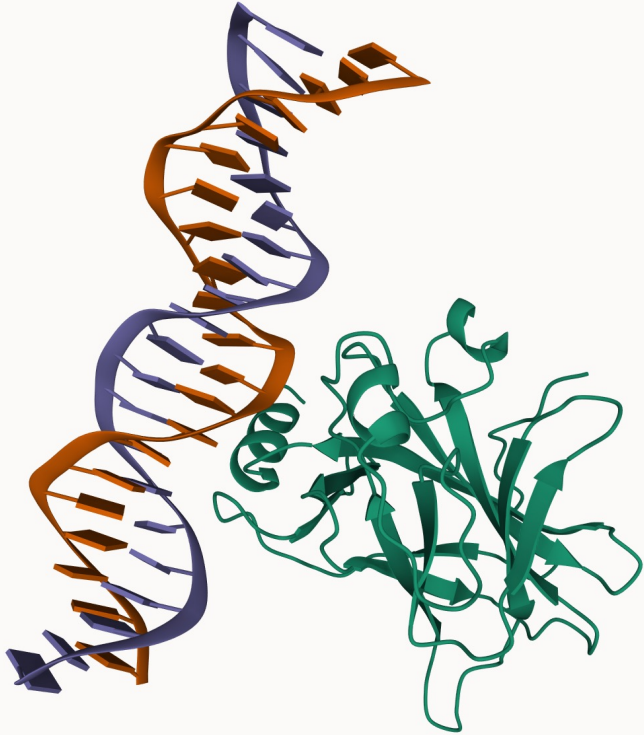


Breakout Session 1: Track B

**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

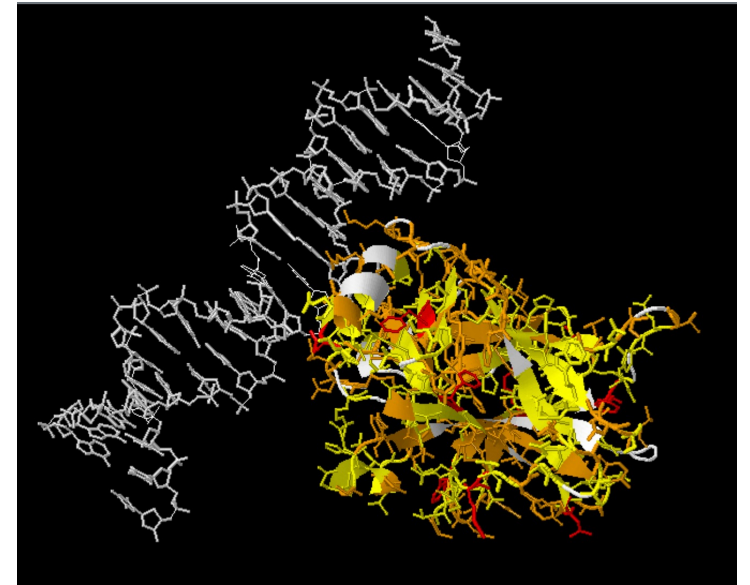
Database compiles *TP53* variant data from 1989

Currently holds information on 24,547 *TP53* variants

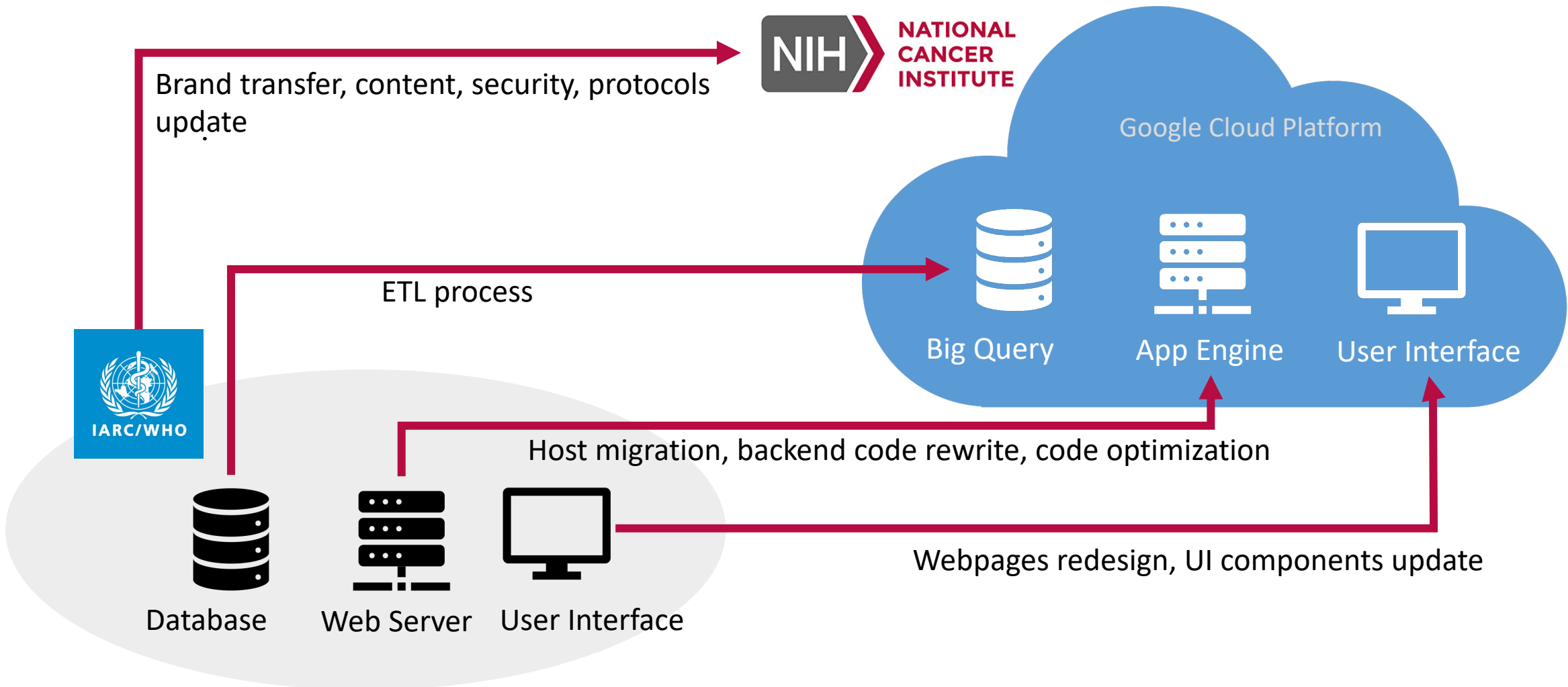
Database includes:

- *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



Transfer of Website and Database into the Cloud



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

<https://mitelmandatabase.isb-cgc.org>



Cases Cytogenetics Search Result

View Overall Chromosomal Imbalances: **121** View SQL Statement Download

Show 10 entries

Morphology	Topography	Karyotype	Case No	Reference	View Case
<input type="checkbox"/> Acute basophilic leukemia	39-40,X,-Y,t(2;6)(q27;q27),add(3)(q27),del(5)(p11),del(12)(p11),-13,-14,-15,-19		2	Gligoudis et al 2001, Eur J Haematol	View Karyotype Details
<input type="checkbox"/> Acute erythroleukemia (FAB type M6)	34,X,-Y,-1,-6,-8,t(9;22)(p11;q11),-10,-11,-12,-15,-19,-21,-del(22)(p15.3q22),-9,-10,der(11)t(9;11)(p12;p15),del(13)(q22),-17,-18		19	Pedersen et al 1997, Acta Haematol	View Karyotype Details
<input type="checkbox"/> Acute erythroleukemia (FAB type M6)	35-42,XX,r(13),-5,der(7)t(6;7)(q22;q17)(p15.3q22),-9,-10,der(11)t(9;11)(p12;p15),del(13)(q22),-17,-18		12	Cigudosa et al 2003, Genes Chromosomes Cancer	View Karyotype Details
<input type="checkbox"/> Acute erythroleukemia (FAB type M6)	37-48,X,-Y,-del(5)(q21),-8,-10,-11,-12,-13,-14,-15,-16,-17,-18,-19,-20,-21,-22,-23,-24,-25,-26,-27,-28,-29,-30,-31,-32,-33,-34,-35,-36,-37,-38,-39,-40,-41,-42,-43,-44,-45,-46,-47,-48,-49,-50,-51,-52,-9,-11,del(5)(q12q33),der(8)t(8;11)		9	Patnisk et al 2010, Am J Hematol	View Karyotype Details
<input type="checkbox"/> Acute erythroleukemia (FAB type M6)	37-48,X,-Y,-del(5)(q21),-8,-10,-11,-12,-13,-14,-15,-16,-17,-18,-19,-20,-21,-22,-23,-24,-25,-26,-27,-28,-29,-30,-31,-32,-33,-34,-35,-36,-37,-38,-39,-40,-41,-42,-43,-44,-45,-46,-47,-48,-49,-50,-51,-52,-9,-11,del(5)(q12q33),der(8)t(8;11)		1	Cigudosa et al 2003,	View Karyotype Details

All data is publicly available in BigQuery.

Google Cloud Explorer showing a BigQuery dataset named 'mitelmandb'. The Explorer pane shows a folder structure with 'prod' containing various tables like 'AuthorReference', 'Cytobands_hg38', 'CytoConverted', 'CytoConvertedLog', 'Cytogen', 'CytogenInv', 'CytogenInvValid', 'KaryAbnorm', 'KaryBit', 'KaryBreak', 'KaryClone', 'Koder', and 'MolBiolClassAssoc'.

Code editor showing a notebook titled 'Mitelman_Cytogenetics_Subsets.ipynb'. The notebook content includes:

```
1 SELECT DISTINCT
2   c.RefNo,
3   c.CaseNo,
4   c.InvNo,
5   Reference.Abbreviation,
6   Reference.Journal,
7   KoderT.Benaming AS Morph,
8   KoderT.Benaming AS Topo,
9   c.KaryShort,
10  c.KaryLong
11 FROM
12   mitelmandb.prod.CytogenInv c,
13   mitelmandb.prod.Reference Ref,
14   mitelmandb.prod.Cytogen Cytogen,
15   LEFT JOIN mitelmandb.prod.KoderT
16   ON
17   (Cytogen.Morph = KoderT.Kod AND KoderT.Benaming = Cytogen.Morph)
18   LEFT JOIN mitelmandb.prod.KoderT
19   ON
20   (Cytogen.Topo = KoderT.Kod AND KoderT.Benaming = Cytogen.Topo)
21   WHERE
22   Cytogen.RefNo = c.RefNo
23   AND Cytogen.CaseNo = c.CaseNo
24   AND c.RefNo = Reference.RefNo
25   AND c.CaseNo = Reference.CaseNo
26   AND c.InvNo = KaryBit.InvNo
27   AND c.InvNo = KaryBit.InvNo
28   AND c.InvNo = KaryBit.InvNo
29   AND c.InvNo = KaryBit.InvNo
30   AND c.InvNo = KaryBit.InvNo
```

The notebook also contains text describing the purpose of the query and the goal of the exercise.



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

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

The screenshot shows the homepage of the TP53 Database. At the top, there is a navigation bar with links for 'The TP53 Database', 'About', 'User Manual', 'Other Resources', 'Events', and 'Release Notes'. Below the navigation bar, a blue banner contains the text: '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'. A light purple announcement box follows, stating: '[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'. The main content area features six green-bordered cards, each with a plus sign icon and a title: 'Functional / Structural Data', 'Tumor Variants', 'Germline Variants', 'Cell Lines', 'Mouse Models', and 'Experimentally Induced Variants'. Each card contains a brief description of the data it represents.

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

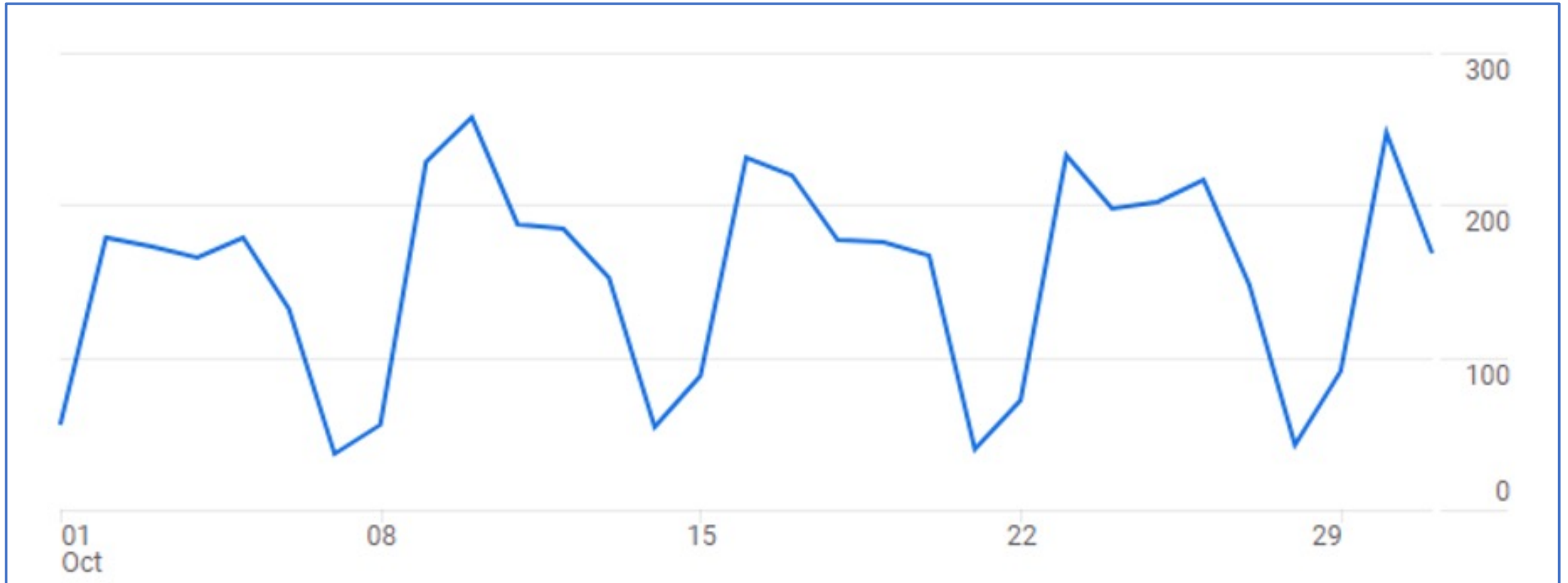
Application is now

- 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

The TP53 Database interface includes a search bar, navigation tabs (About, User Manual, Other Resources, Events, Release Notes), and a search filter for 'Functional / Structural Data: by Gene Variants'. The search results table displays columns for Genomic Description, cDNA (hg18), Protein Description, Exon No, Effect, Class, DNE, TA, LOP, GYGD, Somatic Count, Germline Count, CellLine Count, TCGA KOSG, GENE Count, Validated SNP, ClinVar, and COSMIC. The statistics section features a bar chart for 'Exon/Intron Distribution (N = 10,264)' and a pie chart for 'Variant Pattern (N = 10,264)'. The 3D Viewer section includes a 3D model of a protein structure and a list of structural impacts of variants.

Genomic Description	cDNA (hg18)	Protein Description	Exon No	Effect	Class	DNE	TA	LOP	GYGD	Somatic Count	Germline Count	CellLine Count	TCGA KOSG	GENE Count	Validated SNP	ClinVar	COSMIC
p.7																	
g.76985920C>T	c.11090G>A	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920A>A	c.11090C>T	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920G>A	c.11091G>C	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920T>A	c.11028A>G	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920T>G	c.11028A>C	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920G>T	c.11029C>A	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920A>A	c.11029C>T	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920G>T	c.11024G>A	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985920A>A	c.11026C>T	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	no		
g.76985940A>A	c.11052C>T	p.7	11-exon	NA	NA	NA	NA	NA	NA	0	0	0	0	0	validated		

TP53 Database Usage



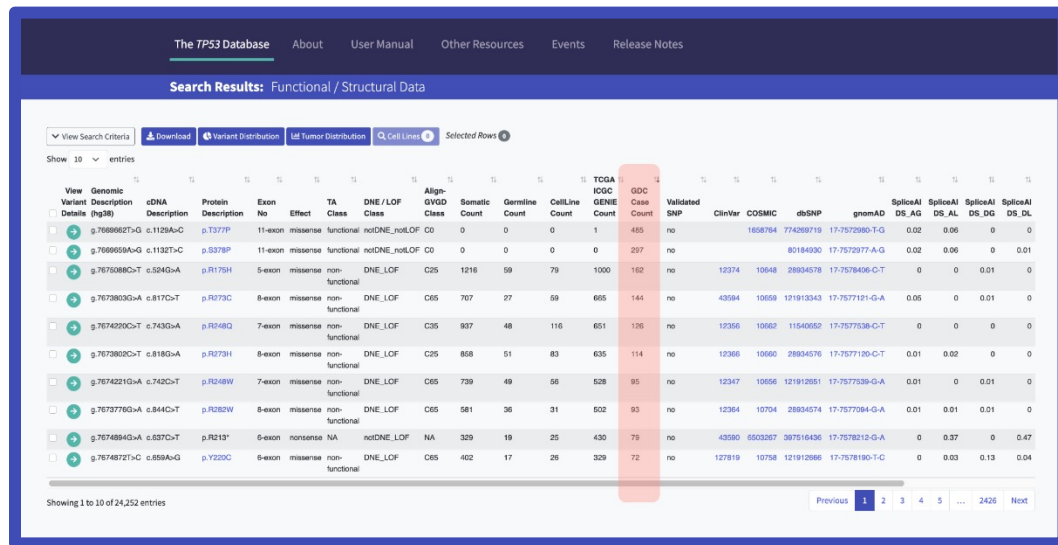
Future Development:

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

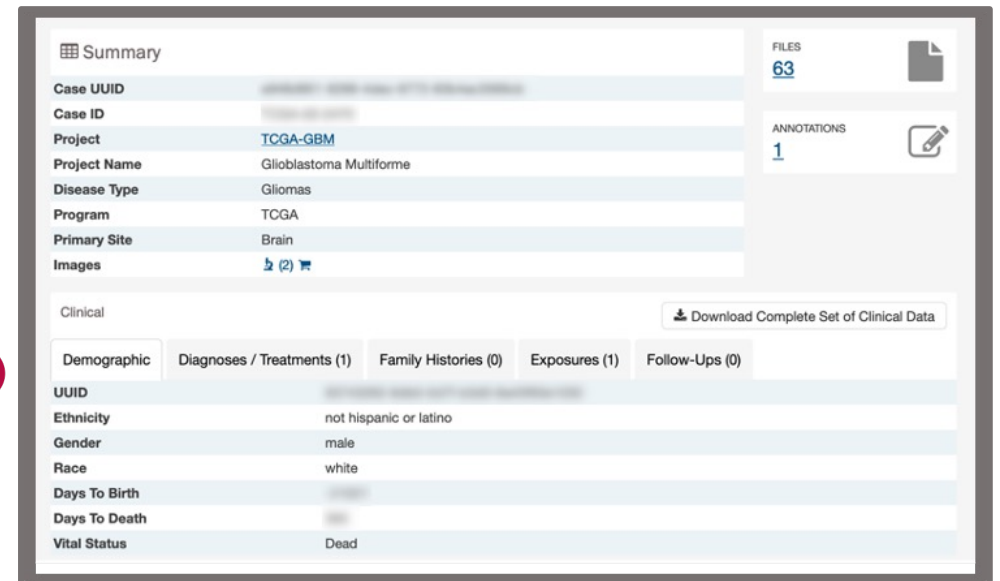
With *TP53* now part of the CRDC, we can use the data to inform analyses of CRDC data



The screenshot shows the 'The TP53 Database' search results for 'Functional / Structural Data'. The table includes columns for variant details and GDC case counts. A red box highlights the 'GDC Case Count' column.

Variant Description	Genomic Description	cDNA Description	Protein Description	Exon No	Effect	TA Class	DNE / LOF Class	Allo-GVGD Class	Somatic Count	GermLine Count	CellLine Count	TCGA ICDC GENE Count	GDC Case Count	Validated SNP	ClinVar	COSMIC	dbSNP	gnomAD	DS_AG	DS_AL	DS_DG	DS_DL
g.7686962T>G c.1123A>G	p.T377P			11	missense	functional	ncDNE_ncLOF	C0	0	0	0	1	465	no		1658794	774269719	17-7572980-T-G	0.02	0.06	0	0
g.7686959A>G c.1132T>C	p.S378P			11	missense	functional	ncDNE_ncLOF	C0	0	0	0	0	297	no		80184930	17-7572977-A-G	0.02	0.06	0	0.01	
g.7676088C>T c.524G>A	p.R175H			5	missense	non-functional	DNE_LOF	C25	1216	59	79	1000	182	no	12374	10648	28934578	17-7578406-G-T	0	0	0.01	0
g.7673803G>A c.817C>T	p.R273C			8	missense	non-functional	DNE_LOF	C65	707	27	59	655	144	no	43994	10659	121913343	17-757121-G-A	0.06	0	0.01	0
g.7674229C>T c.743G>A	p.R248Q			7	missense	non-functional	DNE_LOF	C35	937	48	116	651	126	no	12366	10662	11540052	17-7577038-C-T	0	0	0	0
g.7673802C>T c.818G>A	p.R273H			8	missense	non-functional	DNE_LOF	C25	858	51	83	635	114	no	12366	10660	28934576	17-7577120-G-T	0.01	0.02	0	0
g.7674221G>A c.742C>T	p.R248W			7	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	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.527C>T	p.R213*			6	non-sense	NA	ncDNE_LOF	NA	329	19	25	430	79	no	40390	6503267	397516436	17-7578212-G-A	0	0.37	0	0.47
g.7674827T>C c.859A>G	p.Y220C			6	missense	non-functional	DNE_LOF	C65	402	17	26	329	72	no	12719	10758	121912686	17-7578190-T-C	0	0.03	0.13	0.04

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



The screenshot shows the 'Summary' page for a case in the Genomic Data Commons. It includes fields for Case UUID, Case ID, Project (TCGA-GBM), Project Name (Glioblastoma Multiforme), Disease Type (Gliomas), Program (TCGA), Primary Site (Brain), and Images (2). Clinical data is also visible, including Demographic, Diagnoses / Treatments (1), Family Histories (0), Exposures (1), and Follow-Ups (0). The patient's vital status is listed as 'Dead'.

Field	Value
Case UUID	[Redacted]
Case ID	[Redacted]
Project	TCGA-GBM
Project Name	Glioblastoma Multiforme
Disease Type	Gliomas
Program	TCGA
Primary Site	Brain
Images	2 (2)
Clinical	[Download Complete Set of Clinical Data]
Demographic	[Redacted]
Diagnoses / Treatments (1)	[Redacted]
Family Histories (0)	[Redacted]
Exposures (1)	[Redacted]
Follow-Ups (0)	[Redacted]
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



Elaine Lee

William Longabaugh

Boris Aguilar

Lauren Hagen

Lauren Wolfe

Mi Tian

Suzanne Paquette

Ilya Shmulevich

GENERAL DYNAMICS
Information Technology

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/>