# GDC Visualization Suite, User Acceptance Testing, April 2017

This guide describes the new data visualization features that can be accessed at the GDC Data Portal. Visualization features include detailed descriptions of mutations, genes, and their frequency; graphical representation of mutation positions, and dynamic survival analysis plots. Additionally, new API endpoints are available for users to programmatically retrieve the data used to generate the visualization features.

The mutation-based visualization features are derived from open-access MAF files that were produced by GDC variant-calling pipelines. The format of these MAF files was developed by and for the GDC and is outlined in the MAF Format documentation.

GDC Portal Home Page

The GDC Portal home page is the entry way to accessing data in the Genomic Data Commons.
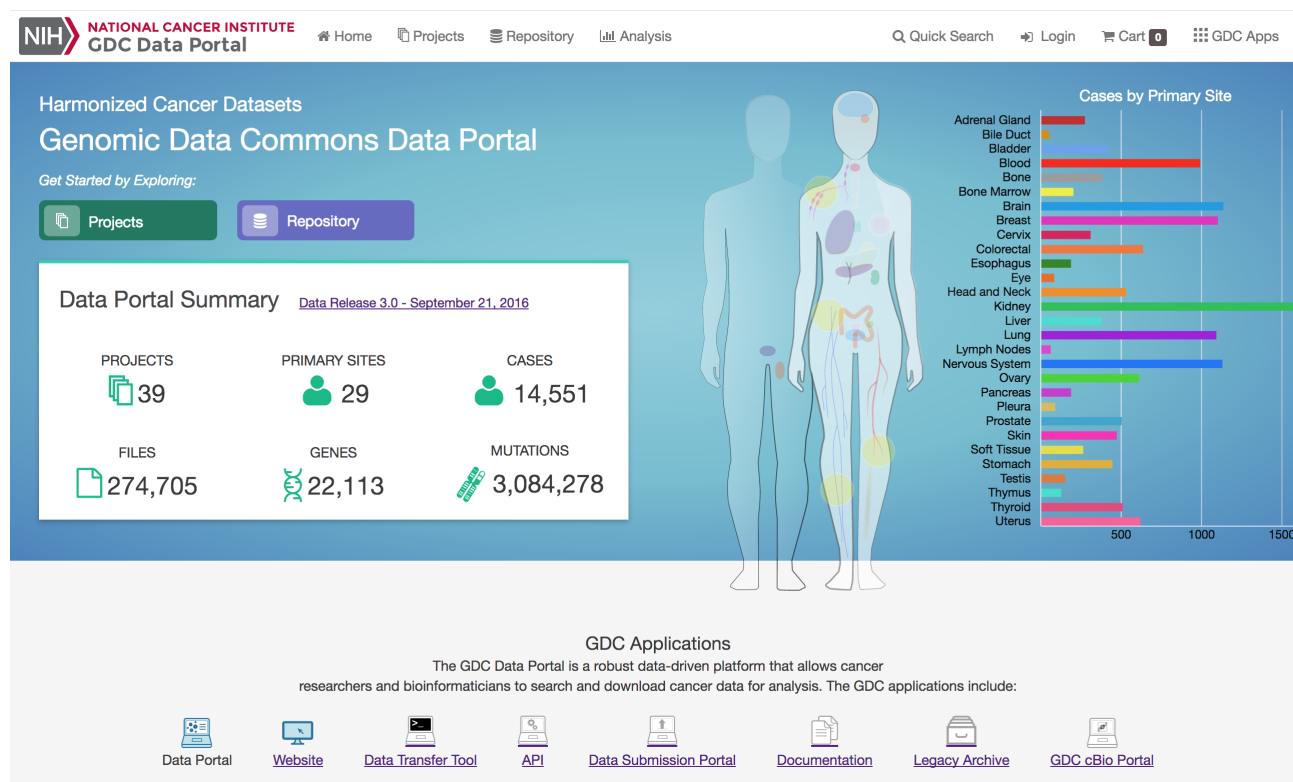
The Genomic Data Commons can be queried by project or by primary site (i.e Brain). Data can be narrowed down in a few ways listed below.

- **Projects**: The projects link directs users to the Project List Page, which gives an overall summary of project-level information.

- **Repository**: The repository link allows users to see the data files available at the GDC and apply file/case filters to narrow down their search.

- **Human Outline**: The home page displays a human anatomical outline that can be used to refine their search. Choosing an associated organ will direct the user to a listing of all projects associated with that primary site. For example, clicking on the human brain will show only cases and projects associated with brain cancer (TCGA-GBM and TCGA-LGG). The number of cases associated with each primary site is also displayed here and separated by project.

Project List Page

The project list page displays statistics about the projects that are available at the GDC.
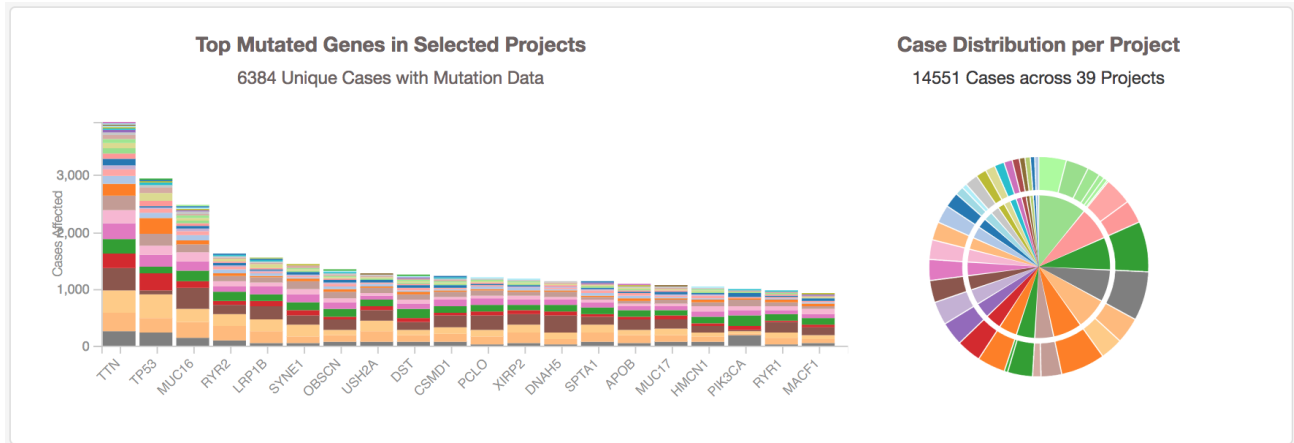
Top Mutated Genes in Selected Projects

This dynamically generated bar graph shows the ten genes with the most mutations across all projects. The bars represent the frequency of each mutation and is broken down into different colored segments by project and disease type. The graphic is updated as filters are applied for projects, programs, disease types, and data categories available in the project.

Hovering the cursor over each bar will display information about the number of cases affected by the disease type and clicking on each bar will bring the user to the Gene Summary Page page for the gene associated with the mutation.

Case Distribution per Project

A pie graph displays the relative number of cases for each primary site (inner circle), which is further divided by project (outer circle). Hovering the cursor over each portion of the graph will display the primary site or project with the number of associated cases. Filtering projects at the left panel will update the pie chart.

Project Table

Showing **1 - 39** of **39** projects

ID	Disease Type	Primary Site	Program	Cases	Available Cases per Data Category							Files
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
TARGET-NBL	Neuroblastoma	Nervous System	TARGET	1,127	270	151	216	0	0	0	0	2,802
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098	1,098	1,097	1,044	1,096	1,095	1,097	1,098	27,207
TARGET-AML	Acute Myeloid Leukemia	Blood	TARGET	988	299	272	8	0	0	0	0	1,869
TARGET-WT	High-Risk Wilms Tumor	Kidney	TARGET	652	128	128	34	0	0	0	0	1,320
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617	406	166	396	593	423	596	617	9,657

General information about each project is displayed below the graphs. The numbers of available cases per data category for each project are displayed as links. Clicking on a specific project will bring a user to that project's Project Detail Page, which provides an overview of all cases, files and annotations available for the project.

Project Detail Page

Each project has a detail page that provides an overview of all available cases, files, and annotations available. Clicking on the numbers in the summary table will display the corresponding data.

PR TCGA-BRCA

⬇ Download manifest   ⬇ Download Clinical   ⬇ Download Biospecimen

⊞ Summary

Project ID	TCGA-BRCA
Project Name	Breast Invasive Carcinoma
Disease Type	Breast Invasive Carcinoma
Primary Site	Breast
Program	TCGA

CASES
1,098

FILES
27,207

ANNOTATIONS
78

Cases and File Counts by Experimental Strategy		
Experimental Strategy	**Cases**	**Files**
■ Genotyping Array	1,096	4,446
■ Methylation Array	1,095	1,234
■ WXS	1,050	10,823
■ RNA-Seq	1,092	4,888
■ miRNA-Seq	1,079	3,621

Cases and File Counts by Data Category		
Data Category	**Cases**	**Files**
■ Raw Sequencing Data	1,098	4,604
■ Transcriptome Profiling	1,097	6,080
■ Simple Nucleotide Variation	1,044	8,648
■ Copy Number Variation	1,096	4,446
■ DNA Methylation	1,095	1,234
■ Clinical	1,097	1,097
■ Biospecimen	1,098	1,098

Three download buttons in the top right corner of the screen allow the user to download the entire project dataset, along with the associated project metadata:

- **Download 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.
- **Download Clinical**: Downloads clinical metadata about all cases in the project.
- **Download Biospecimen**: Downloads biospecimen metadata associated with all cases in the project.

Most Frequently Mutated Genes

The project detail page also reports the genes that have somatic mutations in the greatest numbers of cases in a graphical and tabular format.

Most Frequently Mutated Genes

Distribution of Most Frequently Mutated Genes

Overall Survival Plot
1044 Cases in TCGA-BRCA

Showing **1 - 10** of **19,038** genes

Symbol	Name	Cytoband	# Affected Cases in TCGA-BRCA	# Affected Cases Across all Projects	Survival Analysis
TP53	tumor protein p53	17p13.1	344 / 1044 (32.95%)	3,864 / 14,551 (26.55%) ◀	
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	3q26.32	329 / 1044 (31.51%)	1,373 / 14,551 (9.44%) ◀	
TTN	titin	2q31.2	251 / 1044 (24.04%)	3,931 / 14,551 (27.02%) ◀	
MUC16	mucin 16, cell surface associated	19p13.2	149 / 1044 (14.27%)	2,467 / 14,551 (16.95%) ◀	
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	16q22.1	141 / 1044 (13.51%)	372 / 14,551 (2.56%) ◀	
GATA3	GATA binding protein 3	10p14	135 / 1044 (12.93%)	356 / 14,551 (2.45%) ◀	
KMT2C	lysine (K)-specific methyltransferase 2C	7q36.1	96 / 1044 (9.20%)	992 / 14,551 (6.82%) ◀	
RYR2	ryanodine receptor 2 (cardiac)	1q43	92 / 1044 (8.81%)	1,617 / 14,551 (11.11%) ◀	
MAP3K1	mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase	5q11.2	85 / 1044 (8.14%)	321 / 14,551 (2.21%) ◀	
SYNE1	spectrin repeat containing, nuclear envelope 1	6q25.2	80 / 1044 (7.66%)	1,609 / 14,551 (11.06%) ◀	

The top of this section also contains a survival plot of all the cases within the specified project, in addition to a bar graph of the most frequently mutated genes. Hovering over each bar in the plot will display information about the number of cases affected. Users may choose to download the underlying data in JSON or TSV format or an image of the graph in SVG or PNG format by clicking the `download` icon at the top of each graph.

Below these graphs is a tabular view of the genes affected, which includes the following information:

- **Symbol:** The gene symbol, which links to the Gene Summary Page
- **Name:** Full name of the gene
- **Cytoband:** The location of the mutation on the chromosome in terms of Giemsa-stained samples.
- **# Affected Cases in Project:** The number of cases within the project that contain a mutation on this gene
- **# Affected Cases Across all Projects:** The number of cases within all the projects in the GDC that contain a mutation on this gene. Clicking the red arrow will display the cases broken down by project
- **# Mutations:** The number of SSMs (simple somatic mutations) detected in that gene
- **Survival Analysis:** An icon that, when clicked, will plot the survival rate between cases in the project with mutated and non-mutated forms of the gene

Survival Analysis

Survival analysis is used to analyze the occurrence of event data over time. In the GDC, survival analysis is performed on the mortality of the cases. Survival analysis requires:

- Data on the time to a particular event (days to death or last follow up)

- - - Fields: **diagnoses.days_to_death** and **diagnoses.days_to_last_follow_up**
  - Information on whether the event has occurred (alive/deceased)
    - Fields: **diagnoses.vital_status**
  - Data split into different categories or groups (i.e. gender, etc.)
    - Fields: **demographic.gender**

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

$$S(t_i) = \prod_{t_j \le t}\left(1 - \frac{d_i}{n_i}\right)$$

Where:

- $S(t_i)$ is the estimated survival probability for any particular one of the t time periods
- n_i is the number of subjects at risk at the beginning of time period t_i
- and d_i is the number of subjects who die during time period t_i

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

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

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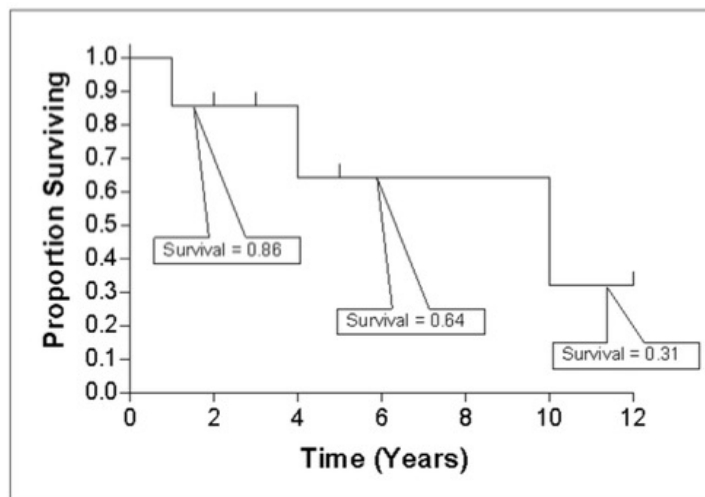
OncoGrid

The project detail page includes an OncoGrid plot of the cases with the most mutations, for the top 50 mutated genes affected by high impact mutations. Genes displayed on the left of the grid (Y-axis) correspond to individual cases on the bottom of the grid (X-axis).
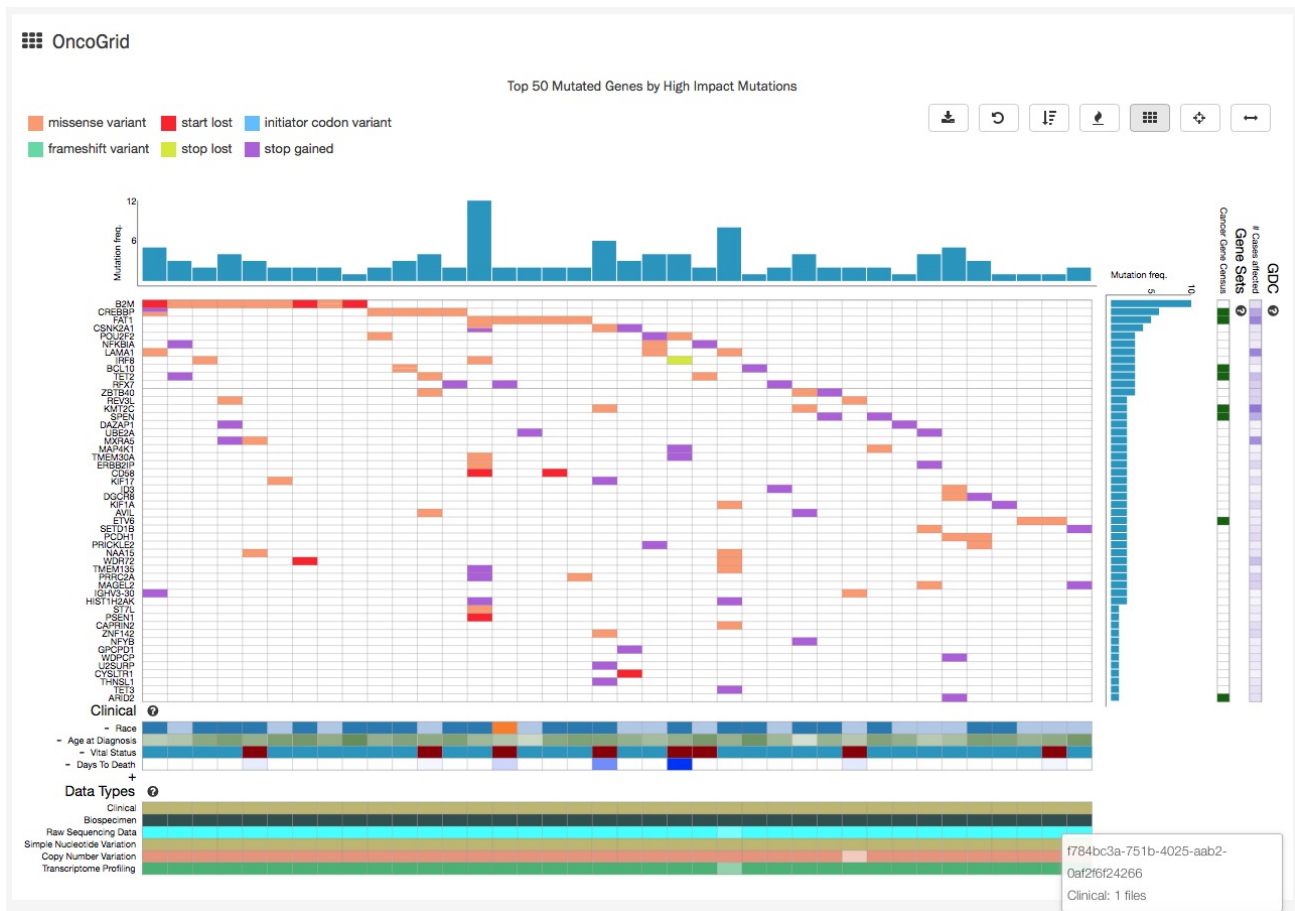
The grid is color-coded with a legend at the top left which describes what type of mutation consequence is observed for each gene/case combination. Clinical information and the available data for each case are available at the bottom of the grid.

The right side of the grid displays additional information about the genes:

- **Gene Sets:** Describes whether a gene is part of the Gene Census. (The cancer Gene Census is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer)
- **GDC:** Heat-map of all cases in the GDC affected with a mutation in this gene

OncoGrid Options

To facilitate readability and comparisons, drag-and-drop can be used to reorder the gene rows. Double clicking a row in the "# Cases Affected" bar at the right side of the graphic will bring the user to the respective Gene Summary Page. Hovering over a cell will display information about the mutation such as its ID, affected case, and biological consequence. Clicking on the cell will bring the user to the respective Mutation Summary page.

A tool bar at the top right of the graphic allows the user to export the data as a JSON object, PNG image, or SVG image. Seven buttons are available in this toolbar:

- **Download:** Users can choose to export the contents either to a static image file (PNG or SVG format) or the underlying data in JSON format
- **Refresh:** Sets all OncoGrid rows, columns, and zoom levels back to their initial positions
- **Cluster Data:** Clusters the rows and columns to place mutated genes with the same cases and cases with the same mutated genes together
- **Toggle Heatmap:** The view can be toggled between cells representing mutation consequences or number of mutations in each gene
- **Toggle Gridlines:** Turn the gridlines on and off
- **Toggle Crosshairs:** Turns crosshairs on, so that users can zoom into specific sections of the OncoGrid

- **Fullscreen:** Turns Fullscreen mode on/off

Most Frequent Mutations

The project detail page also displays the 20 most frequent mutations in the project as a bar graph that indicates the number of cases that are affected by each mutation. Hovering over each bar in the plot will show information about the number of cases affected.
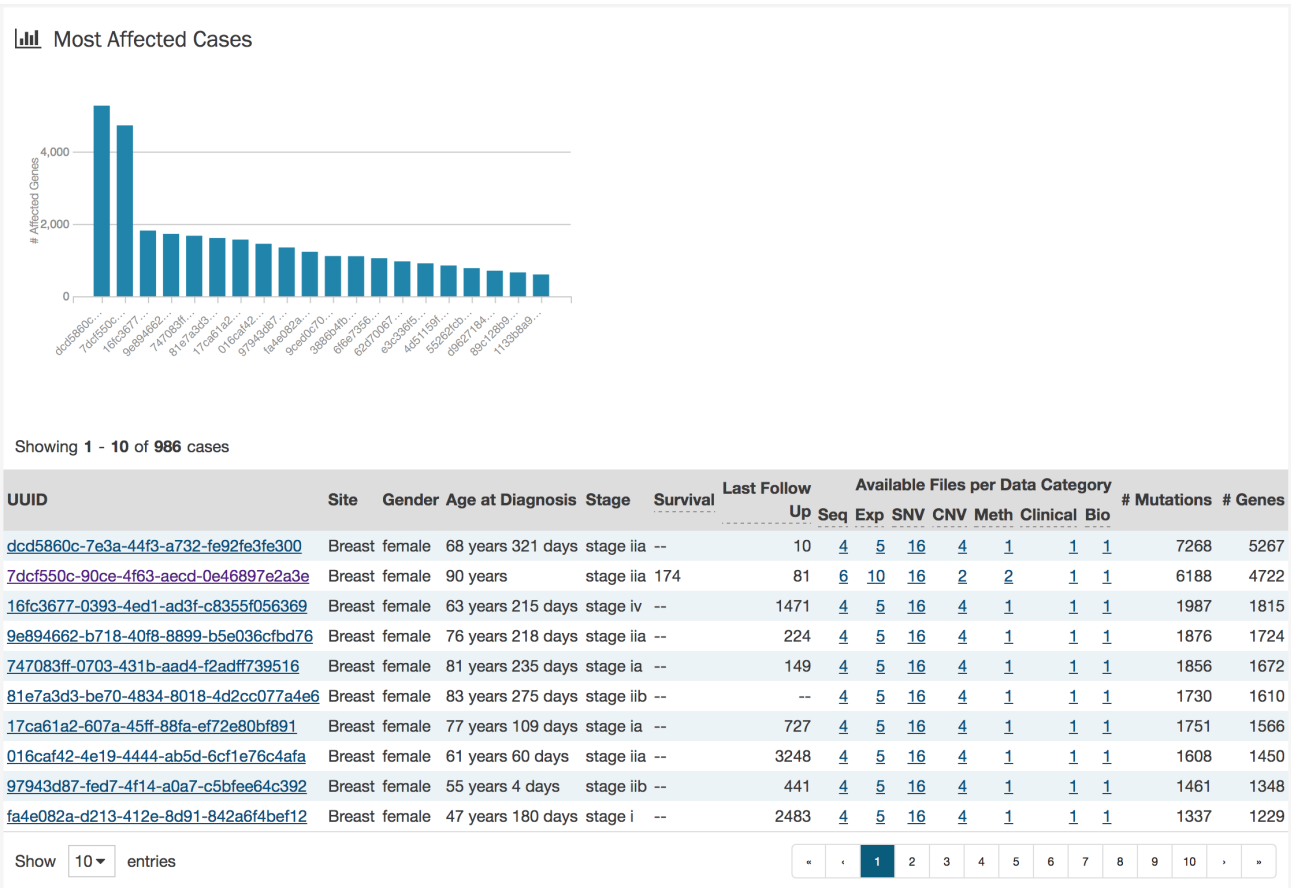
📊 Most Frequent Mutations

Distribution of Most Frequent Mutations

Overall Survival Plot
1044 Cases in TCGA-BRCA
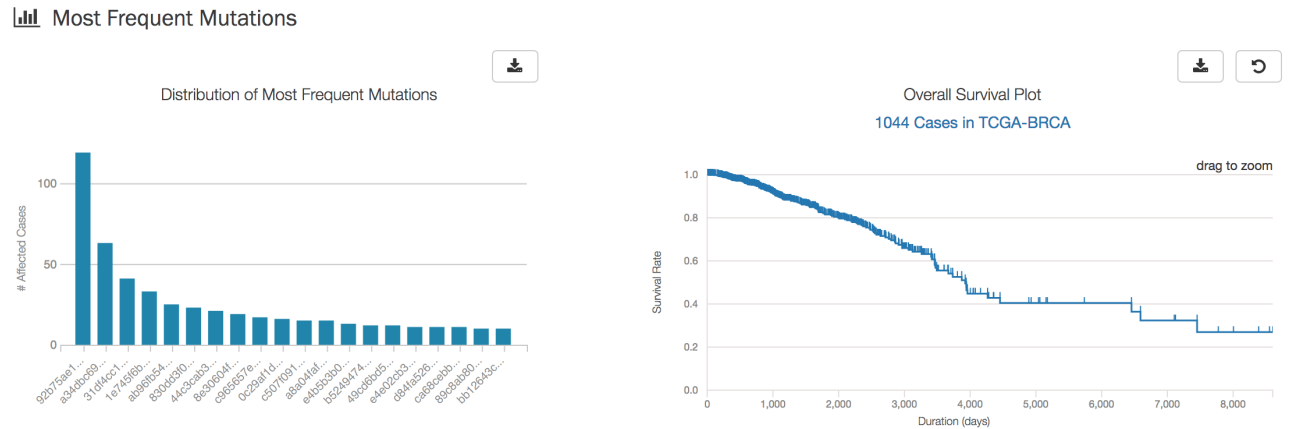
Showing **1 - 10** of **126,949** mutations

ID	DNA Change	Type	Consequences	# Affected Cases in TCGA-BRCA	# Affected Cases Across all Projects	Impact (VEP)	Survival Analysis
92b75ae1...	chr3:g.179234297 A>G	Substitution	**Missense** ENSG00000121879 p.H1047R	119 / 986 (12.07%)	230 / 10,193 (2.26%) ◄	M	📈
a34dbc69...	chr3:g.179218303 G>A	Substitution	**Missense** ENSG00000121879 p.E545K	63 / 986 (6.39%)	254 / 10,193 (2.49%) ◄	M	📈
31df4cc1...	chr3:g.179218294 G>A	Substitution	**Missense** ENSG00000121879 p.E542K	41 / 986 (4.16%)	157 / 10,193 (1.54%) ◄	M	📈
1e745f6b...	chr1:g.76576946_ 76576947insAAA C	Insertion	**Intron** ENSG00000184005	33 / 986 (3.35%)	75 / 10,193 (0.74%) ◄	MO	📈
ab96fb54...	chr14:g.10478021 4C>T	Substitution	**Missense** ENSG00000142208 p.E17K	25 / 986 (2.54%)	53 / 10,193 (0.52%) ◄	M	📈
830dd3f0...	chr2:g.11612342T >G	Substitution	**Intron** ENSG00000196208	23 / 986 (2.33%)	28 / 10,193 (0.27%) ◄	MO	📈

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

- **ID:** A UUID for the mutation assigned by the GDC, when clicked will bring a user to the Mutation Summary Page
- **DNA Change:** The chromosome and starting coordinates of the mutation are displayed along with the nucleotide differences between the reference and tumor allele
- **Type:** A general classification of the mutation
- **Consequences:** The effects the mutation has on the gene coding for a protein (i.e. synonymous, missense, non-coding transcript). A link to the Gene Summary Page for the gene affected by the mutation is included
- **# Affected Cases in Project:** The number of affected cases in the project
- **# Affected Cases in Across all Projects:** The number of affected cases, expressed as number across all projects. Choosing the arrow next to the percentage will display a breakdown of each affected project
- **Impact:** A subjective classification of the severity of the variant consequence. The categories are:
  - **HIGH**: 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**: A non-disruptive variant that might change protein effectiveness
  - **LOW**: Assumed to be mostly harmless or unlikely to change protein behavior
- **Survival Analysis:** An icon that when clicked, will plot the survival rate between the gene's mutated and non-mutated cases

Most Affected Cases

The final section of the project detail page is a display of the top 20 cases in a specified project, with the greatest number of affected genes.

Most Affected Cases

Showing **1 - 10** of **986** cases

UUID	Site	Gender	Age at Diagnosis	Stage	Survival	Last Follow Up	Seq	Exp	SNV	CNV	Meth	Clinical	Bio	# Mutations	# Genes
dcd5860c-7e3a-44f3-a732-fe92fe3fe300	Breast	female	68 years 321 days	stage iia	--	10	4	5	16	4	1	1	1	7268	5267
7dcf550c-90ce-4f63-aecd-0e46897e2a3e	Breast	female	90 years	stage iia	174	81	6	10	16	2	2	1	1	6188	4722
16fc3677-0393-4ed1-ad3f-c8355f056369	Breast	female	63 years 215 days	stage iv	--	1471	4	5	16	4	1	1	1	1987	1815
9e894662-b718-40f8-8899-b5e036cfbd76	Breast	female	76 years 218 days	stage iia	--	224	4	5	16	4	1	1	1	1876	1724
747083ff-0703-431b-aad4-f2adff739516	Breast	female	81 years 235 days	stage ia	--	149	4	5	16	4	1	1	1	1856	1672
81e7a3d3-be70-4834-8018-4d2cc077a4e6	Breast	female	83 years 275 days	stage iib	--	--	4	5	16	4	1	1	1	1730	1610
17ca61a2-607a-45ff-88fa-ef72e80bf891	Breast	female	77 years 109 days	stage ia	--	727	4	5	16	4	1	1	1	1751	1566
016caf42-4e19-4444-ab5d-6cf1e76c4afa	Breast	female	61 years 60 days	stage iia	--	3248	4	5	16	4	1	1	1	1608	1450
97943d87-fed7-4f14-a0a7-c5bfee64c392	Breast	female	55 years 4 days	stage iib	--	441	4	5	16	4	1	1	1	1461	1348
fa4e082a-d213-412e-8d91-842a6f4bef12	Breast	female	47 years 180 days	stage i	--	2483	4	5	16	4	1	1	1	1337	1229

Show 10 ▾ entries

« ‹ 1 2 3 4 5 6 7 8 9 10 › »

Below the bar graph is a table contains information about these cases:

- **UUID:** The UUID of the case, which links to the Case Summary Page
- **Site:** The anatomical location of the site affected
- **Gender:** Text designations that identify gender. Gender is described as the assemblage of properties that distinguish people on the basis of their societal roles
- **Age at Diagnosis:** Age at the time of diagnosis expressed in number of days since birth
- **Stage:** The extent of a cancer in the body. Staging is usually based on the size of the tumor, whether lymph nodes contain cancer, and whether the cancer has spread from the original site to other parts of the body. The accepted values for tumor_stage depend on the tumor site, type, and accepted staging system
- **Survival:** The number of days until death
- **Last Follow Up:** Time interval from the date of last follow up to the date of initial pathologic diagnosis, represented as a calculated number of days
- **Available Files per Data Category:** Five columns displaying the number of files available in each of the five data categories. These link to the files for the specific case.
- **Genes:** The number of genes affected by mutations for the case

Case Summary Page

The Case Summary Page displays case details including the project and disease information, data files that are available for that case, and the experimental strategies employed. A button in the top-right corner of the page allows the user to add all files associated with the case to the file cart.

CA 7dcf550c-90ce-4f63-aecd-0e46897e2a3e

🛒 Add all files to the cart

⊞ Summary

Case UUID	7dcf550c-90ce-4f63-aecd-0e46897e2a3e
Case Submitter ID	TCGA-AC-A23H
Project ID	TCGA-BRCA
Project Name	Breast Invasive Carcinoma
Disease Type	Breast Invasive Carcinoma
Program	TCGA
Primary Site	Breast

FILES
38 📄

ANNOTATIONS
0 ✏️

File Counts by Experimental Strategy 🥧

Experimental Strategy	Files
Genotyping Array	2
Methylation Array	2
WXS	18
RNA-Seq	8
miRNA-Seq	6

File Counts by Data Category 🥧

Data Category	Files
Raw Sequencing Data	6
Transcriptome Profiling	10
Simple Nucleotide Variation	16
Copy Number Variation	2
DNA Methylation	2
Clinical	1
Biospecimen	1

Clinical and Biospecimen Information

The page also provides clinical and biospecimen information about that case. Links to export clinical and biospecimen information in JSON format are provided.

Clinical ⬇ Export

Demographic	Diagnoses / Treatments (1)	Family Histories (0)	Exposures (1)

ID	d40005ad-6bb7-5b32-9f2d-2ab2394dd0ba
Ethnicity	not hispanic or latino
Gender	female
Race	white
Year Of Birth	1919
Year Of Death	--

Biospecimen ⬇ Export

🔍 Search   Collapse All

- ⊟ **Samples**
  - ⚒ TCGA-AC-A23H-11A ▸
    - ⊟ **Portions**
      - ⚒ TCGA-AC-A23H-11A-12
        - ⊟ **Analytes**
          - ⚒ TCGA-AC-A23H-11A-12D
            - ⊟ **Aliquots**
              - ⚒ TCGA-AC-A23H-11A-12D-A17G-09
              - ⚒ TCGA-AC-A23H-11A-12D-A161-05
              - ⚒ TCGA-AC-A23H-11A-12D-A160-01
              - ⚒ TCGA-AC-A23H-11A-12D-A159-09
              - ⚒ TCGA-AC-A23H-11A-12D-A158-02
          - ⚒ TCGA-AC-A23H-11A-12R
            - ⊟ **Aliquots**
              - ⚒ TCGA-AC-A23H-11A-12R-A156-13
              - ⚒ TCGA-AC-A23H-11A-12R-A157-07

Submitter ID	TCGA-AC-A23H-11A
Sample ID	7df59ca8-7e51-4581-9d7f-8bba0395ce17
Sample Type	Solid Tissue Normal
Sample Type Id	11
Tissue Type	--
Tumor Code	--
Tumor Code Id	--
Oct Embedded	false
Shortest Dimension	--
Intermediate Dimension	--
Longest Dimension	--
Is Ffpe	false
Pathology Report Uuid	--
Tumor Descriptor	--
Current Weight	--
Initial Weight	70
Composition	--

For clinical records that support multiple records of the same type (Diagnoses, Family Histories, or Exposures), a UUID of the record is provided on the left hand side of the corresponding tab, allowing the user to select the entry of

interest.

Biospecimen Search

A search filter just below the biospecimen section can be used to find and filter biospecimen data. The wildcard search will highlight entities in the tree that match the characters typed. This will search both the case submitter ID, as well as the additional metadata for each entity. For example, searching 'Primary Tumor' will highlight samples that match that type.

Biospecimen			⬇ Export
🔍 primary tumor	Collapse All		

Tree view:
- ⊟ Samples
  - ⚗ TCGA-AC-A23H-11A
    - ⊟ Portions
      - ⚗ TCGA-AC-A23H-11A-12
        - ⊟ Analytes
          - ⚗ TCGA-AC-A23H-11A-12D
            - ⊞ Aliquots
          - ⚗ TCGA-AC-A23H-11A-12R
            - ⊞ Aliquots
          - ⚗ TCGA-AC-A23H-11A-12W
            - ⊞ Aliquots
        - ⊞ Slides
  - ⚗ TCGA-AC-A23H-01A ▸
    - ⊟ Portions
      - ⚗ TCGA-AC-A23H-01A-11
        - ⊞ Analytes
        - ⊟ Slides
          - ⚗ TCGA-AC-A23H-01A-01-TS1

Submitter ID	TCGA-AC-A23H-01A
Sample ID	d7e3b628-d5fd-4e79-9c4a-6409330fb8a7
Sample Type	Primary Tumor
Sample Type Id	01
Tissue Type	--
Tumor Code	--
Tumor Code Id	--
Oct Embedded	false
Shortest Dimension	--
Intermediate Dimension	--
Longest Dimension	--
Is Ffpe	false
Pathology Report Uuid	A7C7D409-D086-4A9B-8C8F-E7E231D5891D
Tumor Descriptor	--
Current Weight	--
Initial Weight	70
Composition	--
Time Between Clamping And Freezing	--
Time Between Excision And Freezing	--
Days To Sample Procurement	--
Freezing Method	--
Preservation Method	--
Days To Collection	478
Portions	1
Status	4

Gene Summary Page

The Gene Summary Page describes each gene with mutation data featured at the GDC and provides results related to the analyses that are performed on these genes.

Summary

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

TP53

Summary

Symbol	TP53
Name	tumor protein p53
Synonyms	LFS1
	p53
Type	protein_coding
Location	chr17:7661779-7687550 (GRCh38)
Strand	–
Description	This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycl...

▾ *more*

External References

Entrez Gene	☐ 7157
Uniprotkb Swissprot	☐ P04637
Hgnc	☐ HGNC:11998
Omim Gene	☐ 191170

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

External References

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

Cancer Distribution

📊 Cancer Distribution

21 CASES AFFECTED BY 21 MUTATIONS ACROSS 21 PROJECTS
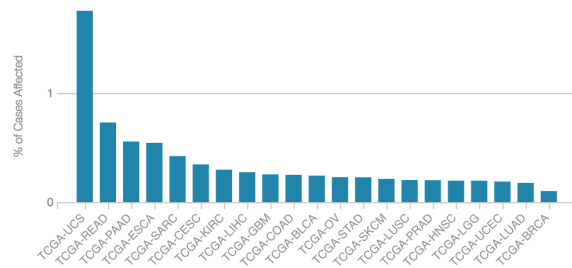
Project ID	Disease Type	Site	# Affected Cases	# Mutations
TCGA-THYM	Thymoma	Thymus	1 / 123 (0.81%)	1
TCGA-READ	Rectum Adenocarcinoma	Colorectal	1 / 137 (0.73%)	1
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	1 / 180 (0.56%)	1
TCGA-ESCA	Esophageal Carcinoma	Esophagus	1 / 184 (0.54%)	1
TCGA-SARC	Sarcoma	Soft Tissue	1 / 237 (0.42%)	1
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	1 / 364 (0.27%)	1
TCGA-GBM	Glioblastoma Multiforme	Brain	1 / 393 (0.25%)	1
TCGA-COAD	Colon Adenocarcinoma	Colorectal	1 / 400 (0.25%)	1
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	1 / 412 (0.24%)	1
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary	1 / 436 (0.23%)	1

A table and bar graph displayed that shows how many cases are affected by mutations within the gene as a ratio and percentage. Each row/bar represents the number of cases for each project.
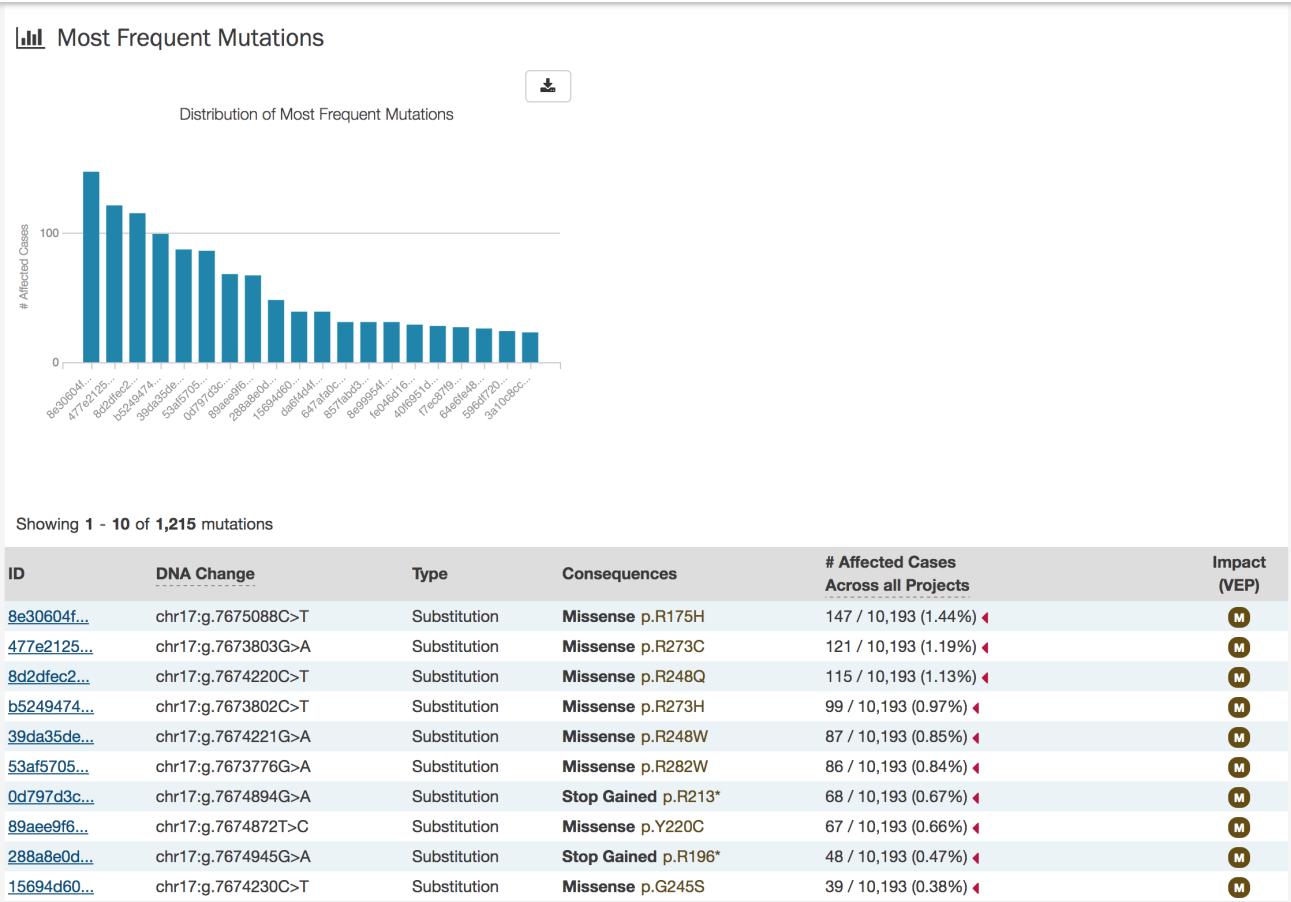
Protein Viewer
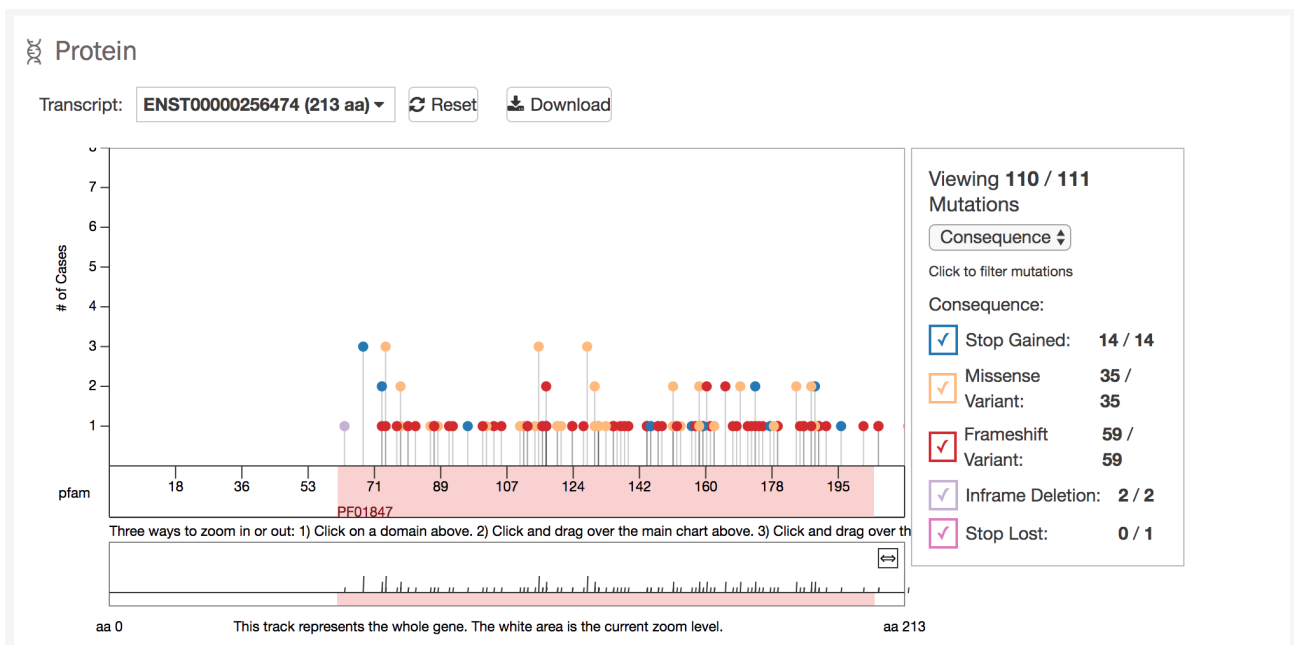
Mutations and their frequency across cases are mapped to a graphical visualization of protein-coding regions with a lollipop plot. Pfam domains are highlighted along the x-axis to assign functionality to specific protein-coding regions. The bottom track represents a view of the full gene length. Different transcripts can be selected by using the drop-down menu above the plot.

The panel to the right of the plot allows the plot to be filtered by mutation consequences or impact. The plot will dynamically change as filters are applied. Mutation consequence and impact is denoted in the plot by color.

The plot can be viewed at different zoom levels by clicking and dragging across the x-axis, clicking and dragging across the bottom track, or double clicking the pfam domain IDs. The `Reset` button can be used to bring the zoom level back to its original position. The plot can also be exported as a PNG image, SVG image or as JSON formatted text by choosing the `Download` button above the plot.

Most Frequent Mutations

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

Showing **1** - **10** of **1,215** mutations

ID	DNA Change	Type	Consequences	# Affected Cases Across all Projects	Impact (VEP)
8e30604f...	chr17:g.7675088C>T	Substitution	**Missense** p.R175H	147 / 10,193 (1.44%) ◀	M
477e2125...	chr17:g.7673803G>A	Substitution	**Missense** p.R273C	121 / 10,193 (1.19%) ◀	M
8d2dfec2...	chr17:g.7674220C>T	Substitution	**Missense** p.R248Q	115 / 10,193 (1.13%) ◀	M
b5249474...	chr17:g.7673802C>T	Substitution	**Missense** p.R273H	99 / 10,193 (0.97%) ◀	M
39da35de...	chr17:g.7674221G>A	Substitution	**Missense** p.R248W	87 / 10,193 (0.85%) ◀	M
53af5705...	chr17:g.7673776G>A	Substitution	**Missense** p.R282W	86 / 10,193 (0.84%) ◀	M
0d797d3c...	chr17:g.7674894G>A	Substitution	**Stop Gained** p.R213*	68 / 10,193 (0.67%) ◀	M
89aee9f6...	chr17:g.7674872T>C	Substitution	**Missense** p.Y220C	67 / 10,193 (0.66%) ◀	M
288a8e0d...	chr17:g.7674945G>A	Substitution	**Stop Gained** p.R196*	48 / 10,193 (0.47%) ◀	M
15694d60...	chr17:g.7674230C>T	Substitution	**Missense** p.G245S	39 / 10,193 (0.38%) ◀	M

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

- **ID:** A UUID Code for the mutation assigned by the GDC, when clicked will bring a user to the Mutation Summary Page
- **DNA Change:** The chromosome and starting coordinates of the mutation are displayed along with the nucleotide differences between the reference and tumor allele
- **Type:** A general classification of the mutation
- **Consequences:** The effects the mutation has on the gene coding for a protein (i.e. synonymous, missense, non-coding transcript)
- **# Affected Cases:** The number of affected cases, expressed as number across all projects. Choosing the arrow next to the percentage will expand the selection with a breakdown of each affected project
- **Impact:** A subjective classification of the severity of the variant consequence. The categories are:
  - **HIGH**: 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**: A non-disruptive variant that might change protein effectiveness
  - **LOW**: Assumed to be mostly harmless or unlikely to change protein behavior

Mutation Summary Page

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

Summary

MU chr3:g.10142050C>A

⊞ Summary			🗐 External References	
ID	966829c5-20d6-52f7-ae12-3c69810cc61b		Entrez Gene	⊡ 7428
DNA Change	chr3:g.10142050C>A		Uniprotkb Swissprot	⊡ P40337
Type	SNP		Hgnc	⊡ HGNC:12687
Reference Genome Assembly	GRCh38		Omim Gene	⊡ 608537
Allele In The Reference Assembly	C			
Functonal Impact	High			

- **ID:** A unique identifier (UUID) for this mutation
- **DNA Change:** Denotes the chromosome number, position, and nucleotide change of the mutation
- **Type:** A broad categorization of the mutation
- **Reference Genome Assembly:** The reference genome in which the chromosomal position refers to
- **Allele in the Reference Assembly:** The nucleotide(s) that compose the site in the reference assembly
- **Functional Impact:** A subjective classification of the severity of the variant consequence. The categories are:
  - **HIGH**: 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**: A non-disruptive variant that might change protein effectiveness
  - **LOW**: Assumed to be mostly harmless or unlikely to change protein behavior

External References

A separate panel contains links to information about the gene that the mutation effects in external databases. These databases include: Entrez, Hugo Gene Nomenclature Committee, Online Mendelian Inheritance in Man, and Uniprotkb SwissProt.

Consequences

⊞ Consequences					
Gene	**AA Change**	**Consequence**	**Coding DNA Change**	**Strand**	**Transcript(s)**
VHL	p.Ser68Ter	stop_gained	c.203C>A	**+**	⊡ ENST00000345392 ⊡ **ENST00000256474** ⊡ ENST00000477538

The consequences of the mutation are displayed in a table. The fields that detail each mutation are listed below:

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

Cancer Distribution

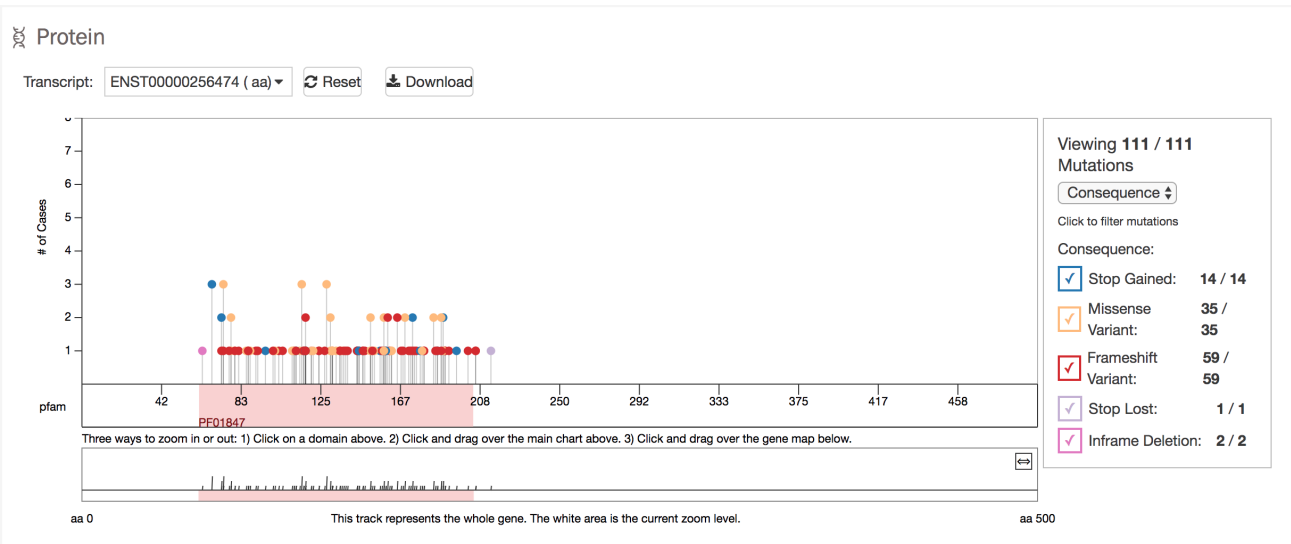|			
📊 Cancer Distribution			
THIS MUTATION AFFECTS 3 DISTINCT CASES ACROSS 1 CANCER PROJECTS			
Project ID	**Disease Type**	**Site**	**# Affected Cases**
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	3 / 336 (0.89%)

A bar graph detailing the number of cases affected by the mutation across all projects is generated dynamically. Below the graph is a table with information about how the mutation affects each project, which can be exported as a JSON object. The table contains the following fields:

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

Protein Viewer

The protein viewer displays a plot representing the position of mutations along the polypeptide chain associated with the mutation. The y-axis represents the number of cases that exhibit each mutation, whereas the x-axis represents the polypeptide chain sequence. Pfam domains that were identified along the polypeptide chain are identified with colored rectangles labeled with pfam IDs. See the Gene Summary Page for additional details about the protein viewer.

Visualization API Endpoint Additions

The GDC Visualization Suite uses the same API as the rest of the Data Portal and takes advantage of several new endpoints:

- **ssms:** The simple somatic mutation (`ssms`) endpoint allows users to access information about each somatic point mutation. For example, a `ssm` would represent the transition of C to T at position 52000 of chromosome 1.
- **ssm_occurrences:** A SSM entity as applied to a single instance (case). An example of a `ssm occurrence` would be that the transition of C to T at position 52000 of chromosome 1 occurred in patient TCGA-XX-XXXX.
- **genes:** The `genes` endpoint allows users to access in-depth information about each gene.

The methods for retrieving information from these endpoint are very similar to those used for the `cases` and `files` endpoints. These methods are explored in depth in the [API Search and Retrieval](#) documentation. The `_mapping` parameter can also be used with each of these endpoints to generate a list of potential fields. For example:

```
https://api.gdc.cancer.gov/ssms/_mapping
```

While it is not an endpoint, the `observation` entity is featured in the visualization section of the API. The `observation` entity provides information from the MAF file, such as read depth and normal genotype, that supports the validity of the associated `ssm`.

Analysis Endpoints

In addition the `ssms`, `ssm_occurrences`, and `genes` endpoints mentioned previously, several `analysis` endpoints were designed to quickly retrieve specific datasets used for visualization display.

- **analysis/survival**
- **analysis/top_cases_counts_by_genes**
- **analysis/top_mutated_genes_by_project**
- **analysis/top_mutated_cases_by_gene**
- **analysis/top_mutated_cases_by_ssm**
- **analysis/mutated_cases_count_by_project**

For a set of API example queries, go to:

https://github.com/NCIP/gdc-docs/blob/Visualization-Docs/docs/Visualization_Documentation/API_Endpoints.md

GDC MAF Format v.1.0.0

Introduction

Mutation Annotation Format (MAF) is a tab-delimited text file with aggregated mutation information from VCF Files and are generated on a project-level. MAF files are produced through the [Somatic Aggregation Workflow](#) The GDC produces MAF files at two permission levels: **protected** and **somatic** (or open-access). One MAF file is produced per variant calling pipeline per GDC project. MAFs are produced by aggregating the GDC annotated VCF files generated from one pipeline for one project.

Annotated VCF files often have variants reported on multiple transcripts whereas the MAF files generated from the VCFs (*protected.maf) only report the most critically affected one. Somatic MAFs (*somatic.maf), which are also known as [Masked Somatic Mutation](#) files, are further processed to remove lower quality and potential germline variants. For tumor samples that contain variants from multiple combinations of tumor-normal aliquot pairs, only one pair is selected in the Somatic MAF based on their sample type. Somatic MAFs are publicly available and can be freely distributed within the boundaries of the [GDC Data Access Policies](#).

The GDC MAF file format is based on the [TCGA Mutation Annotation Format](#) specifications, with additional columns included.

Note: The criteria for allowing mutations into open-access are purposefully implemented to overcompensate and filter out germline variants. If omission of true-positive somatic mutations is a concern, the GDC recommends using protected MAFs.

Somatic MAF File Generation

The process for modifying a protected MAF into a somatic MAF is as follows:

- Aliquot Selection: only one tumor-normal pair are selected for each tumor sample based on the plate number, sample type, analyte type and other features extracted from tumor TCGA aliquot barcode.
- Low quality variant filtering and germline masking:
    i. Variants with **Mutation_Status != 'Somatic'** or **GDC_FILTER = 'Gapfiller', 'ContEst', 'multiallelic', 'nonselectedaliquot', 'BCR_Duplicate' or 'BadSeq'** are **removed**.
    ii. Remaining variants with **GDC_Valid_Somatic = True** are **included** in the Somatic MAF.
    iii. Remaining variants with **FILTER != 'panel_of_normals' or PASS** are **removed**. Note that the `FILTER != panel_of_normals` value is only relevant for the variants generated from the MuTect2 pipeline.
    iv. Remaining variants with **MC3_Overlap = True** are **included** in the Somatic MAF.
    v. Remaining variants with **GDC_FILTER = 'ndp', 'NonExonic', 'bitgt', or 'gdc_pon'** are **removed**.
    vi. Remaining variants with **SOMATIC != null** are **included** in the Somatic MAF.
    vii. Remaining variants with **dbSNP_RS = 'novel' or null** are **included** in the Somatic MAF.
    viii. Remaining variants are **removed**.
- Removal of the following columns:
    - vcf_info
    - vcf_format
    - vcf_tumor_gt
    - vcf_normal_gt
    - GDC_Valid_Somatic
- Set values to be blank in the following columns that may contain information about germline genotypes:
    - Match_Norm_Seq_Allele1
    - Match_Norm_Seq_Allele2
    - Match_Norm_Validation_Allele1, Match_Norm_Validation_Allele2
    - n_ref_count, n_alt_count
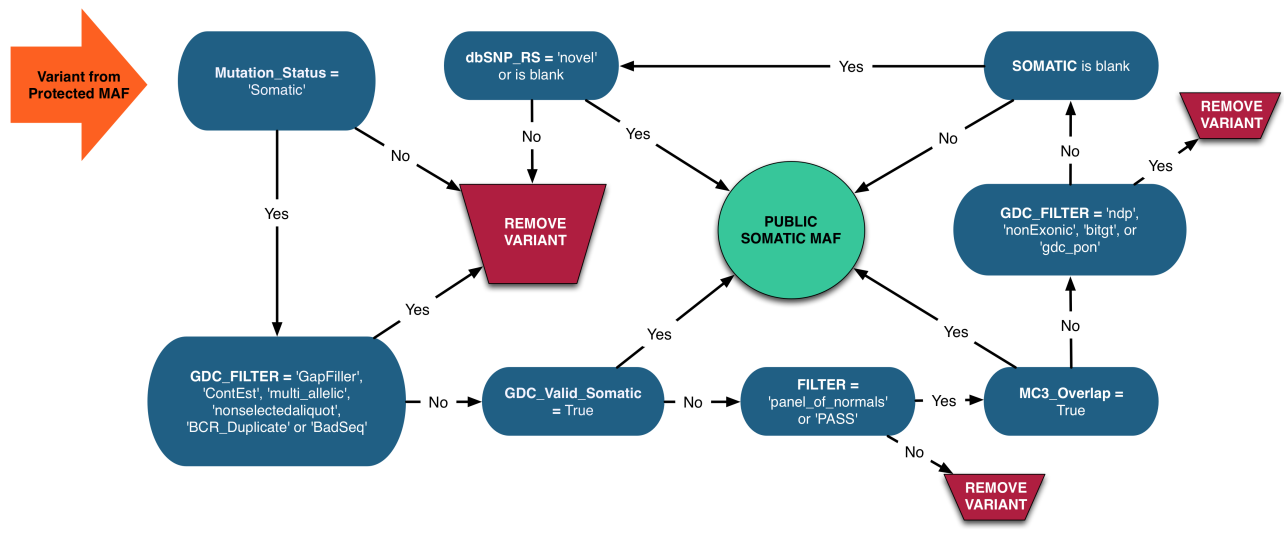
Protected MAF File Structure

The table below describes the columns in a protected MAF and their definitions. Note that the somatic (open-access) MAF structure is the same except for having the last six columns removed.

Column	Description
1 - Hugo_Symbol	HUGO symbol for the gene (HUGO symbols are always in all caps). "Unknown" is used for regions that do not correspond to a gene
2 - Entrez_Gene_Id	Entrez gene ID (an integer). "0" is used for regions that do not correspond to a gene region or Ensembl ID
3 - Center	One or more genome sequencing center reporting the variant
4 - NCBI_Build	The reference genome used for the alignment (GRCh38)
5 - Chromosome	The affected chromosome (chr1)
6 - Start_Position	Lowest numeric position of the reported variant on the genomic reference sequence. Mutation start coordinate
7 - End_Position	Highest numeric genomic position of the reported variant on the genomic reference sequence. Mutation end coordinate
8 - Strand	Genomic strand of the reported allele. Currently, all variants will report the positive strand: '+'
9 - Variant_Classification	Translational effect of variant allele
10 - Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
11 - Reference_Allele	The plus strand reference allele at this position. Includes the deleted sequence for a deletion or "-" for an insertion
12 - Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases
13 - Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
14 - dbSNP_RS	The rs-IDs from the dbSNP database, "novel" if not found in any database used, or null if there is no dbSNP record, but it is found in other databases
15 - dbSNP_Val_Status	The dbSNP validation status is reported as a semicolon-separated list of statuses. The union of all rs-IDs is taken when there are multiple
16 - Tumor_Sample_Barcode	Aliquot barcode for the tumor sample
17 - Matched_Norm_Sample_Barcode	Aliquot barcode for the matched normal sample
18 - Match_Norm_Seq_Allele1	Primary data genotype. Matched normal sequencing allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases (cleared in somatic MAF)
19 - Match_Norm_Seq_Allele2	Matched normal sequencing allele 2
20 - Tumor_Validation_Allele1	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases

Column	Description
21 - Tumor_Validation_Allele2	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 2
22 - Match_Norm_Validation_Allele1	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases (cleared in somatic MAF)
23 - Match_Norm_Validation_Allele2	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 2 (cleared in somatic MAF)
24 - Verification_Status	Second pass results from independent attempt using same methods as primary data source. Generally reserved for 3730 Sanger Sequencing
25 - Validation_Status	Second pass results from orthogonal technology
26 - Mutation_Status	An assessment of the mutation as somatic, germline, LOH, post transcriptional modification, unknown, or none. The values allowed in this field are constrained by the value in the Validation_Status field
27 - Sequencing_Phase	TCGA sequencing phase (if applicable). Phase should change under any circumstance that the targets under consideration change
28 - Sequence_Source	Molecular assay type used to produce the analytes used for sequencing. Allowed values are a subset of the SRA 1.5 library_strategy field values. This subset matches those used at CGHub
29 - Validation_Method	The assay platforms used for the validation call
30 - Score	Not in use
31 - BAM_File	Not in use
32 - Sequencer	Instrument used to produce primary sequence data
33 - Tumor_Sample_UUID	GDC aliquot UUID for tumor sample
34 - Matched_Norm_Sample_UUID	GDC aliquot UUID for matched normal sample
35 - HGVSc	The coding sequence of the variant in HGVS recommended format
36 - HGVSp	The protein sequence of the variant in HGVS recommended format. "p.=" signifies no change in the protein
37 - HGVSp_Short	Same as the HGVSp column, but using 1-letter amino-acid codes
38 - Transcript_ID	Ensembl ID of the transcript affected by the variant
39 - Exon_Number	The exon number (out of total number)
40 - t_depth	Read depth across this locus in tumor BAM
41 - t_ref_count	Read depth supporting the reference allele in tumor BAM
42 - t_alt_count	Read depth supporting the variant allele in tumor BAM
43 - n_depth	Read depth across this locus in normal BAM
44 - n_ref_count	Read depth supporting the reference allele in normal BAM (cleared in somatic MAF)
45 - n_alt_count	Read depth supporting the variant allele in normal BAM (cleared in somatic MAF)

Column	Description
46 - all_effects	A semicolon delimited list of all possible variant effects, sorted by priority ([SYMBOL,Consequence,HGVSp_Short,Transcript_ID,RefSeq])
47 - Allele	The variant allele used to calculate the consequence
48 - Gene	Stable Ensembl ID of affected gene
49 - Feature	Stable Ensembl ID of feature (transcript, regulatory, motif)
50 - Feature_type	Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature (or blank)
51 - Consequence	Consequence type of this variation; sequence ontology terms
52 - cDNA_position	Relative position of base pair in the cDNA sequence as a fraction. A "-" symbol is displayed as the numerator if the variant does not appear in cDNA
53 - CDS_position	Relative position of base pair in coding sequence. A "-" symbol is displayed as the numerator if the variant does not appear in coding sequence
54 - Protein_position	Relative position of affected amino acid in protein. A "-" symbol is displayed as the numerator if the variant does not appear in coding sequence
55 - Amino_acids	Only given if the variation affects the protein-coding sequence
56 - Codons	The alternative codons with the variant base in upper case
57 - Existing_variation	Known identifier of existing variation
58 - ALLELE_NUM	Allele number from input; 0 is reference, 1 is first alternate etc.
59 - DISTANCE	Shortest distance from the variant to transcript
60 - TRANSCRIPT_STRAND	The DNA strand (1 or -1) on which the transcript/feature lies
61 - SYMBOL	The gene symbol
62 - SYMBOL_SOURCE	The source of the gene symbol
63 - HGNC_ID	Gene identifier from the HUGO Gene Nomenclature Committee if applicable
64 - BIOTYPE	Biotype of transcript
65 - CANONICAL	A flag (YES) indicating that the VEP-based canonical transcript, the longest translation, was used for this gene. If not, the value is null
66 - CCDS	The CCDS identifier for this transcript, where applicable
67 - ENSP	The Ensembl protein identifier of the affected transcript
68 - SWISSPROT	UniProtKB/Swiss-Prot accession
69 - TREMBL	UniProtKB/TrEMBL identifier of protein product
70 - UNIPARC	UniParc identifier of protein product
71 - RefSeq	RefSeq identifier for this transcript
72 - SIFT	The SIFT prediction and/or score, with both given as prediction (score)
73 - PolyPhen	The PolyPhen prediction and/or score
74 - EXON	The exon number (out of total number)
75 - INTRON	The intron number (out of total number)
76 - DOMAINS	The source and identifier of any overlapping protein domains

Column	Description
77 - GMAF	Non-reference allele and frequency of existing variant in 1000 Genomes
78 - AFR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined African population
79 - AMR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined American population
80 - ASN_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined Asian population
81 - EAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined East Asian population
82 - EUR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined European population
83 - SAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined South Asian population
84 - AA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP African American population
85 - EA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP European American population
86 - CLIN_SIG	Clinical significance of variant from dbSNP
87 - SOMATIC	Somatic status of each ID reported under Existing_variation (0, 1, or null)
88 - PUBMED	Pubmed ID(s) of publications that cite existing variant
89 - MOTIF_NAME	The source and identifier of a transcription factor binding profile aligned at this position
90 - MOTIF_POS	The relative position of the variation in the aligned TFBP
91 - HIGH_INF_POS	A flag indicating if the variant falls in a high information position of a transcription factor binding profile (TFBP) (Y, N, or null)
92 - MOTIF_SCORE_CHANGE	The difference in motif score of the reference and variant sequences for the TFBP
93 - IMPACT	The impact modifier for the consequence type
94 - PICK	Indicates if this block of consequence data was picked by VEP's pick feature (1 or null)
95 - VARIANT_CLASS	Sequence Ontology variant class
96 - TSL	Transcript support level, which is based on independent RNA analyses
97 - HGVS_OFFSET	Indicates by how many bases the HGVS notations for this variant have been shifted
98 - PHENO	Indicates if existing variant is associated with a phenotype, disease or trait (0, 1, or null)
99 - MINIMISED	Alleles in this variant have been converted to minimal representation before consequence calculation (1 or null)
100 - ExAC_AF	Global Allele Frequency from ExAC
101 - ExAC_AF_Adj	Adjusted Global Allele Frequency from ExAC

Column	Description
102 - ExAC_AF_AFR	African/African American Allele Frequency from ExAC
103 - ExAC_AF_AMR	American Allele Frequency from ExAC
104 - ExAC_AF_EAS	East Asian Allele Frequency from ExAC
105 - ExAC_AF_FIN	Finnish Allele Frequency from ExAC
106 - ExAC_AF_NFE	Non-Finnish European Allele Frequency from ExAC
107 - ExAC_AF_OTH	Other Allele Frequency from ExAC
108 - ExAC_AF_SAS	South Asian Allele Frequency from ExAC
109 - GENE_PHENO	Indicates if gene that the variant maps to is associated with a phenotype, disease or trait (0, 1, or null)
110 - FILTER	Copied from input VCF
111 - CONTEXT	The reference allele per VCF specs, and its five flanking base pairs
112 - src_vcf_id	GDC UUID for the input VCF file
113 - tumor_bam_uuid	GDC UUID for the tumor bam file
114 - normal_bam_uuid	GDC UUID for the normal bam file
115 - case_id	GDC UUID for the case
116 - GDC_FILTER	GDC filters applied universally across all MAFs
117 - COSMIC	Overlapping COSMIC variants
118 - MC3_Overlap	Indicates whether this region overlaps with an MC3 variant for the same sample pair
119 - GDC_Validation_Status	GDC implementation of validation checks. See notes section (#5) below for details
120 - GDC_Valid_Somatic	True or False (not in somatic MAF)
121 - vcf_region	Colon separated string containing the CHROM, POS, ID, REF, and ALT columns from the VCF file (e.g., chrZ:20:rs1234:A:T) (not in somatic MAF)
122 - vcf_info	INFO column from VCF (not in somatic MAF)
123 - vcf_format	FORMAT column from VCF (not in somatic MAF)
124 - vcf_tumor_gt	Tumor sample genotype column from VCF (not in somatic MAF)
125 - vcf_normal_gt	Normal sample genotype column from VCF (not in somatic MAF)

Notes About GDC MAF Implementation

1. Column #4 **NCBI_Build** is GRCh38 by default
2. Column #32 **Sequencer** includes the sequencers used. If different sequencers were used to generate normal and tumor data, the normal sequencer is listed first.
3. Column #60 VEP name "STRAND" is changed to **TRANSCRIPT_STRAND** to avoid confusion with Column#8 "Strand"
4. Column #122-125 **vcf_info, vcf_format, vcf_tumor_gt, and vcf_normal_gt** are the corresponding columns from the VCF files. Including them facilitates parsing specific variant information.
5. Column #119 **GDC_Validation_Status**: GDC also collects TCGA validation sequences. It compares these with variants derived from Next-Generation Sequencing data from the same sample and populates the comparison result in "GDC_Validation_Status".

- "Valid", if the alternative allele(s) in the tumor validation sequence is(are) the same as GDC variant call
- "Invalid", if none of the alternative allele(s) in the tumor validation sequence is the same as GDC variant call
- "Inconclusive" if two alternative allele exists, and one matches while the other does not
- "Unknown" if no validation sequence exists

6. Column #120 **GDC_Valid_Somatic** is TRUE if GDC_Validation_Status is "Valid" and the variant is "Somatic" in validation calls. It is FALSE if these criteria are not met