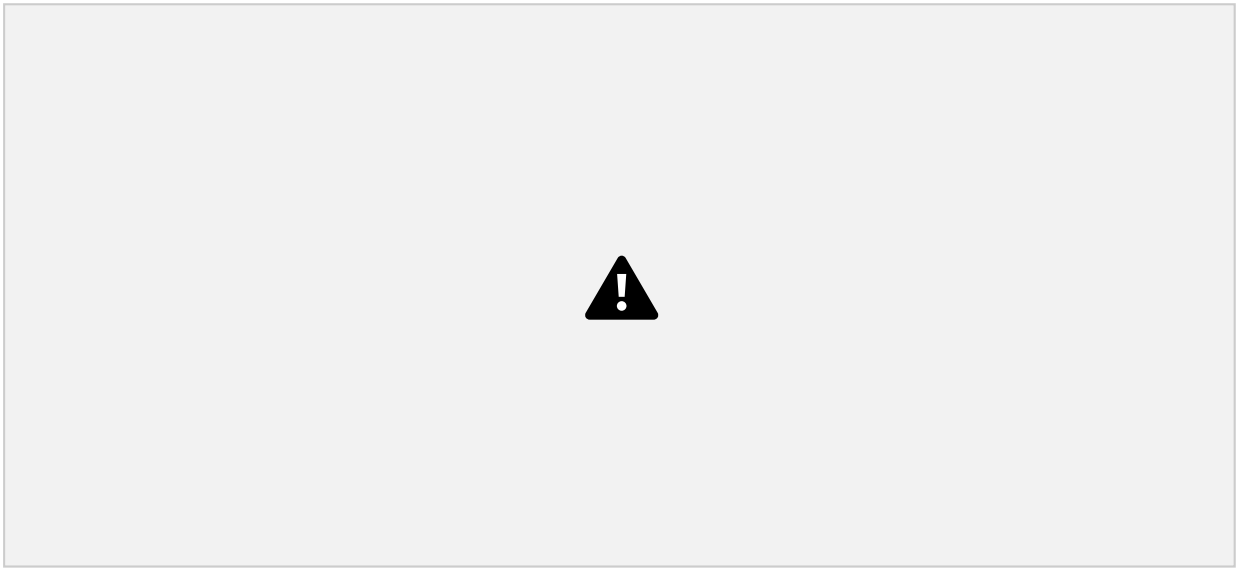


In Figure (c), the reads from (b) are realigned to the alternative allele using Clustal Omega (ClustalO) by clicking on the link showing the number of reads aligned to the alternative allele. This provides an intuitive view confirming the accuracy of the classification of reads to the designated allele. 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](#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.

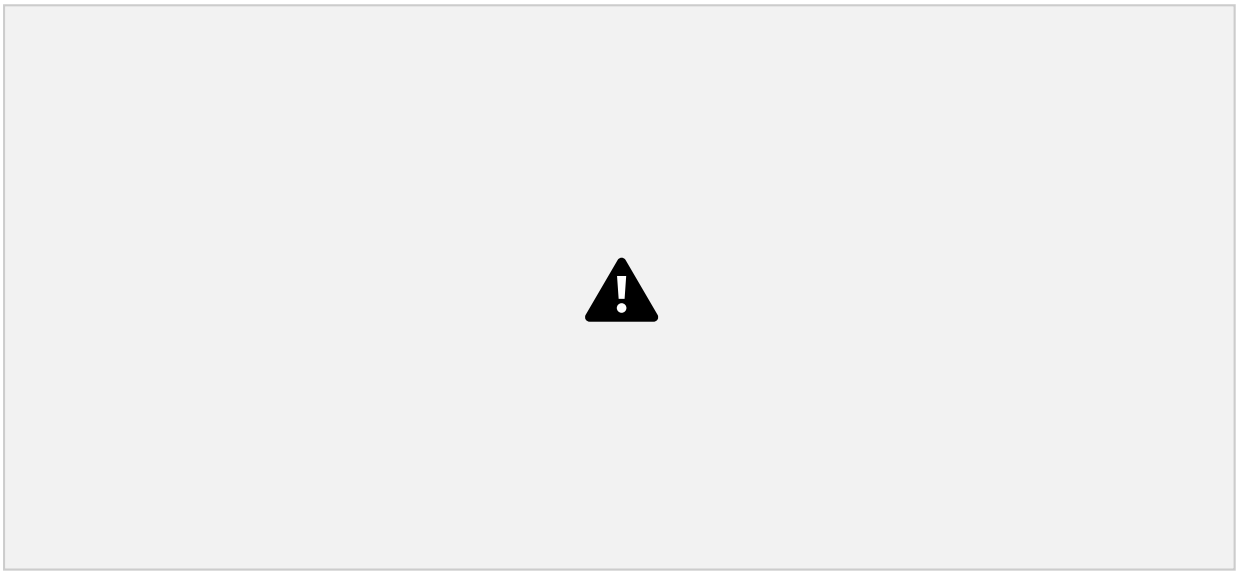


Introduction to OncoMatrix

The GDC OncoMatrix is a handy tool to visualize coding mutations (Simple Somatic Mutations, or SSM) and copy number variations (here onward referred to as CNV).
Each row is a gene and each cell, or a column represents a case. At any point in the tutorial, hover over a symbol or icon for two-three seconds to display more information about that icon.

Accessing the Matrix Chart

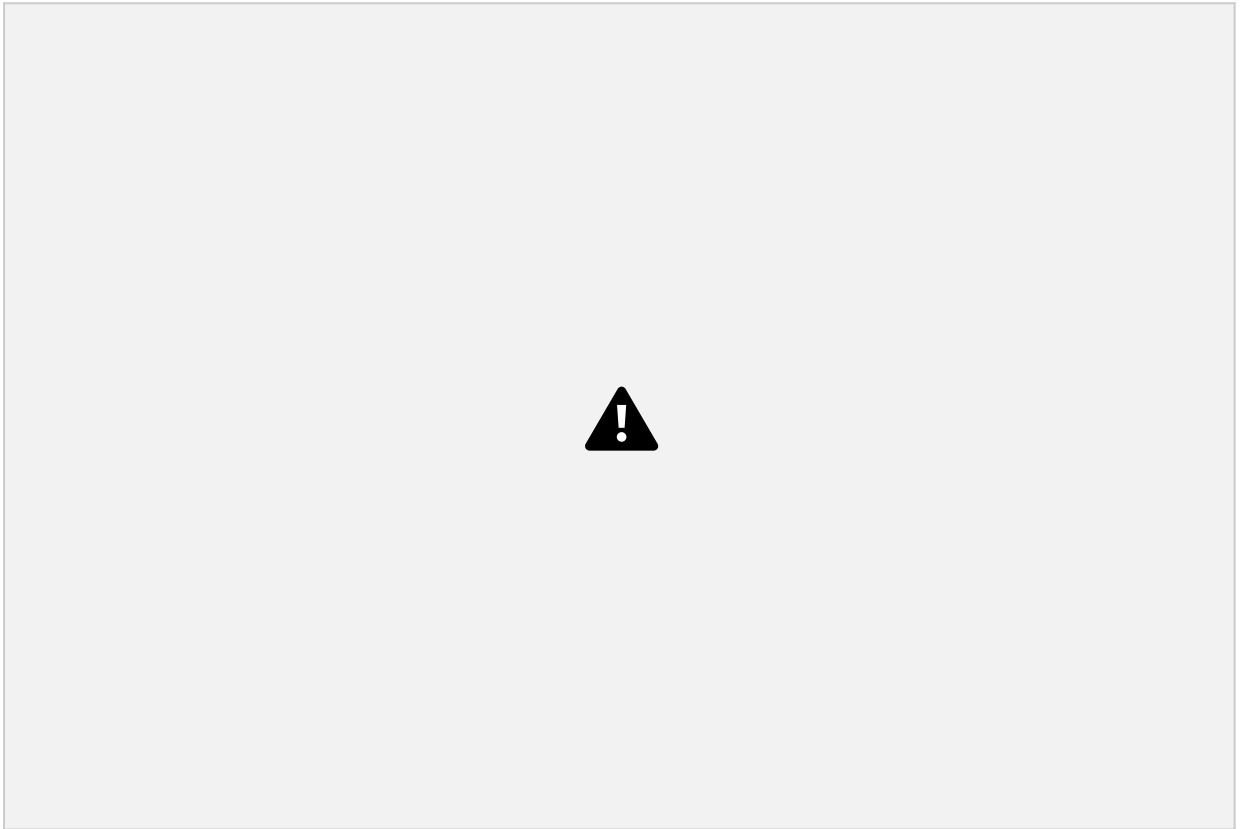
At the Analysis Center, click on the “OncoMatrix” card to launch the app.



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

OncoMatrix 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`.

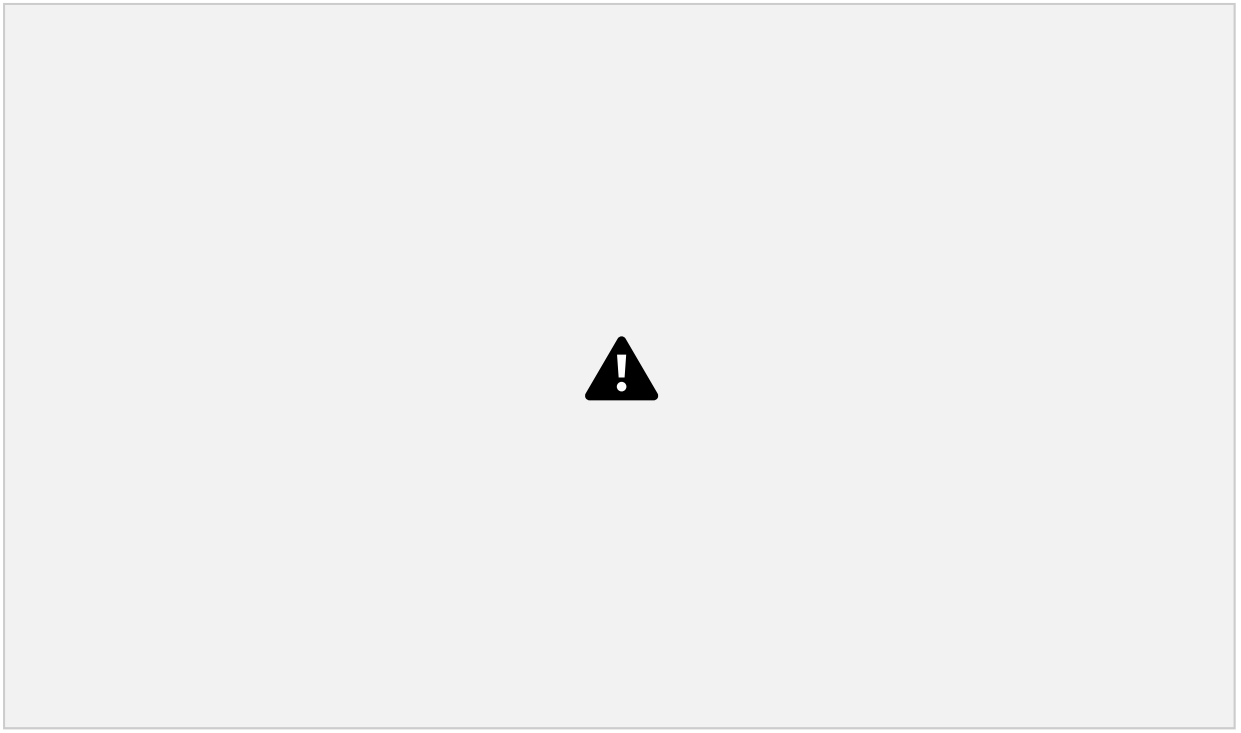


Each of the features and functionalities are described in detail in the following sections.

Matrix plot

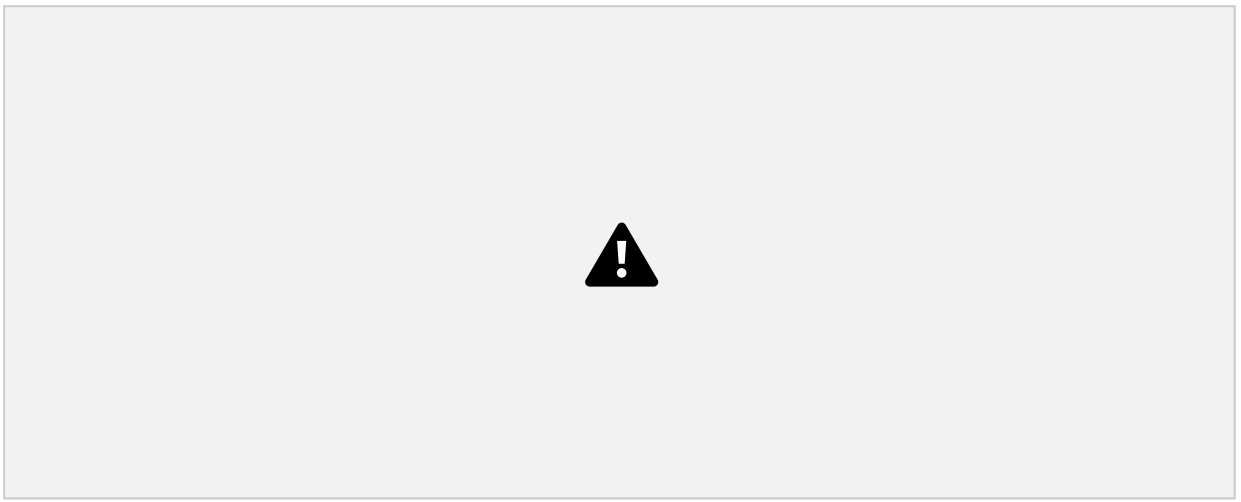
Hovering on sample columns

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 below.



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 below and leave the mouse.



This allows for an automatic zoom as shown below. 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 the 2000 cases.