## Breakout Session 1: Track B

# NCI CRDC Cloud Transfer of TP53 Website and Database

Mr. William Longabaugh
*Senior Software Engineer, Institute for Systems Biology*

# NCI CRDC Cloud Transfer of *TP53* Website and Database

**William Longabaugh**

Senior Software Engineer, Institute for Systems Biology

Jan 17 2024

# Funding

- We received funds from *“FY2021 Request for ODSS Funds to Catalyze Migration to and Usage of the Cloud via the STRIDES Initiative (HVD 21)”*
- Google cloud credits were provided to us to support cloud operations underlying our migration of the IARC WHO TP53 database (now retired) to become part of the ISB-CGC Cloud Resource, a component of the Cancer Research Data Commons (CRDC)
- Additionally, the credits covered cloud operation costs of our development, test, and production tier Google cloud projects until September 2023

Thank you to the Office of Data Science Strategy

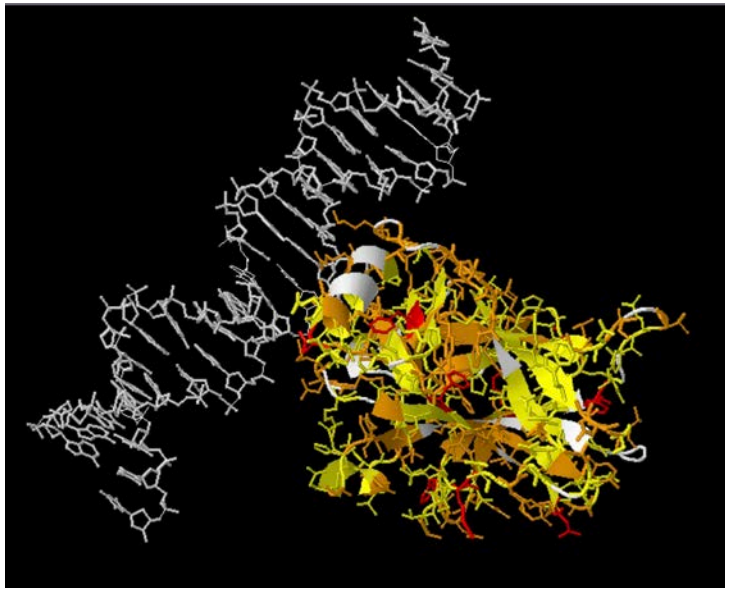
# The *TP53* Database: Aim and Scope

Database compiles *TP53* variant data from 1989

Currently holds information on 24,547 *TP53* variants

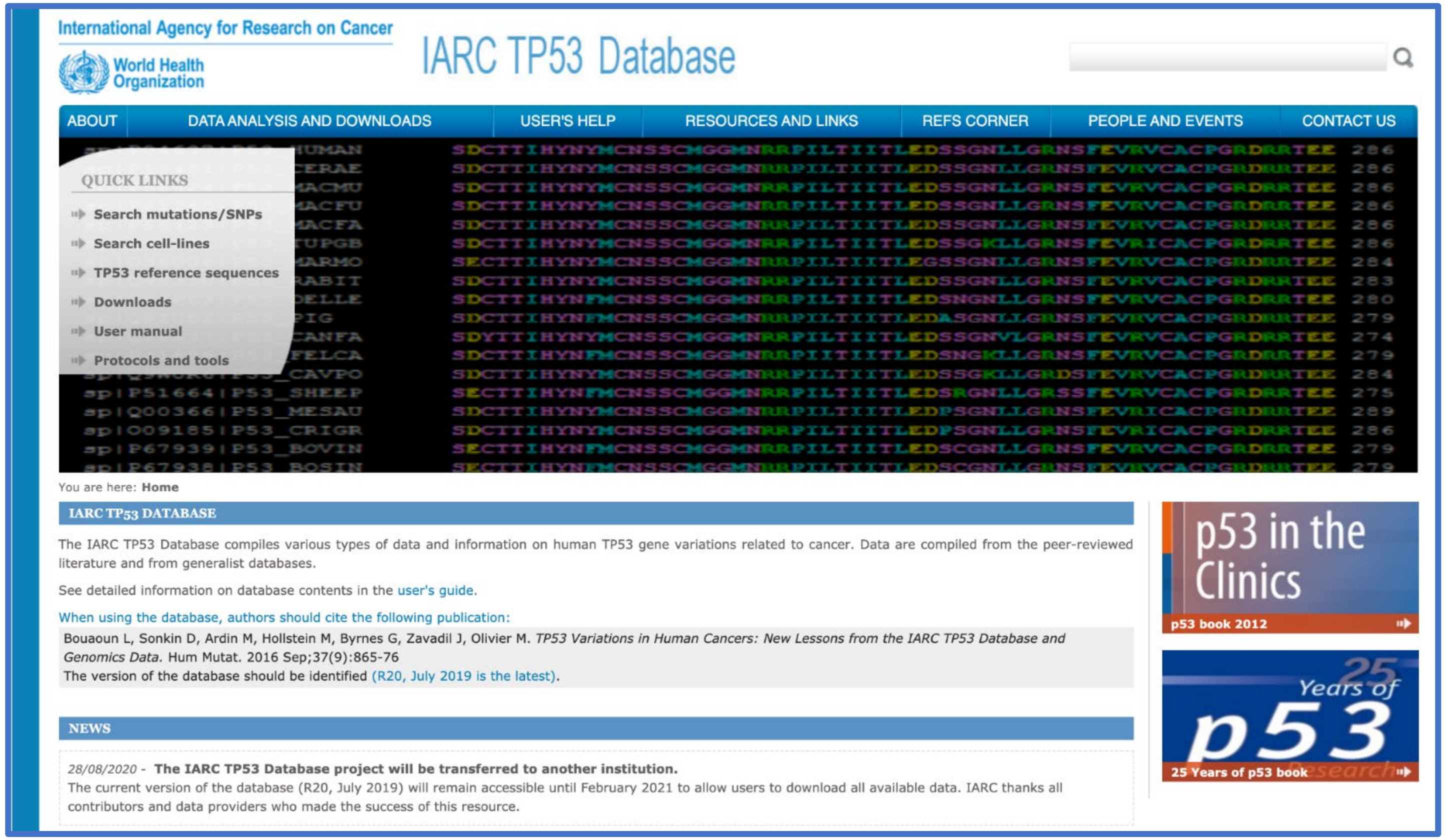
Database includes:

- *TP53* **functional** and **structural** data
- *TP53* **tumor** variants in sporadic cancer
- *TP53* **germline** variants in cancer patients, families with cancers
- *TP53* gene status in human **cell-lines**
- **Mouse models** with engineered *p53*
- **Experimentally-induced** *TP53* variants



Holds information on *TP53* variants for a broad range of scientists and clinicians who work in different research areas

# IARC *TP53* Database



You are here: **Home**

**IARC TP53 DATABASE**

The IARC TP53 Database compiles various types of data and information on human TP53 gene variations related to cancer. Data are compiled from the peer-reviewed literature and from generalist databases.

See detailed information on database contents in the user's guide.

When using the database, authors should cite the following publication:

Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, Olivier M. *TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data*. Hum Mutat. 2016 Sep;37(9):865-76
The version of the database should be identified (R20, July 2019 is the latest).

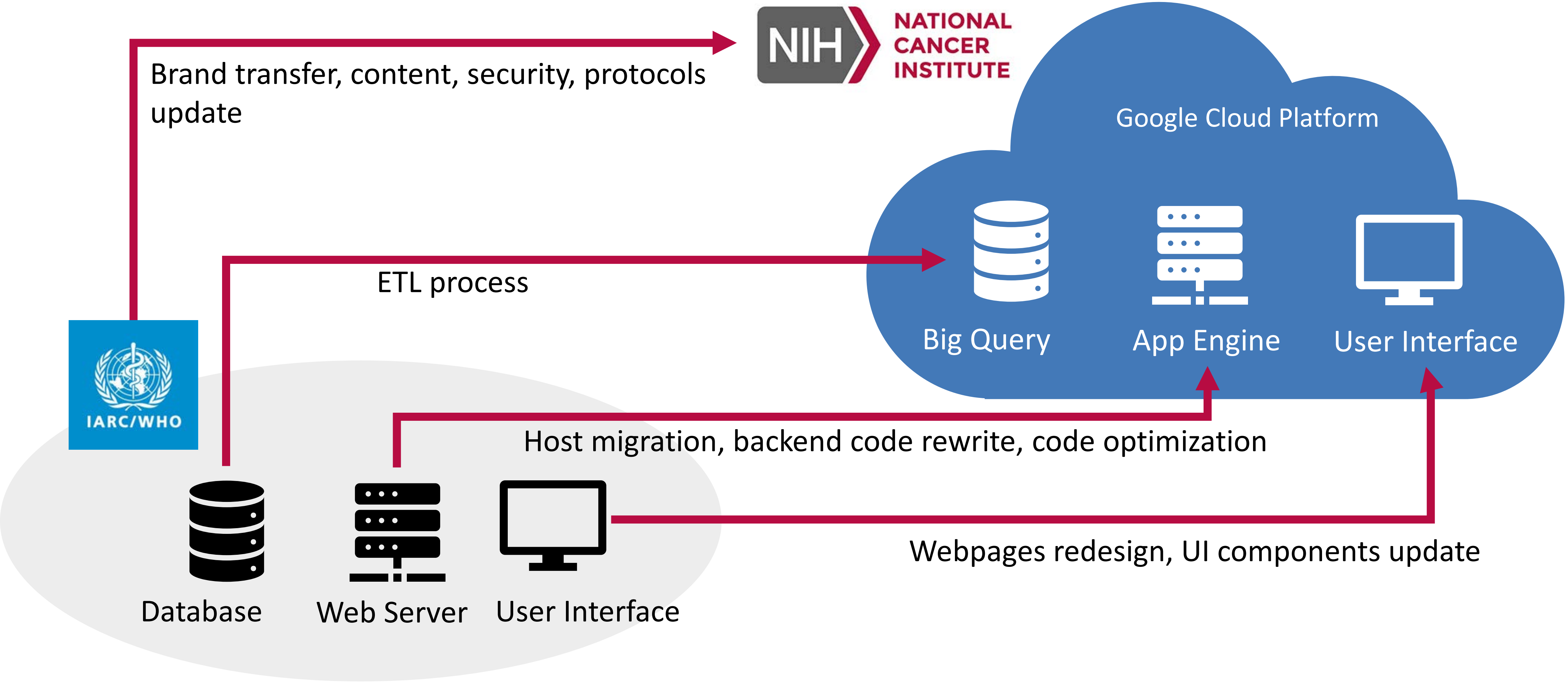
**NEWS**

*28/08/2020* - **The IARC TP53 Database project will be transferred to another institution.**
The current version of the database (R20, July 2019) will remain accessible until February 2021 to allow users to download all available data. IARC thanks all contributors and data providers who made the success of this resource.

IARC TP53 Database Website in 2020

The original ***TP53* database** was initiated in 1991, further developed and maintained by WHO's **International Agency for Research on Cancer** until 2021.

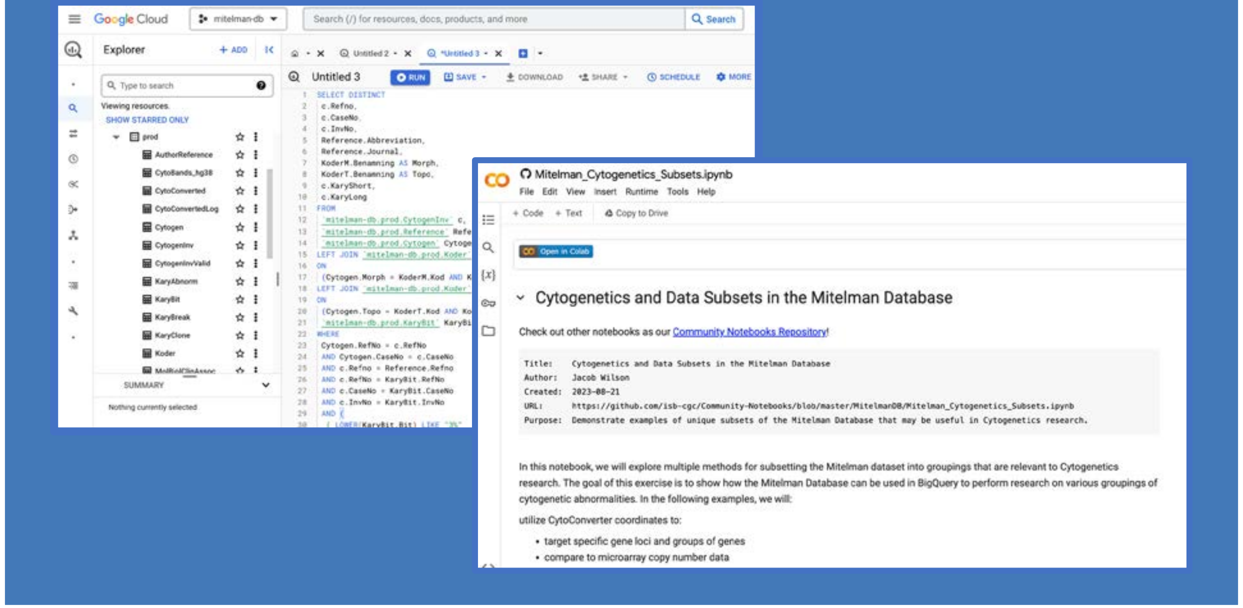
# Transfer of Website and Database into the Cloud


Brand transfer, content, security, protocols update
NIH NATIONAL CANCER INSTITUTE
Google Cloud Platform
ETL process
Big Query
App Engine
User Interface
IARC/WHO
Host migration, backend code rewrite, code optimization
Webpages redesign, UI components update
Database
Web Server
User Interface

# Transfer of Website and Database into the Clouds: **Mitelman Database**

- The **Mitelman Database** was part of CGAP (Cancer Genome Anatomy Project, NCI)
- That website was retired on 2019
- ISB-CGC was responsible for transferring all web components to the Google Cloud Platform
- The application has been further developed for advanced queries and additional features

All data is **publicly available in BigQuery**.



```
SELECT DISTINCT
 c.Refno,
 c.CaseNo,
 c.InvNo,
 Reference.Abbreviation,
 Reference.Journal,
 KoderM.Benamning AS Morph,
 KoderT.Benamning AS Topo,
 c.KaryShort,
 c.KaryLong
FROM
 `mitelman-db.prod.CytogenInv` c,
 `mitelman-db.prod.Reference` Refe
 `mitelman-db.prod.Cytogen` Cytoge
LEFT JOIN `mitelman-db.prod.Koder`
ON
 (Cytogen.Morph = KoderM.Kod AND K
LEFT JOIN `mitelman-db.prod.Koder`
ON
 (Cytogen.Topo = KoderT.Kod AND Ko
 `mitelman-db.prod.KaryBit` KaryBi
WHERE
 Cytogen.RefNo = c.RefNo
 AND Cytogen.CaseNo = c.CaseNo
 AND c.Refno = Reference.Refno
 AND c.RefNo = KaryBit.RefNo
 AND c.CaseNo = KaryBit.CaseNo
 AND c.InvNo = KaryBit.InvNo
 AND (
  ( LOWER(KaryBit.Bit) LIKE "%"
```

CO
Mitelman_Cytogenetics_Subsets.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Copy to Drive
Open in Colab

Cytogenetics and Data Subsets in the Mitelman Database

Check out other notebooks as our Community Notebooks Repository!

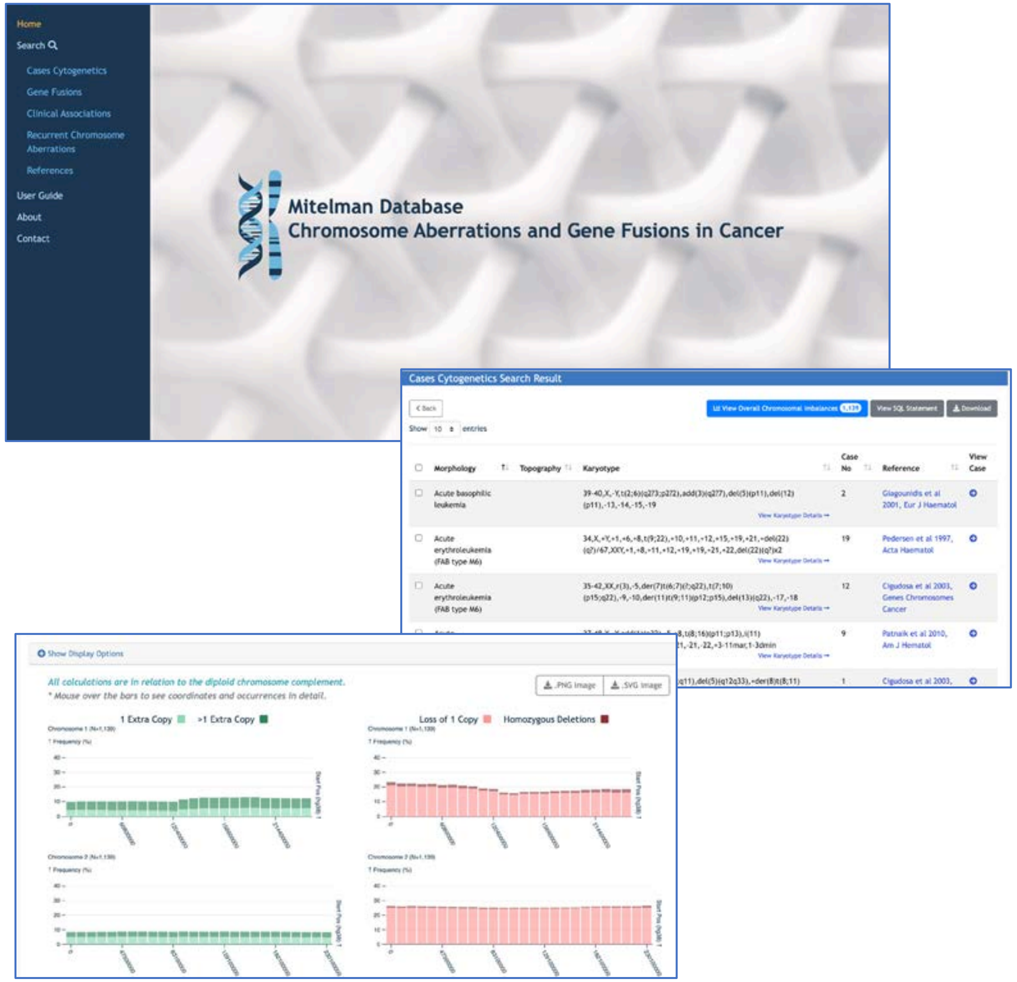
```
Title:    Cytogenetics and Data Subsets in the Mitelman Database
Author:   Jacob Wilson
Created:  2023-08-21
URL:      https://github.com/isb-cgc/Community-Notebooks/blob/master/MitelmanDB/Mitelman_Cytogenetics_Subsets.ipynb
Purpose:  Demonstrate examples of unique subsets of the Mitelman Database that may be useful in Cytogenetics research.
```

In this notebook, we will explore multiple methods for subsetting the Mitelman dataset into groupings that are relevant to Cytogenetics research. The goal of this exercise is to show how the Mitelman Database can be used in BigQuery to perform research on various groupings of cytogenetic abnormalities. In the following examples, we will:

utilize CytoConverter coordinates to:

- target specific gene loci and groups of genes
- compare to microarray copy number data

https://mitelmandatabase.isb-cgc.org



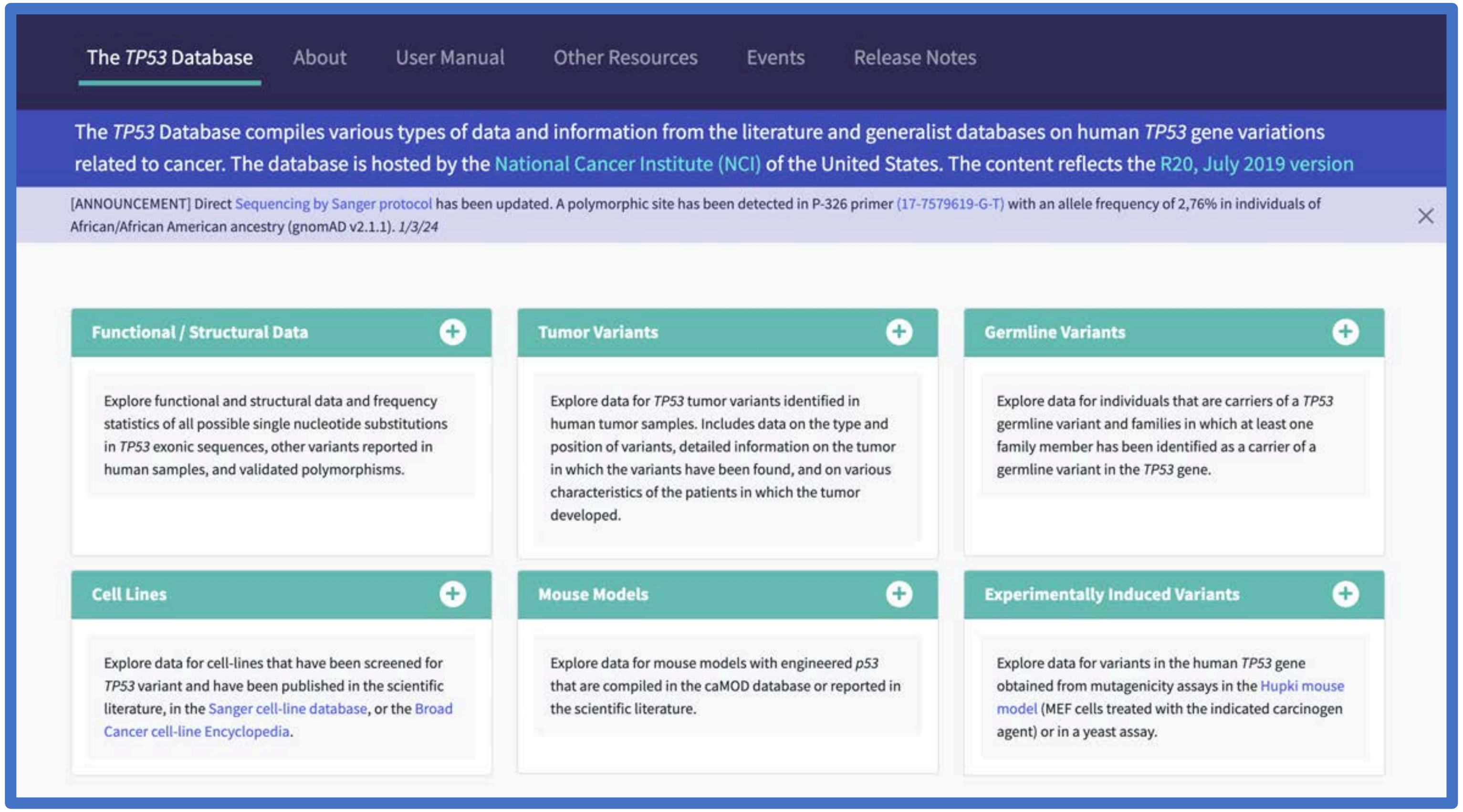
Cases Cytogenetics Search Result

| ☐ | Morphology | Topography | Karyotype | Case No | Reference | View Case |
|---|---|---|---|---|---|---|
| ☐ | Acute basophilic leukemia | | 39-40,X,-Y,t(2;6)(q27?3;p27?2),add(3)(q27?),del(5)(p11),del(12)(p11),-13,-14,-15,-19<br>View Karyotype Details → | 2 | Glagounidis et al 2001, Eur J Haematol | ● |
| ☐ | Acute erythroleukemia (FAB type M6) | | 34,X,+Y,+1,+6,+8,t(9;22),+10,+11,+12,+15,+19,+21,+del(22)(q?)/67,XXY,+1,+8,+11,+12,+19,+19,+21,+22,del(22)(q?)x2<br>View Karyotype Details → | 19 | Pedersen et al 1997, Acta Haematol | ● |
| ☐ | Acute erythroleukemia (FAB type M6) | | 35-42,XX,r(3),-5,der(7)t(6;7)(?;q22),t(7;10)(p15;q22),-9,-10,der(11)t(9;11)(p12;p15),del(13)(q22),-17,-18<br>View Karyotype Details → | 12 | Cigudosa et al 2003, Genes Chromosomes Cancer | ● |
| ☐ | Acute | | 8,t(8;16)(p11;p13),i(11)<br>21,-21,-22,+3-11mar,1-3dmin<br>View Karyotype Details → | 9 | Patnaik et al 2010, Am J Hematol | ● |
| | | | q11),del(5)(q12q33),+der(8)t(8;11) | 1 | Cigudosa et al 2003, | ● |

*All calculations are in relation to the diploid chromosome complement.*
** Mouse over the bars to see coordinates and occurrences in detail.*

# Transfer of Website and Database into the Clouds:
# The *TP53* Database

https://tp53.isb-cgc.org



The *TP53* Database compiles various types of data and information from the literature and generalist databases on human *TP53* gene variations related to cancer. The database is hosted by the National Cancer Institute (NCI) of the United States. The content reflects the R20, July 2019 version

[ANNOUNCEMENT] Direct Sequencing by Sanger protocol has been updated. A polymorphic site has been detected in P-326 primer (17-7579619-G-T) with an allele frequency of 2,76% in individuals of African/African American ancestry (gnomAD v2.1.1). *1/3/24*

**Functional / Structural Data**

Explore functional and structural data and frequency statistics of all possible single nucleotide substitutions in *TP53* exonic sequences, other variants reported in human samples, and validated polymorphisms.

**Tumor Variants**

Explore data for *TP53* tumor variants identified in human tumor samples. Includes data on the type and position of variants, detailed information on the tumor in which the variants have been found, and on various characteristics of the patients in which the tumor developed.

**Germline Variants**

Explore data for individuals that are carriers of a *TP53* germline variant and families in which at least one family member has been identified as a carrier of a germline variant in the *TP53* gene.

**Cell Lines**

Explore data for cell-lines that have been screened for *TP53* variant and have been published in the scientific literature, in the Sanger cell-line database, or the Broad Cancer cell-line Encyclopedia.

**Mouse Models**

Explore data for mouse models with engineered *p53* that are compiled in the caMOD database or reported in the scientific literature.

**Experimentally Induced Variants**

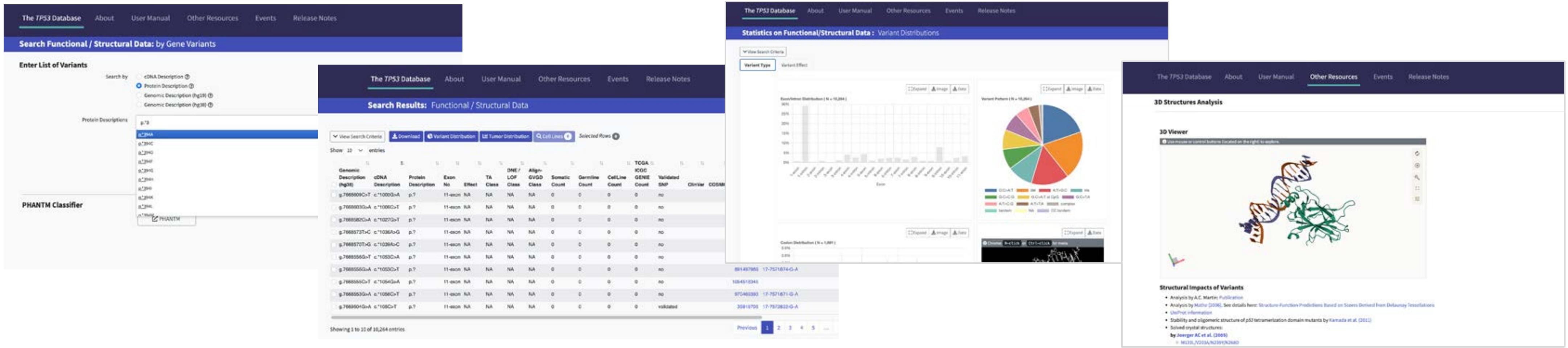
Explore data for variants in the human *TP53* gene obtained from mutagenicity assays in the Hupki mouse model (MEF cells treated with the indicated carcinogen agent) or in a yeast assay.

- The TP53 Database of NCI was launched in 2021 with all of its web components operating under **Google Cloud Platform**.
- All web queries are directly run in **BigQuery.**

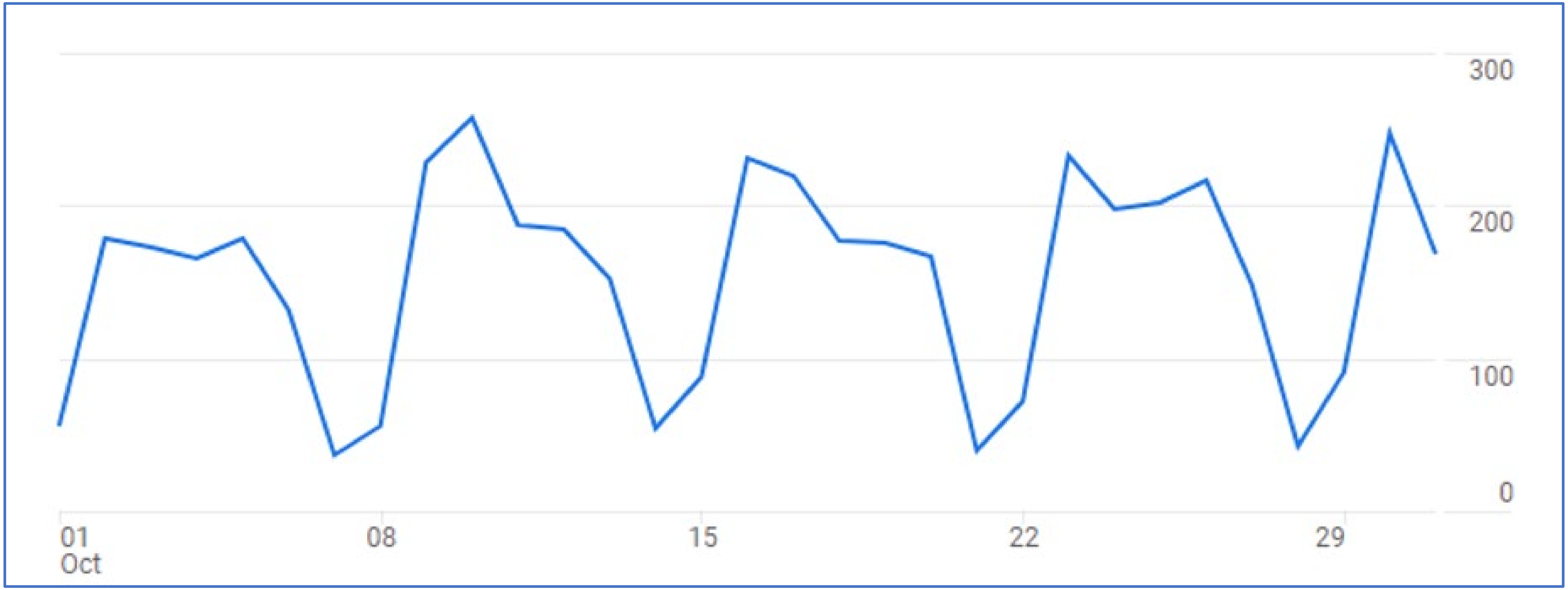
# The *TP53* Database of NCI

Application is now

- Faster to search or run analyses
- Easier to navigate
- Secure
- Shares the same development, deployment, hosting, testing, and security framework with other ISB-CGC components



# *TP53* Database Usage

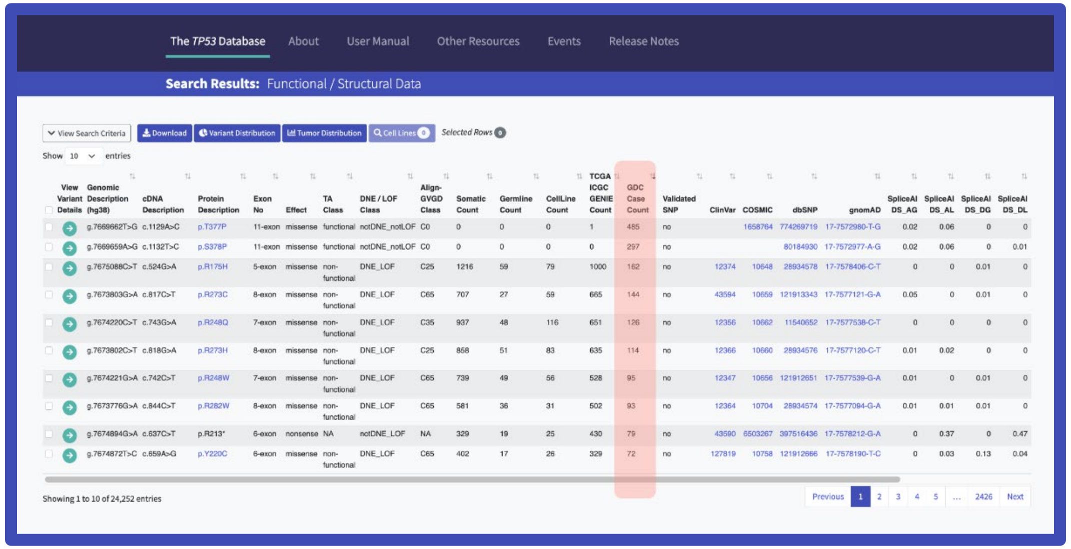


# Future Development:
# Easy Access to *TP53* dataset in BigQuery

- The current BigQuery tables are not yet public (*cf.* Mitelman Database)
- The current data tables are too complex
  - The data is extracted from 70 tables, which have over 500 columns all together
  - Need to optimize the data by trimming fields that are not related to *TP53* variants
  - Need to remove extraneous columns that were never exposed
- Making the data in BigQuery public will make it easily accessible to any researcher or clinician
- The field of the data analysis can then be easily expanded with arbitrary queries

# Future Development:
# Linking *TP53* variant data with GDC case data

With TP53 now part of the CRDC, we can use the data to inform analyses of CRDC data



**Search Results:** Functional / Structural Data

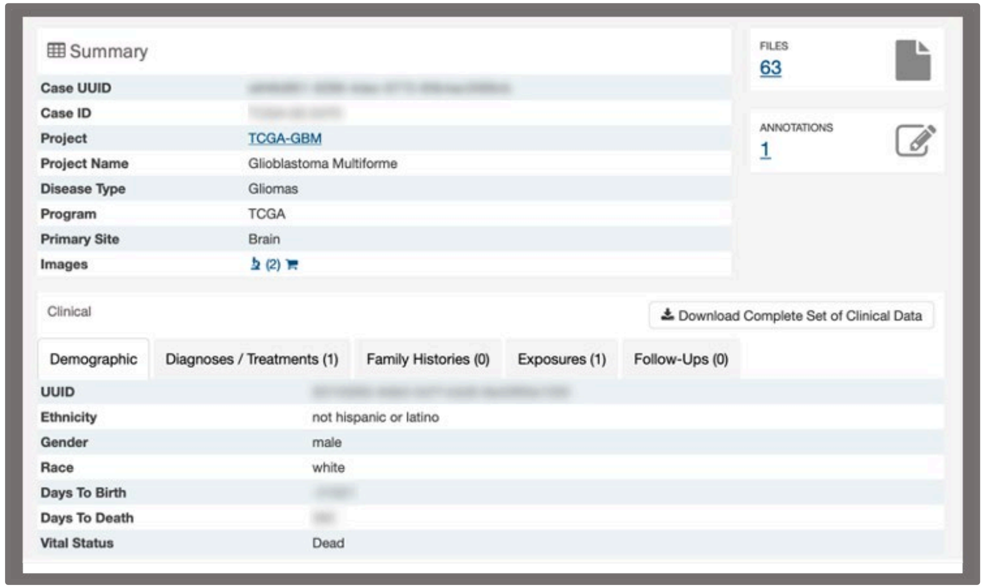
Show 10 entries

| ☐ | View Variant Details | Genomic Description (hg38) | cDNA Description | Protein Description | Exon No | Effect | TA Class | DNE / LOF Class | Align-GVGD Class | Somatic Count | Germline Count | CellLine Count | TCGA ICGC GENIE Count | GDC Case Count | Validated SNP | ClinVar | COSMIC | dbSNP | gnomAD | SpliceAI DS_AG | SpliceAI DS_AL | SpliceAI DS_DG | SpliceAI DS_DL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ☐ | → | g.7669662T>G | c.1129A>C | p.T377P | 11-exon | missense | functional | notDNE_notLOF | C0 | 0 | 0 | 0 | 1 | 485 | no |  | 1658764 | 774269719 | 17-7572980-T-G | 0.02 | 0.06 | 0 | 0 |
| ☐ | → | g.7669659A>G | c.1132T>C | p.S378P | 11-exon | missense | functional | notDNE_notLOF | C0 | 0 | 0 | 0 | 0 | 297 | no |  |  | 80184930 | 17-7572977-A-G | 0.02 | 0.06 | 0 | 0.01 |
| ☐ | → | g.7675088C>T | c.524G>A | p.R175H | 5-exon | missense | non-functional | DNE_LOF | C25 | 1216 | 59 | 79 | 1000 | 162 | no | 12374 | 10648 | 28934578 | 17-7578406-C-T | 0 | 0 | 0.01 | 0 |
| ☐ | → | g.7673803G>A | c.817C>T | p.R273C | 8-exon | missense | non-functional | DNE_LOF | C65 | 707 | 27 | 59 | 665 | 144 | no | 43594 | 10659 | 121913343 | 17-7577121-G-A | 0.05 | 0 | 0.01 | 0 |
| ☐ | → | g.7674220C>T | c.743G>A | p.R248Q | 7-exon | missense | non-functional | DNE_LOF | C35 | 937 | 48 | 116 | 651 | 126 | no | 12356 | 10662 | 11540652 | 17-7577538-C-T | 0 | 0 | 0 | 0 |
| ☐ | → | g.7673802C>T | c.818G>A | p.R273H | 8-exon | missense | non-functional | DNE_LOF | C25 | 858 | 51 | 83 | 635 | 114 | no | 12366 | 10660 | 28934576 | 17-7577120-C-T | 0.01 | 0.02 | 0 | 0 |
| ☐ | → | g.7674221G>A | c.742C>T | p.R248W | 7-exon | missense | non-functional | DNE_LOF | C65 | 739 | 49 | 56 | 528 | 95 | no | 12347 | 10656 | 121912651 | 17-7577539-G-A | 0.01 | 0 | 0.01 | 0 |
| ☐ | → | g.7673776G>A | c.844C>T | p.R282W | 8-exon | missense | non-functional | DNE_LOF | C65 | 581 | 36 | 31 | 502 | 93 | no | 12364 | 10704 | 28934574 | 17-7577094-G-A | 0.01 | 0.01 | 0.01 | 0 |
| ☐ | → | g.7674894G>A | c.637C>T | p.R213* | 6-exon | nonsense | NA | notDNE_LOF | NA | 329 | 19 | 25 | 430 | 79 | no | 43590 | 6503267 | 397516436 | 17-7578212-G-A | 0 | 0.37 | 0 | 0.47 |
| ☐ | → | g.7674872T>C | c.659A>G | p.Y220C | 6-exon | missense | non-functional | DNE_LOF | C65 | 402 | 17 | 26 | 329 | 72 | no | 127819 | 10758 | 121912666 | 17-7578190-T-C | 0 | 0.03 | 0.13 | 0.04 |

Showing 1 to 10 of 24,252 entries

Previous 1 2 3 4 5 … 2426 Next

*Prototype: TP53* variant search results with GDC case info



|  |  |
|---|---|
| Case UUID | [redacted] |
| Case ID | [redacted] |
| Project | TCGA-GBM |
| Project Name | Glioblastoma Multiforme |
| Disease Type | Gliomas |
| Program | TCGA |
| Primary Site | Brain |
| Images | (2) |

|  |  |
|---|---|
| UUID | [redacted] |
| Ethnicity | not hispanic or latino |
| Gender | male |
| Race | white |
| Days To Birth | [redacted] |
| Days To Death | [redacted] |
| Vital Status | Dead |

Genomic Data Common case page

## ISB-CGC

ISB

GENERAL DYNAMICS
Information Technology

**Elaine Lee**
William Longabaugh
Boris Aguilar
Lauren Hagen
Lauren Wolfe
Mi Tian
Suzanne Paquette
Ilya Shmulevich

David Pot
Danna Huffman
Deena Bleich
Fabian Seidl
Jacob Wilson
Poojitha Gundluru
Prema Venkatesan
**Owais Shahzada**

## DCEG

Division of Cancer Epidemiology &
Genetics at the National Cancer Institute

Kelvin de Andrade
Sharon Savage

## Original Team and IARC

Monica Hollstein
Curt C. Harris
Pierre Hainaut
Magali Olivier
Lucile Alteyrac
Jiri Zavadil

## Plus…

Elise Tookmanian, Chimene Kesserwan, James Manfredi, Jessica Hatton, Jennifer Loukissas, Lei Zhou, Megan Frone, Christian Kratz, David Malkin, Pierre Hainaut

https://tp53.isb-cgc.org/