



# An Overview of Cancer Data Repositories

May 24, 2023

Warren A. Kibbe, Ph.D.  
Vice Chair & Professor of  
Biostatistics and Bioinformatics  
Duke University School of Medicine  
[warren.kibbe@duke.edu](mailto:warren.kibbe@duke.edu)



**DukeHealth**

# We will take a tour of:

- AACR Project GENIE
- NCI Genomic Data Commons
- The Gabriela Miller Kids First Data Portal
- dbGaP
- The Human BioMolecular Atlas Program (HuBMAP)
- Human Tumor Atlas Network (HTAN)
- CIViC DB
- MyCancerGenome
- COSMIC
- Childhood Cancer Data Catalog

By no means are these the only or even the most important ones!

# Current sources of data

molecular



genome



pathology



imaging



labs



notes



sensors



Our ability to generate biomedical data continues to grow in terms of variety and volume

# Things to think about



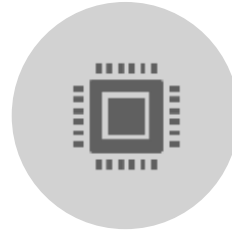
Many different metaphors for building a data repository



Many different methods for making data FAIR (Findable, Accessible, Interoperable, Reusable)



Data wrangling – organizing, formatting, harmonizing, semantically annotating – is hard work and different resources take different approaches



Usability is dependent on the use case, the tools, the problem domain

# Design Considerations



WHERE DO YOU WANT FLEXIBILITY?  
FILE FORMATS?  
DATA TYPES?  
DEFINITIONS?



HOW MUCH CAN YOU ANTICIPATE AT THE BEGINNING?



WHERE IS IT LIKELY THAT YOUR REQUIREMENTS WILL CHANGE?



WHO ARE YOUR USERS?  
BIOLOGISTS?  
CHEMISTS?  
CLINICIANS?  
QUANTS?



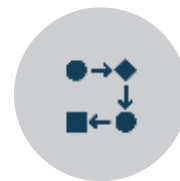
WILL THEY COME TO EXPLORE (HYPOTHESIS GENERATION)?



WILL THEY BRING AN EXISTING QUESTION (HYPOTHESIS ANSWERING)?



DO THEY WANT TO USE THEIR OWN ANALYSIS TOOLS?



DO THEY WANT TO BUILD NEW TOOLS TO EXPLORE THE DATA?

# What is the right data resource to use?



What is the question you want to answer?



Are you exploring how to answer the problem or ready to analyze?



Do you have an analysis plan and workflow defined?



Are your analysis tools appropriate for the resource you want to use?



Do you need to explore a dataset to understand if the data is available?  
Does the resource support the kind of exploration you want to do?

# What is the right data resource to use?

Does it have the right kind of data? (clinical trial data, scRNA-seq, specific cell line, disease area, model system)

What are the access policies? (unrestricted, controlled access)

Level of security required (none, limited, regulated [think HIPAA], very sensitive)

Ability to peruse the metadata?

Can you download the data?

If it is an enclave/closed ecosystem, what tools are supported?

Can you bring your own tools?

# A useful catalog of data resources

<https://datacatalog.ccdi.cancer.gov>

The screenshot shows the homepage of the National Cancer Institute Childhood Cancer Data Initiative Data Catalog. At the top, there is a navigation bar with the NIH logo and the text "NATIONAL CANCER INSTITUTE Childhood Cancer Data Initiative Data Catalog". A search bar labeled "Documentation Search" is located in the top right. Below the navigation bar, there is a green header with links for "Home", "Search Catalog", "Participating Resources", "CCDI Studies", and "About". The main content area features a large image of a man and a young boy, with the boy holding a teddy bear and a stethoscope. The text "Childhood Cancer Data Catalog" is overlaid on the image. Below this, there is a search bar labeled "Search for Datasets" and a featured item section titled "May the Data Be with You" which mentions "3 new resources, 1 new dataset, and many other data updates. Check out all the Catalog changes. Read More". At the bottom, there are four featured resource cards: "Imaging", "The database of Genotypes and Phenotypes", "The Jackson Laboratory PDX Models", and "Therapeutic Research Effective Treatments".

NIH NATIONAL CANCER INSTITUTE  
Childhood Cancer Data Initiative  
Data Catalog

Documentation Search

Home Search Catalog Participating Resources CCDI Studies About

## Childhood Cancer Data Catalog

A searchable database of pediatric data resources, sharing clinical care and research data generated by the pediatric cancer research community.

Search for Datasets

FEATURED ITEMS:  
**May the Data Be with You**  
3 new resources, 1 new dataset, and many other data updates. Check out all the Catalog changes. [Read More](#)

**Imaging**  
which de-identifies and hosts a large image... [READ MORE](#)

**The database of Genotypes and Phenotypes**  
The database of Genotypes and Phenotypes (dbGaP) was developed to archive and... [READ MORE](#)

**The Jackson Laboratory PDX Models**  
Patient-Derived Xenografts (PDX) are generated through the engraftment and passa... [READ MORE](#)

**Therapeutic Research Effective Treatments**  
The Therapeutic Effective Treatments



# Childhood Cancer Data Catalog: TCIA projects

The screenshot shows the NIH National Cancer Institute Childhood Cancer Data Initiative Data Catalog. The page is titled "Search Results" and displays a search bar with the text "Search the Catalog" and a "SUBMIT" button. A filter for "Resources: TCIA" is applied. The search results are displayed in a table view, showing three entries:

Resources	Annotations for Chemotherapy and Radiation Therapy in Treating Young Patients With Newly Diagnosed, Previously Untreated, High-Risk Medulloblastoma/PNET (ACN50332-Tumor-Annotations)	Annotations for Combination Chemotherapy and Radiation Therapy in Treating Young Patients With Newly Diagnosed Hodgkin Lymphoma (AH000831-Tumor-Annotations)	Annotations for Combination Chemotherapy and Surgery in Treating Young Patients With Wilms Tumor (AREN0534-Tumor-Annotations)
<input type="checkbox"/> All of Us			
<input type="checkbox"/> CBTRUS			
<input type="checkbox"/> CCDI			
<input type="checkbox"/> CC55			
<input checked="" type="checkbox"/> CGC			
<input type="checkbox"/> CGCI			
<input type="checkbox"/> CIVIC			
<input type="checkbox"/> CLIC			
<input type="checkbox"/> COG			
<input type="checkbox"/> dbGaP			
<input checked="" type="checkbox"/> DepMap			
<input type="checkbox"/> Fibroregistry			
<input type="checkbox"/> GDC			
<input type="checkbox"/> GEO			
<input type="checkbox"/> Greehey			
<input type="checkbox"/> HCMI			
<input checked="" type="checkbox"/> HitWalker2			

Each entry includes a "Collection" button and a "TCIA" icon. The first entry has a "Case Count: 370" and a "Description" that states: "This dataset contains image annotations derived from the NCI Clinical Trial 'Chemotherapy and Radiation Therapy in Treating Young Patients With Newly Diagnosed, Previously Untreated, High-Risk Medulloblastoma/PNET (ACN50332)'. This curated dataset provides a comprehensive picture of imaging in pediatric patients with newly diagnosed primitive neuroectodermal tumors throughout their treatment and until any potential relapse. This is the largest known dataset of patients with supratentorial primit ...". The second entry has a "Case Count: 169" and a "Description" that states: "This dataset contains image annotations derived from the NCI Clinical Trial 'Combination Chemotherapy and Radiation Therapy in Treating Young Patients With Newly Diagnosed Hodgkin Lymphoma (AH000831)'. The key objective of this project is to generate a large and highly curated imaging dataset of pediatric Hodgkin lymphoma patients with annotations suitable for cancer researchers and AI developers." The third entry has a "Case Count: 169" and a "Description" that states: "This dataset contains image annotations derived from the NCI Clinical Trial 'Combination Chemotherapy and Surgery in Treating Young Patients With Wilms Tumor (AREN0534)'. The key objective of this project is to generate a large and highly curated imaging dataset of pediatric Wilms tumor patients with annotations suitable for cancer researchers and AI developers." The page also includes a "Sort by" dropdown set to "Dataset", "Results per Page: 10", and "Showing 1-10 of 13" results.

# A short scenario



You are interested in exploring non-small cell lung cancer and identifying the top 5 mutations. You want plot either progression free survival or a survival curve for each mutation.



Possible data sources: AACR Project GENIE and the NCI Genomic Data Commons (part of the NCI Cancer Research Data Commons)

# AACR Project GENIE cBioPortal resource

Start with Cohort 13.1 – data from 167,358 samples

The screenshot shows the cBioPortal website interface. At the top, there is a navigation bar with links for Data Sets, Web API, R/MATLAB, Tutorials/Webinars, FAQ, News, Visualize Your Data, and About. The user is logged in as wakibbe@gmail.com. The main content area features a 'Query' section with a search bar and a 'Select Studies for Visualization & Analysis' section. In this section, '1 study selected (167358 samples)' is shown. A list of studies is displayed, with 'GENIE Cohort v13.1-public' selected, showing 167,358 samples. Other studies include 'AACR Project GENIE AKT1 Cohort' (428 samples), 'DFCI-Profile Glioma Cohort 2013-2016 (DFCI, Nat Med 2020)' (1335 samples), 'ERBB2 Cohort (GENIE, 2022)' (313 samples), 'GENIE BPC CRC v2.0-public' (1501 samples), 'GENIE BPC NSCLC v2.0-public' (2004 samples), and 'Metastatic Breast Cancer: 2013-2016 (DFCI, OCR 2020)' (856 samples). Below the list, there are buttons for 'Query By Gene' and 'Explore Selected Studies'. On the right side, there is a 'What's New' section with a tweet from cBioPortal (@cbioportal) dated Apr 18, mentioning a poster presentation at #AACR23!

# AACR Project GENIE cbioportal resource

Select the 24,110 NSCLC

GENIE Cohort v13.1-public  
GENE v13.1-public

Selected: 148,222 patients | 167,358 samples

Cancer Type	#	Freq %
Non-Small Cell Lung Cancer	24,110	14.4%
Breast Cancer	16,004	9.6%
Colorectal Cancer	15,482	9.3%
Glioma	10,074	6.0%
Pancreatic Cancer	6,880	4.1%
Melanoma	6,794	4.1%
Ovarian Cancer	5,095	3.6%
Leukemia	5,931	3.5%
Prostate Cancer	5,731	3.4%
Mature B-Cell Neoplasms	5,515	3.3%
Cancer of Unknown Primary	5,338	3.2%

Cancer Type Detailed	#	Freq %
Lung Adenocarcinoma	18,430	11.0%
Breast Invasive Ductal Carcinoma	9,383	5.6%
Colon Adenocarcinoma	8,949	5.3%
Pancreatic Adenocarcinoma	5,928	3.5%
Prostate Adenocarcinoma	5,581	3.3%
High-Grade Serous Ovarian Cancer	3,587	2.1%
Bladder Urothelial Carcinoma	3,480	2.1%
Melanoma	3,439	2.1%
Colorectal Adenocarcinoma	3,425	2.0%
Acute Myeloid Leukemia	3,092	1.8%
Invasive Breast Carcinoma	2,944	1.8%

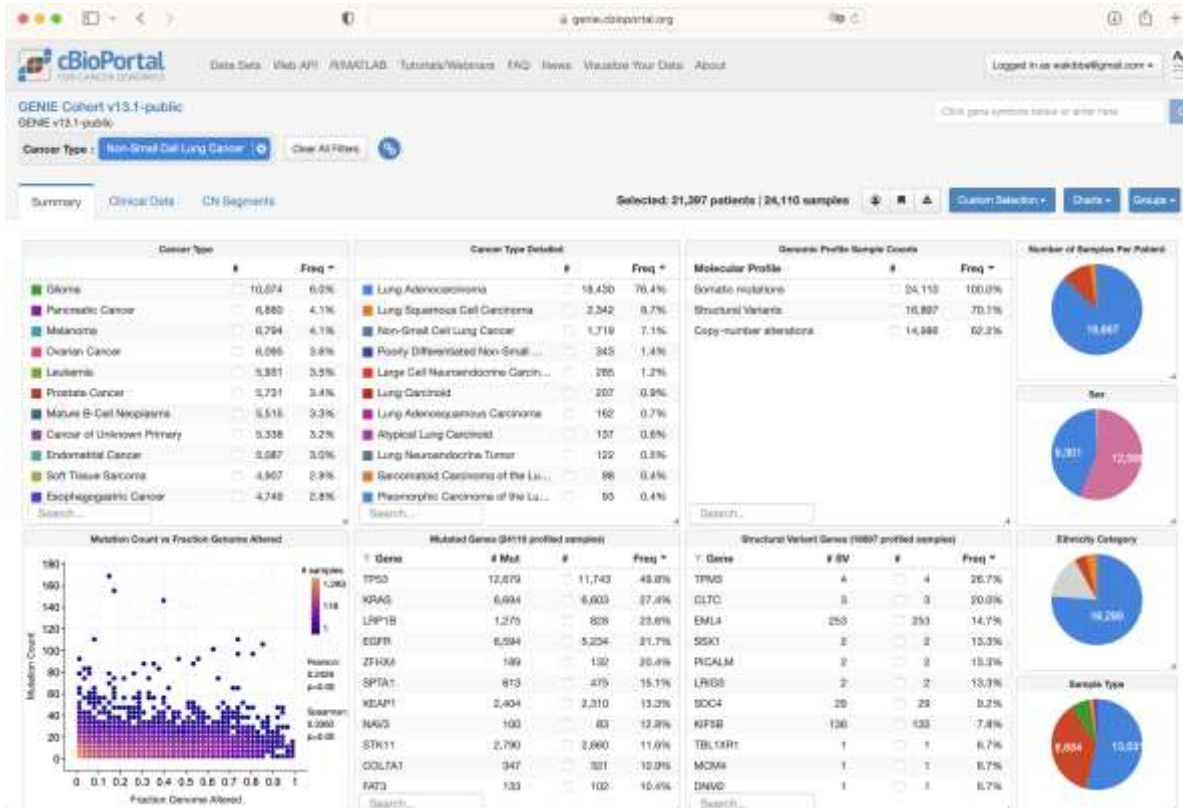
Genomic Profile Sample Counts	#	Freq %
Somatic mutations	167,358	100.0%
Structural Variants	129,396	77.3%
Copy-number alterations	117,069	70.0%

Mutated Genes (167358 profiled samples)	Gene	# Mut	#	Freq %
	TP53	70,349	64,168	38.6%
	KRAS	25,446	24,958	14.9%

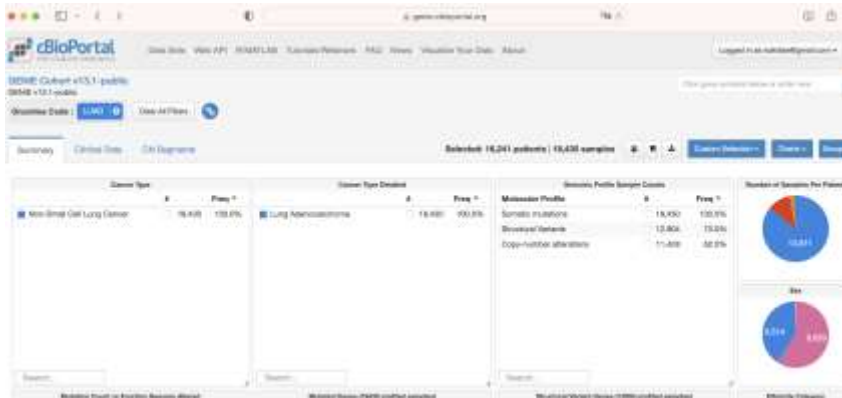
Structural Variant Genes (129396 profiled samples)	Gene	# SV	#	Freq %
	KIAA1549	248	248	30.5%
	RMBP	16	16	8.1%

# AACR Project GENIE cBioportal resource

NSLC View



# AACR Project GENIE cBioportal resource



Lung  
Adenocarcinoma  
18,430 cases

CNA Genes (14988 profiled samples)				
Gene	Cytoband	CNA	#	Freq
CDKN2A	9p21.3	HOMDEL	1,501	10.0%
CDKN2B	9p21.3	HOMDEL	1,399	9.5%
MTAP	9p21.3	HOMDEL	378	7.6%
NKX2-1	14q13.3	AMP	773	5.3%
EGFR	7p11.2	AMP	738	4.9%
MDM2	12q15	AMP	723	4.9%
TRIP13	5p15.33	AMP	80	4.8%
TERT	5p15.33	AMP	603	4.3%
MYC	8q24.21	AMP	633	4.3%
FDXA1	14q21.1	AMP	476	4.0%
NFKBIA	14q13.2	AMP	533	3.7%

Mutated Genes (24110 profiled samples)			
Gene	# Mut	#	Freq
TP53	12,679	11,743	49.8%
KRAS	6,694	6,803	27.4%
LRP1B	1,275	828	23.6%
EGFR	6,594	5,234	21.7%
ZFHX4	189	132	20.4%
SPTA1	613	475	15.1%
KEAP1	2,404	2,310	13.3%
NAV3	100	83	12.8%
STK11	2,790	2,660	11.6%
COL7A1	347	321	10.9%
FAT3	133	102	10.4%

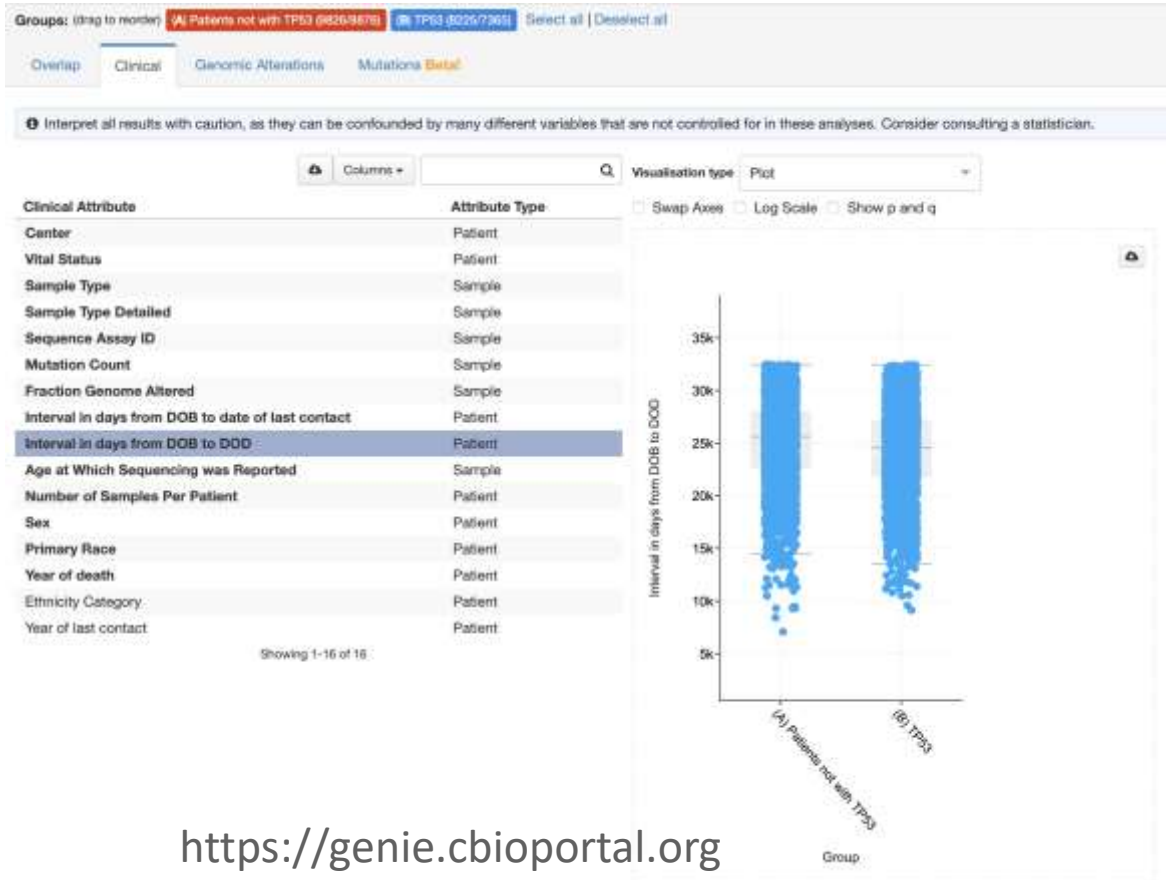
Structural Variant Genes (16987 profiled samples)			
Gene	# SV	#	Freq
TPM3	4	4	26.7%
CLTC	3	3	20.0%
EML4	253	253	14.7%
SSX1	2	2	13.3%
PICALM	2	2	13.3%
LRIG3	2	2	13.3%
SDC4	29	29	9.2%
KIF5B	136	135	7.8%
TBL1XR1	1	1	6.7%
MCM4	1	1	6.7%
DNM2	1	1	6.7%

# AACR Project GENIE cBioportal resource

'Manhattan' plot of mutations in TP53



# AACR Project GENIE cBioPortal resource



Not a survival analysis!

Looks like patients with somatic TP53 mutations die younger

Many potential confounders!

<https://genie.cbioportal.org>



# cBioPortal makes plotting survival data easy – but not with GENIE data! This plot is men vs women with bladder cancer



<https://cbioportal.org>

# Now we will explore the NCI Genomic Data Commons

**NATIONAL CANCER INSTITUTE**  
GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login GDC App

### Harmonized Cancer Datasets

## Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Search: e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-ADG2

#### Data Portal Summary

Data Release 37.0 - March 29, 2023

Category	Count
PROJECTS	78
PRIMARY SITES	68
CASES	86,962
FILES	931,947
GENES	22,501
MUTATIONS	2,885,293

#### Cases by Major Primary Site

Primary Site	Cases
Adrenal Gland	1
Bile Duct	1
Bladder	1
Brain	1
Breast	1
Colon	1
Colon Rectal	1
Esophagus	1
Eye	1
Heart and Esoph	1
Kidney	1
Liver	1
LUNG	11
Lymph Node	1
Nervous System	1
Ovary	1
Pancreas	1
Prostate	1
Skin	1
Soft Tissue	1
Stomach	1
Testis	1
Thyroid	1
Uterus	1

#### GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

- Data Portal
- Website
- API
- Data Transfer Tool
- Documentation
- Data Submission Portal
- Publications

# Genomic Data Commons

Lung cancer – 12,262 cases

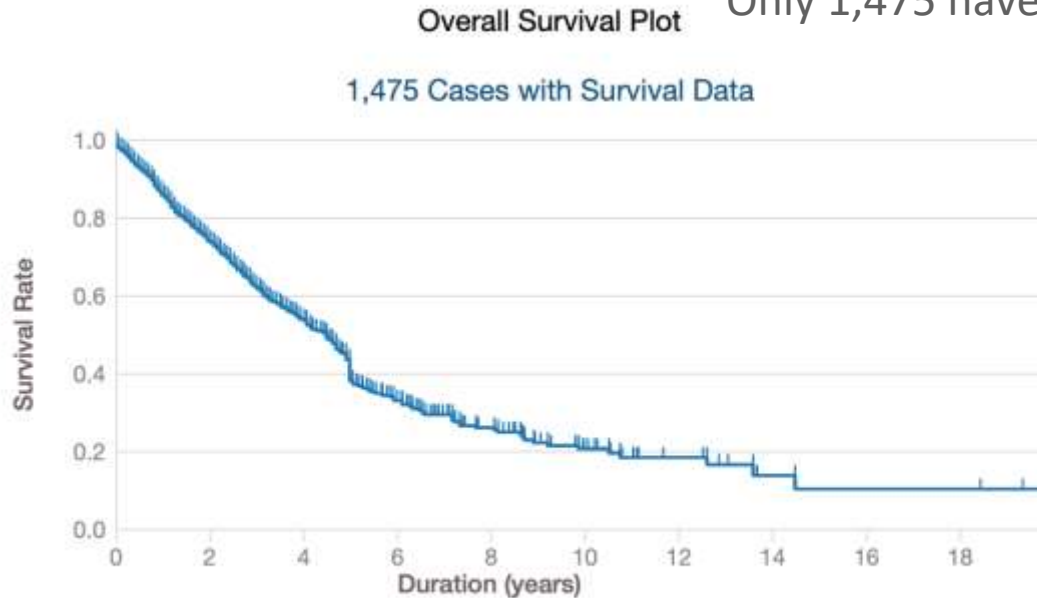
The screenshot displays the GDC Data Portal interface for lung cancer cases. The search filters include 'Primary Site' as 'bronchus and lung', 'Tissue Or Organ Of Origin' as 'lung, nos', and 'lower lobe, lung, nos'. The results show 12,262 cases, 21,318 genes, and 443,974 mutations. An 'Overall Survival Plot' is shown for 1,475 cases with survival data, with a survival rate of approximately 0.25 at 16 years. A table of somatic mutations is displayed below, showing DNA changes, types, consequences, and their impact on survival.

DNA Change	Type	Consequences	# Affected Cases in Cohort	# Affected Cases Across the GDC	Impact	Survival
chr12.p.25245351C>A	Substitution	Missense KRAS G12C	103 / 1,382 (7.4%)	195 / 14,783	High Impact	Survival
chr7.q.50191822T>G	Substitution	Missense EGFR L858R	55 / 1,382 (3.9%)	55 / 14,783	High Impact	Survival
chr12.p.25245350C>A	Substitution	Missense KRAS G12V	54 / 1,382 (3.8%)	287 / 14,783	High Impact	Survival
chr12.p.25245350G>T	Substitution	Missense KRAS G12D	32 / 1,382 (2.3%)	301 / 14,783	High Impact	Survival
chr7.q.55174772delGGAATTAA	Deletion	Inframe Deletion EGFR E746_A750del	38 / 1,382 (2.7%)	39 / 14,783	High Impact	Survival
chr3.q.17921822G>A	Substitution	Missense PIK3CA E545K	23 / 1,382 (1.6%)	212 / 14,783	High Impact	Survival
chr7.q.7970129C>A	Substitution	Missense TP53 R158L	22 / 1,382 (1.6%)	25 / 14,783	High Impact	Survival
chr12.p.25245350C>G	Substitution	Missense KRAS G12A	20 / 1,382 (1.4%)	57 / 14,783	High Impact	Survival
chr17.q.7575994C>A	Substitution	Splice Region TP53 T125=	20 / 1,382 (1.4%)	35 / 14,783	Low Impact	Survival
chr2.p.17921822G>A	Substitution	Missense PIK3CA E542K	18 / 1,382 (1.2%)	195 / 14,783	High Impact	Survival

# Genomic Data Commons

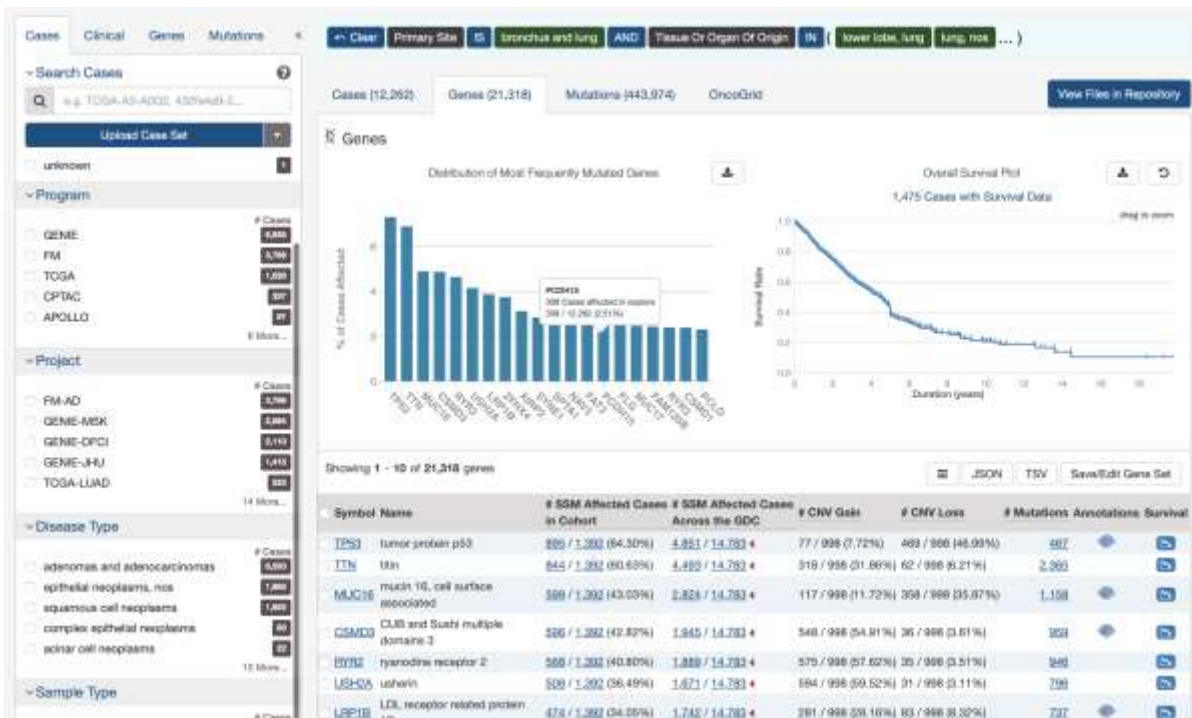
Lung cancer - 12,262 cases

Only 1,475 have survival data

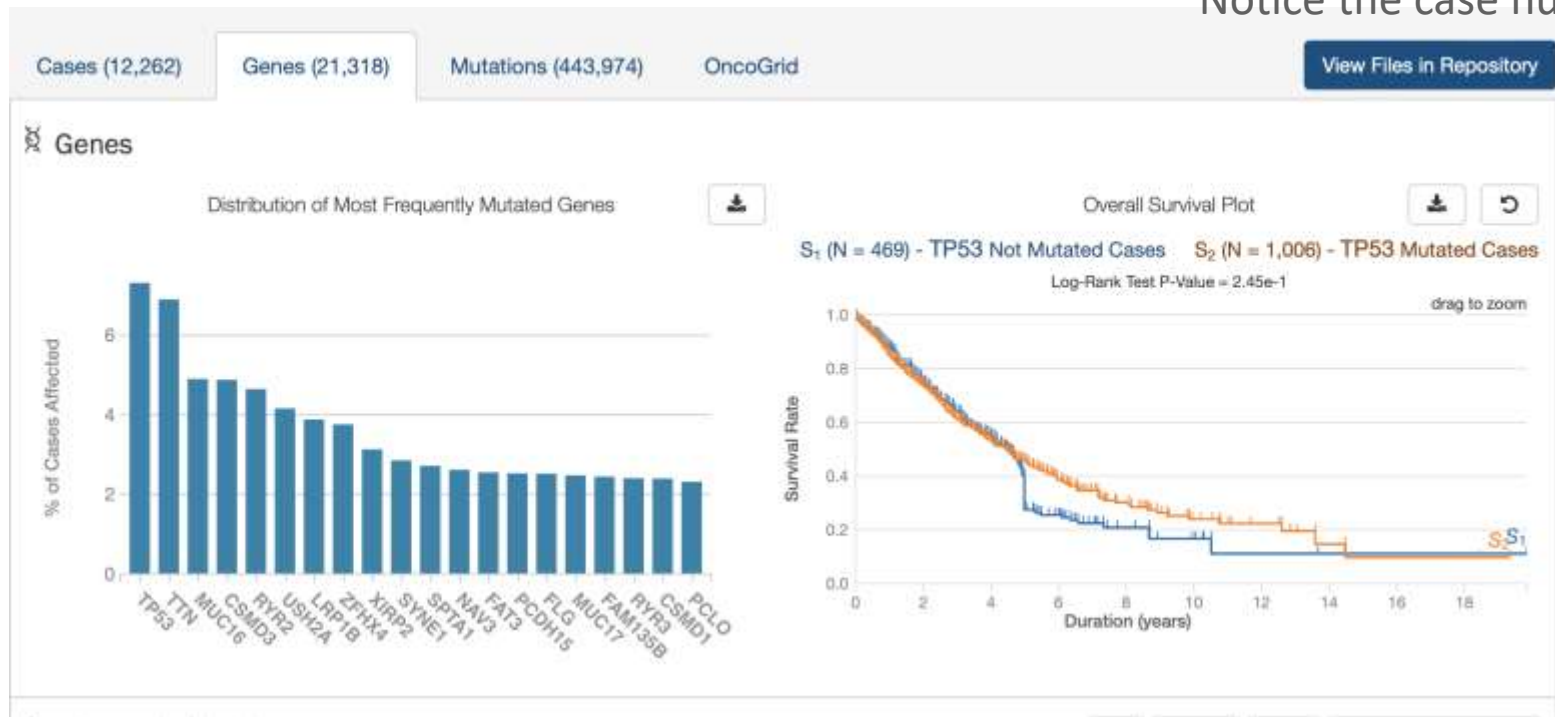


# Genomic Data Commons

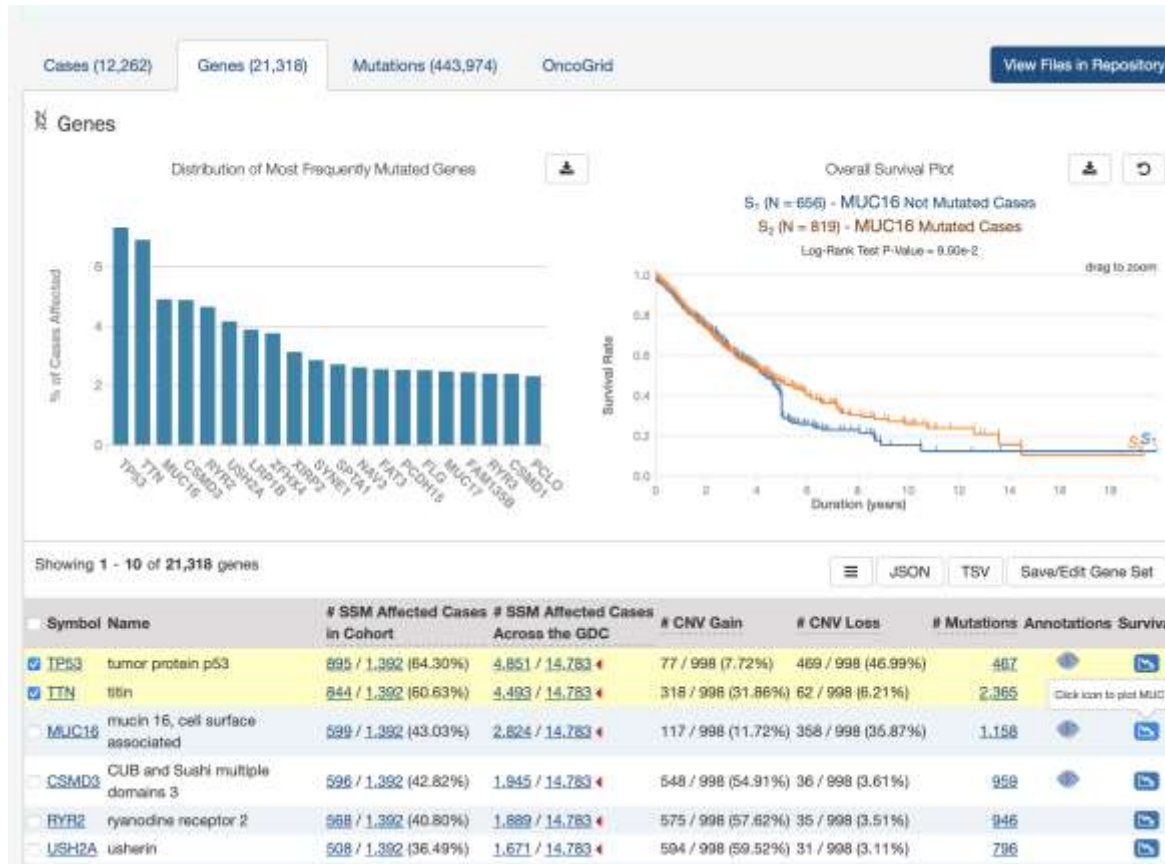
## Genes view



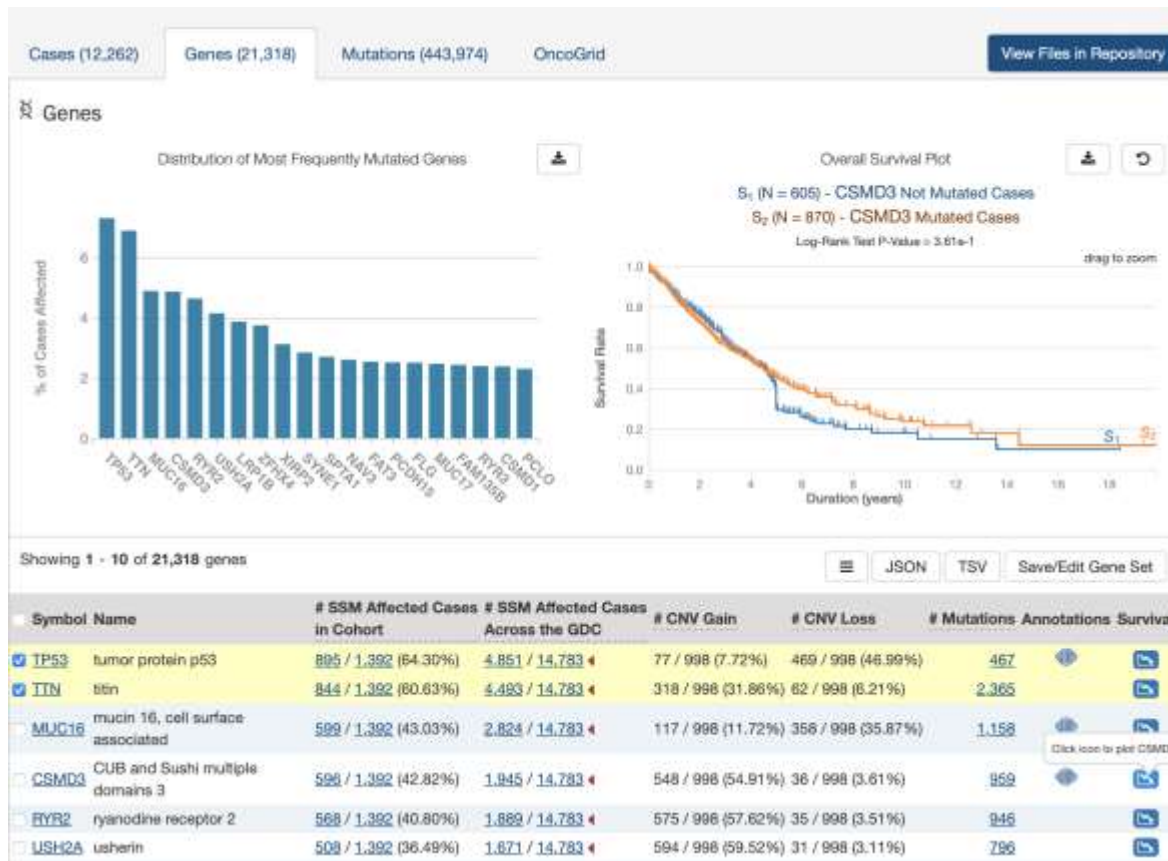
Notice the case numbers

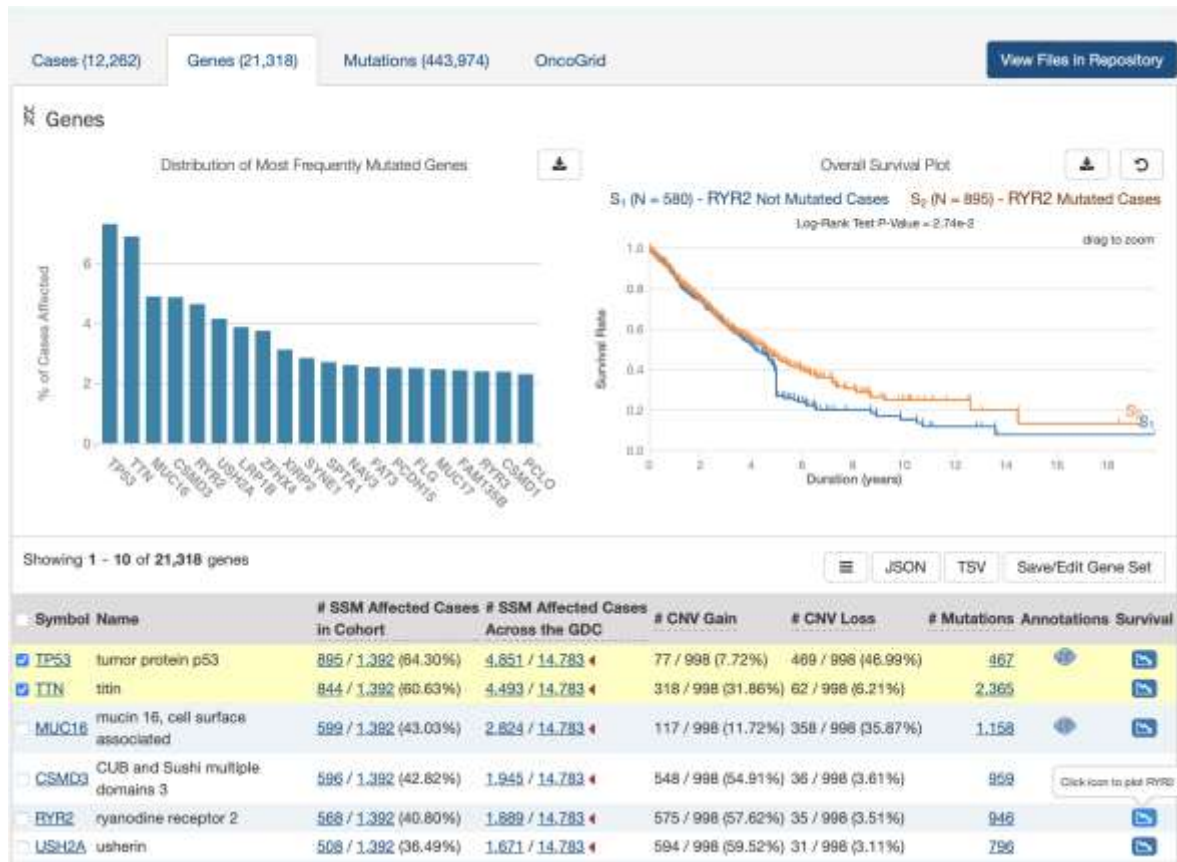


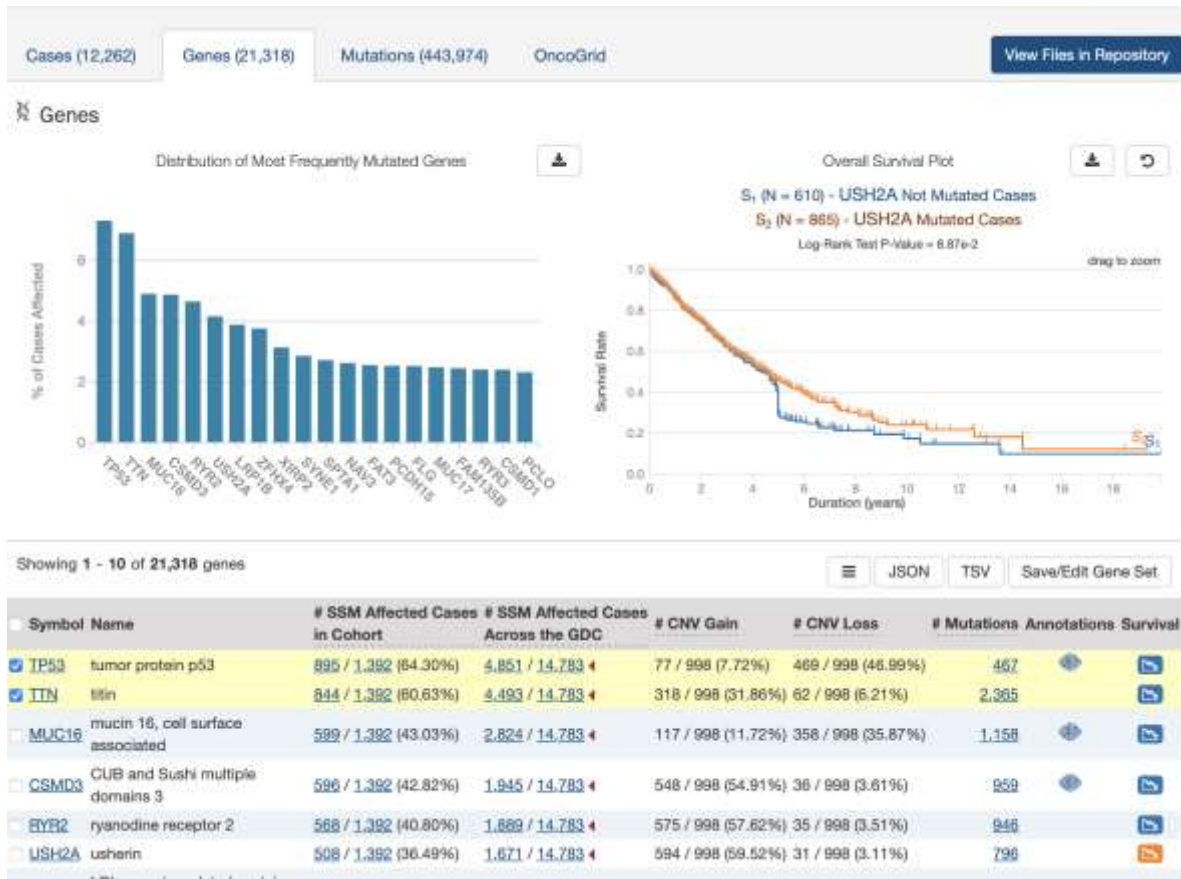










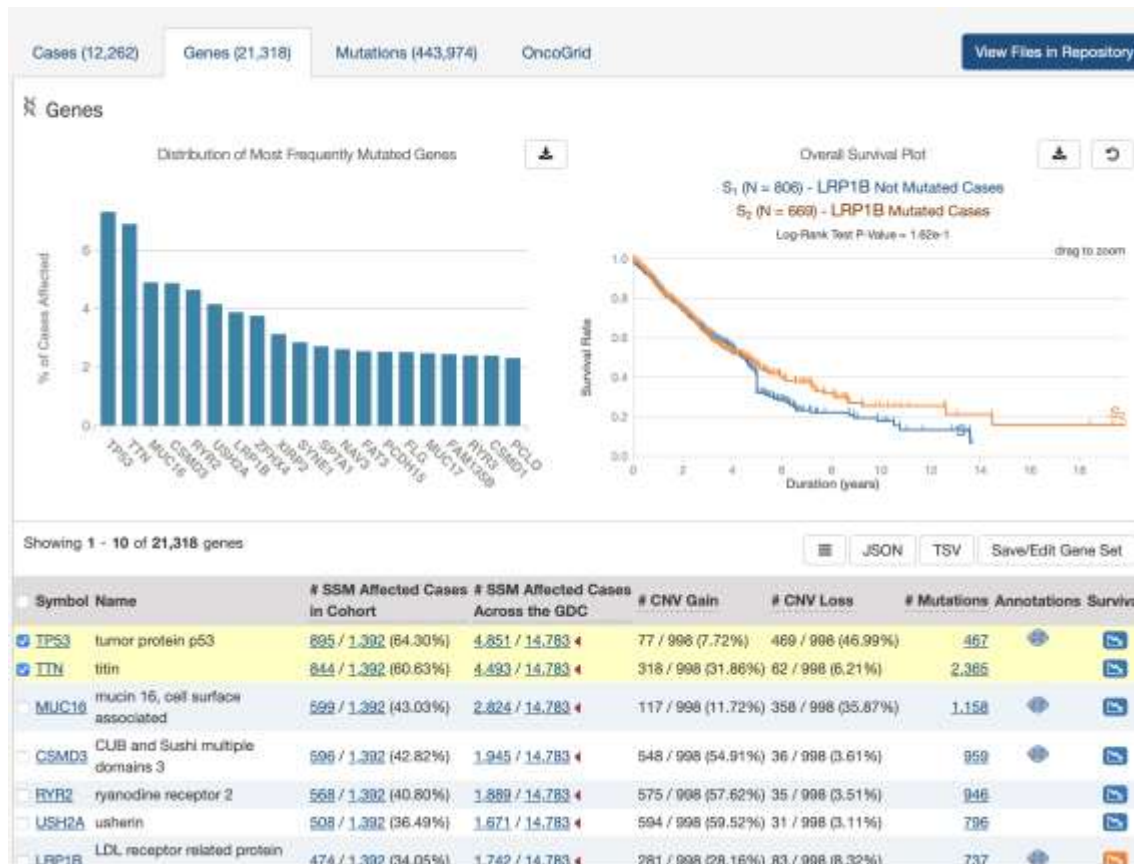


# Genomic Data Commons

LRP1B

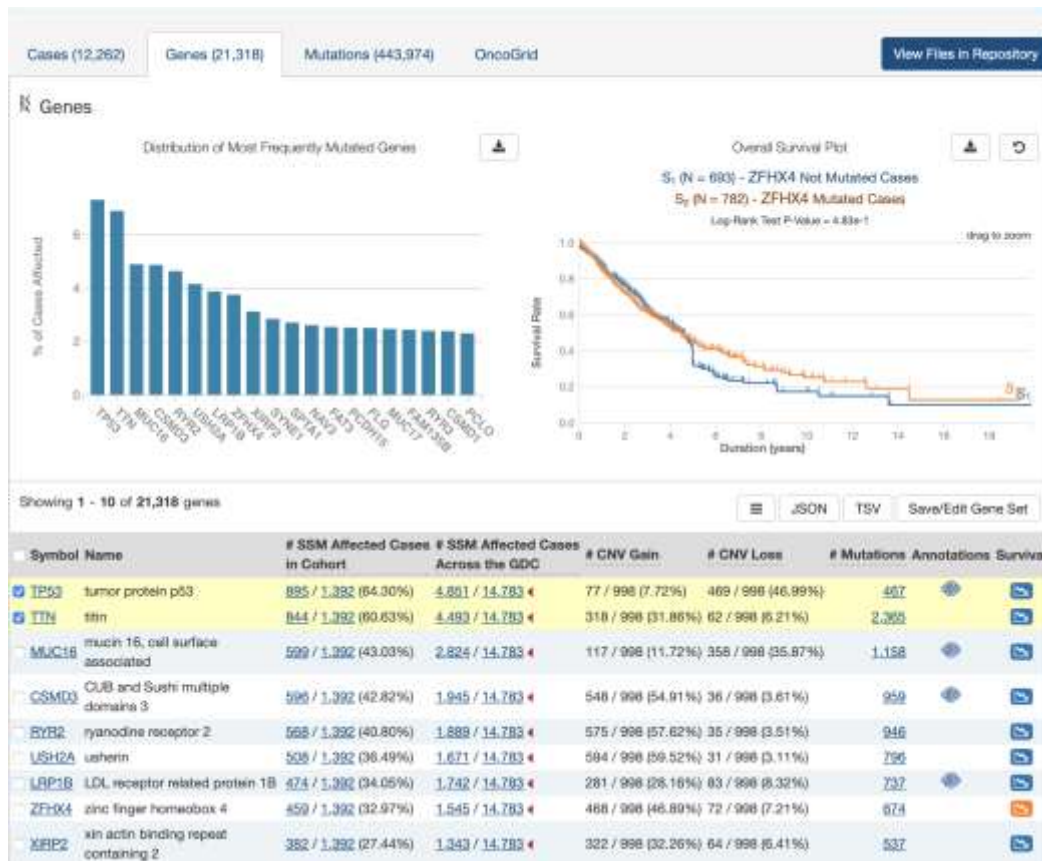
Genes view

Notice the case numbers



Genes view

Notice the case numbers

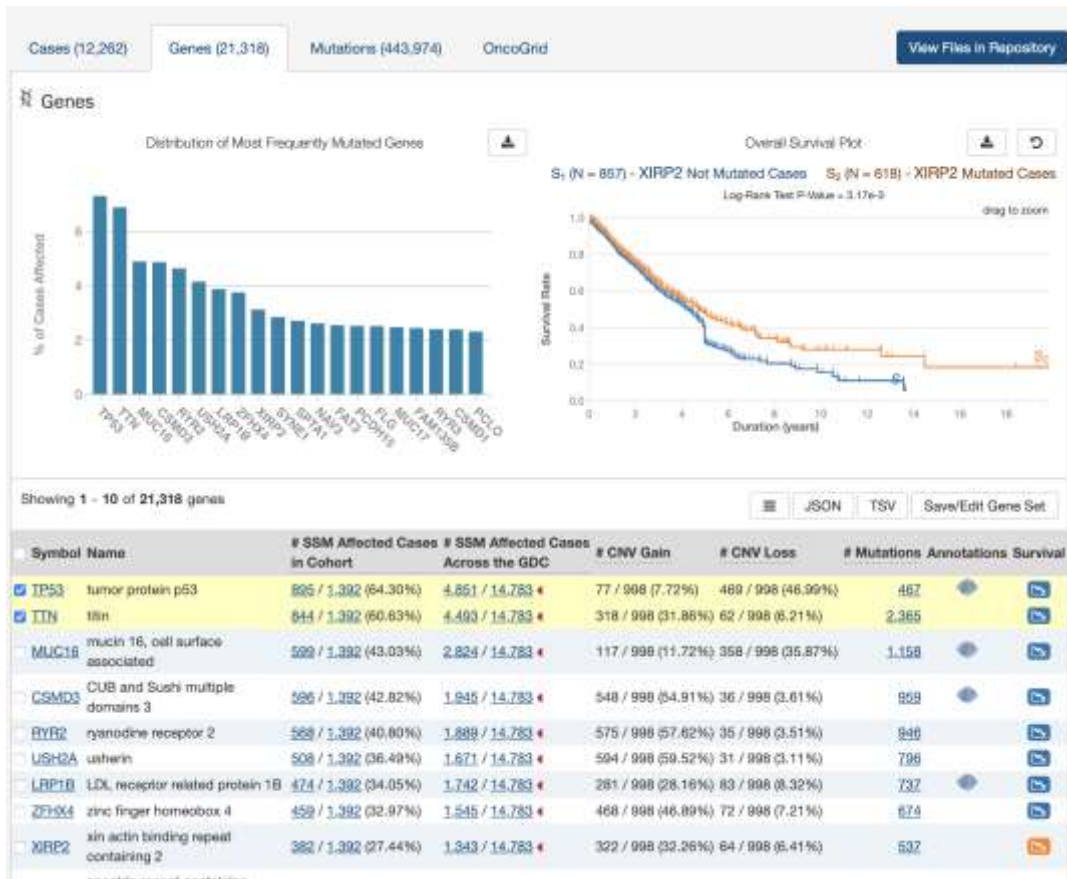


# Genomic Data Commons

XIRP2

Genes view

Notice the case numbers

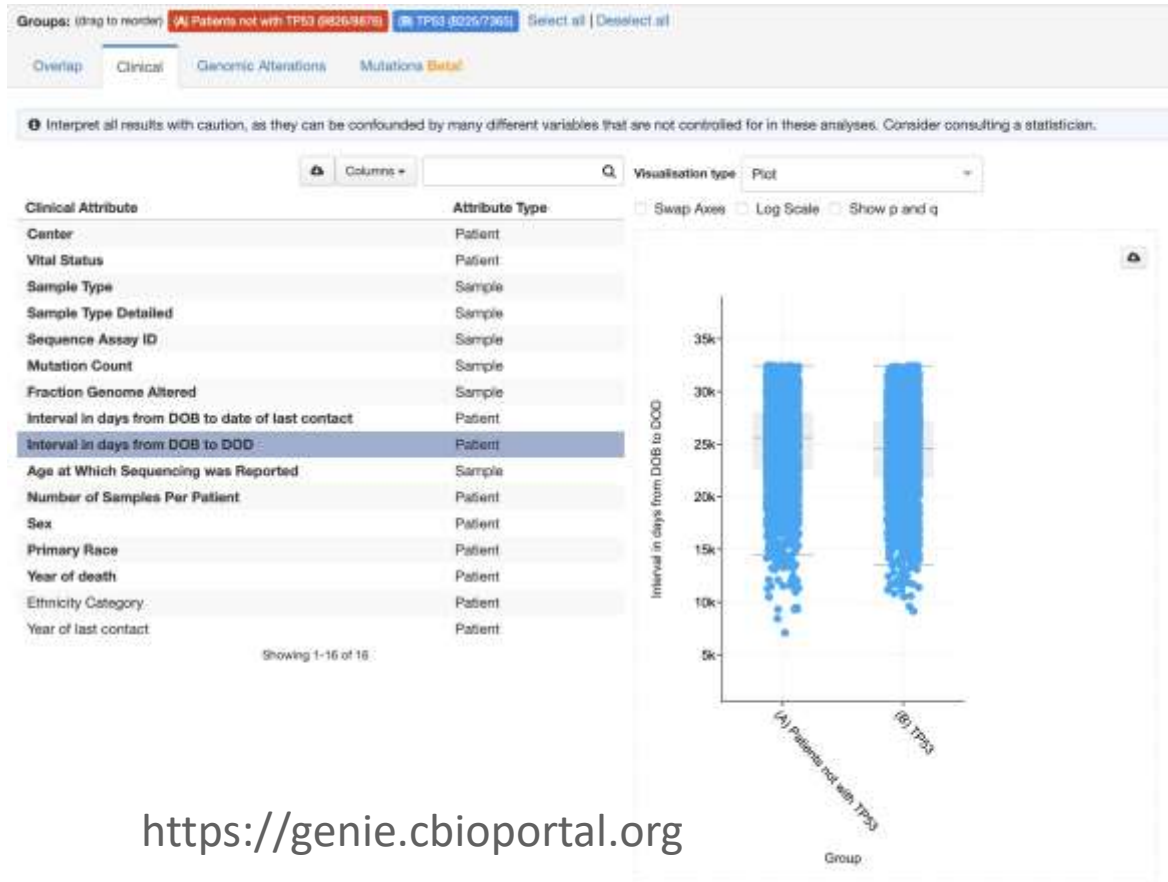


# Survival analysis out of the box

- The fact that mutations in all of those genes appears to convey a survival advantage is puzzling and contradicts the simple longevity curve we got from cBioPortal in GENIE
- Strange drop in survival at about 5 years
- Need a lot more analysis before believing the plot!



# Revisit: AACR Project GENIE cBioPortal resource



Not a survival analysis!

Looks like patients with somatic TP53 mutations die younger

Many potential confounders!

<https://genie.cbioportal.org>



# Survival analysis out of the box

- The fact that mutations in all of those genes appears to convey a survival advantage is puzzling and contradicts the simple longevity curve we got from cBioPortal in GENIE
- There are multiple potential explanations
  - TP53 mutations arise in younger patients and their survival with disease is longer (maybe?)
  - There are one or more confounders in the TCGA data, such as a skew toward late stage disease for patients with TP53 mutations (and the other mutations as well!) (definitely!)
  - Patients are from non-comparable cohorts (probably)
  - Other confounders you can think of?

# Survival analysis out of the box

- These problems and having consistent data entry from all the cases is why clinical trials are the standard for interpreting Progression Free Survival and performing survival analysis
- This is a taste of some of the issues in interpreting real world data – knowing that the cases and controls are balanced for known confounders
- Knowing which questions are appropriate for a given dataset is critical

# Confounders



A story about lung cancer and exercise



Sometimes we overlook the obvious



Don't forget to include known confounders!

# Background on TCGA and GENIE

- The Cancer Genome Atlas (TCGA) is a long running study of all comers across many institutions, however specific institutions contributed specimens and data predominantly from one disease site
- TCGA is biased toward larger tumors and stage 3
- TCGA was research sequencing, so not with therapeutic intent
- TCGA is primarily research whole exome sequencing
- GENIE is also an 'all comers' study. Sequence comes from a variety of in house and commercial clinical sequencing platforms. Primarily targeted sequencing. Based on perceived clinical benefit for the patient

# More on TCGA and GENIE

- Follow-up, therapeutic lines of treatment, PFS, recurrence are not consistently captured
- Diagnosis and stage at specimen collection are fairly consistent

# Good questions for TCGA and GENIE

- Prevalence of mutations in a disease area
- Mutational frequency at time of diagnosis
- Mutational patterns at time of diagnosis
- Age at diagnosis

# Beyond TCGA and GENIE



The next generation repository will need to carefully capture lines of therapy, multiple disease measures, tumor evolution, and more detailed biomarker measurements like scRNAseq, proteomics, metabolomics, tumor microenvironment



Deep characterization of normal tissues, The Human BioMolecular Atlas Program (HuBMAP) <https://commonfund.nih.gov/hubmap>. For cancer progression from precancerous lesions through to advanced disease The Human Tumor Atlas Network (HTAN) <https://humantumoratlas.org> HuBMAP and HTAN are laying the foundation for the next stage of repositories

# HuBMAP <https://portal.hubmapconsortium.org>

## Human BioMolecular Atlas Program

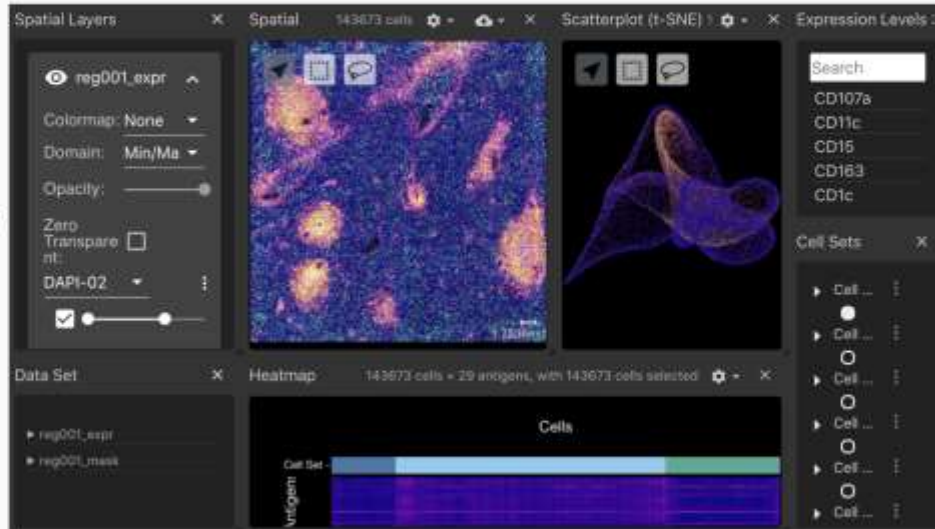
An open, global atlas of the human body at the cellular level

The HuBMAP Data Portal is the central resource for discovery, visualization, and download of single-cell tissue data generated by the consortium. A standardized data curation and processing workflow ensure that only high quality is released.

### Explore spatial single-cell data with Vitessce visualizations

View multi-modal single-cell resolution measurements with reusable interactive components such as a scatterplot, spatial+imaging plot, genome browser tracks, statistical plots, and controller components.

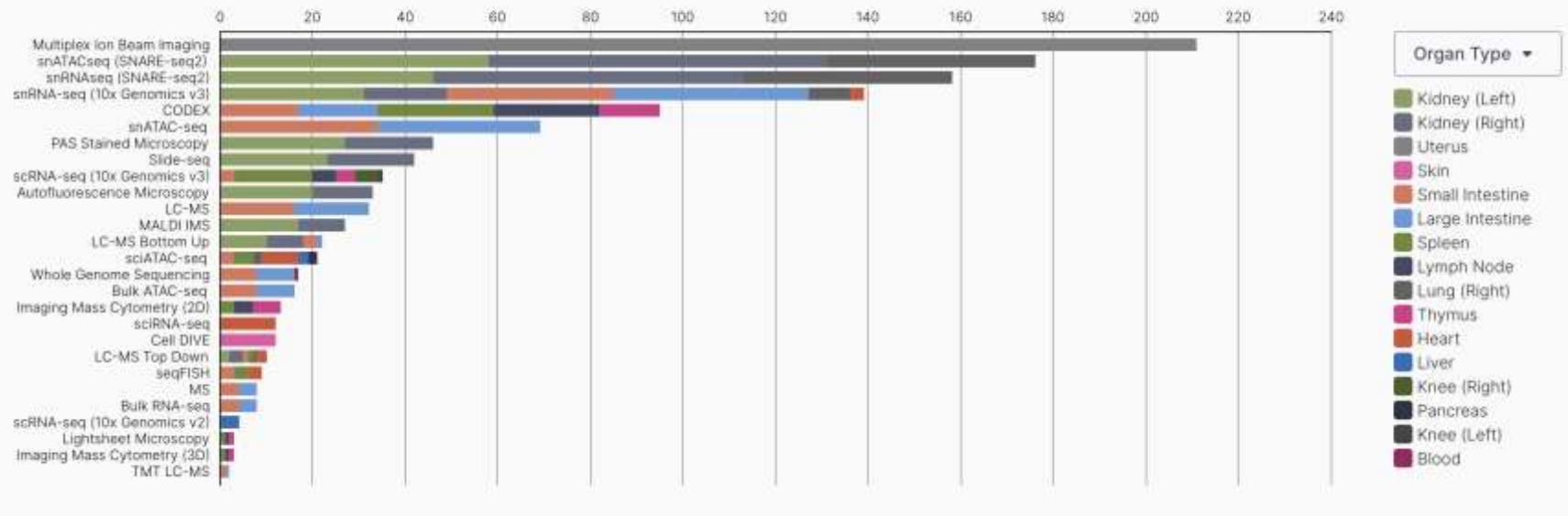
Get Started





# HuBMAP <https://portal.hubmapconsortium.org>

## HuBMAP Datasets



# HTAN <https://humantumoratlas.org>

**HTAN** Human Tumor Atlas Network

EXPLORE ANALYSIS TOOLS MANUAL [ABOUT THE DATA](#) [ABOUT HTAN](#) [SUBMIT DATA](#) [SUPPORT](#) [NEWS](#)

## Human Tumor Atlas Network

HTAN is a National Cancer Institute (NCI)-funded Cancer Moonshot<sup>SM</sup> initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease. (*Cell* April 2020)

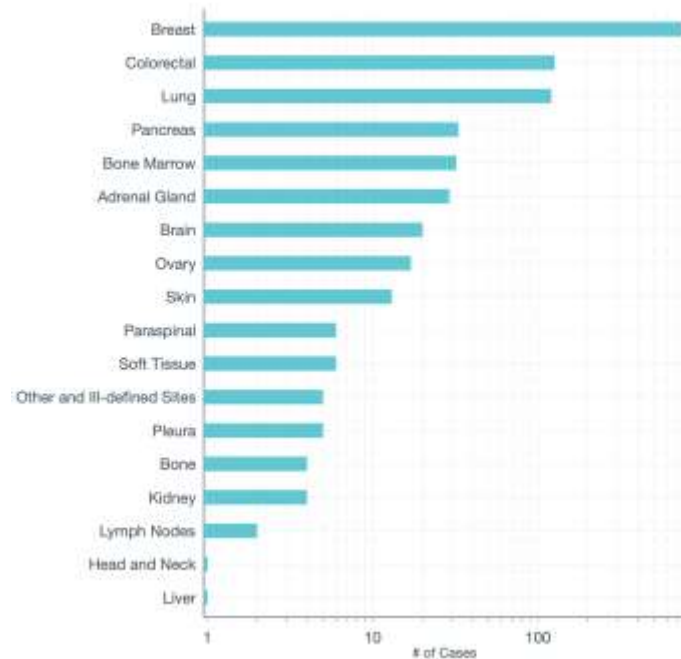
[Explore latest Data](#) [Learn more about HTAN](#)

Data Release V3 (Last updated 2023-04-14)

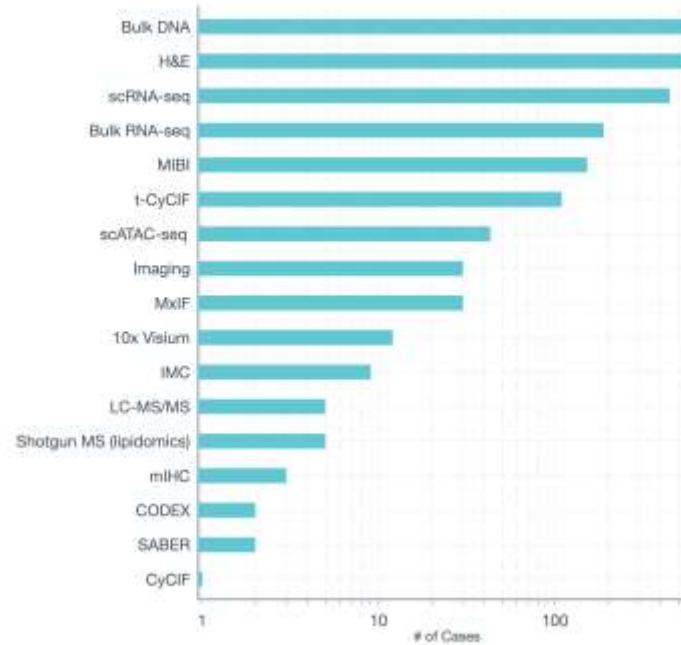
11	50	1311	3750
Atlases	Organs	Cases	Biospecimens

# HTAN <https://humantumoratlas.org>

The latest HTAN data release includes tumors originating from **18** primary tumor sites:



The tumors were profiled with **17** different types of assays:



# Data Access



For the example analyses, we used the public (open access) tier of data from Project GENIE and TCGA.



For detailed sequence and clinical data you generally need to have restricted access




For NIH studies, restricted access data requires submitting a dbGaP access request

An official website of the United States government [Check how you know](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

dbGaP   [Limits](#) [Advanced](#) [Help](#)



**dbGaP**  
The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

**Access dbGaP Data**

- [Advanced Search](#)
- [Controlled Access Data](#)
- [Public FTP Download](#)
- [Collections](#)
- [Summary Statistics](#)

**Resources**

- [dbGaP Data Browser](#)
- [Phenotype-Genotype Integrator](#)
- [dbGaP RSS Feed](#)
- [Software](#)

**Important Links**

- [How to Submit](#)
- [FAQ](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

**Latest Studies**

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
<a href="#">phs02172.v1.p1</a> Gabriela Miller Kids First Pediatric Research Project in Microbiota in Hispanic Populations	Version 1: passed embargo	<a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">P</a>	403	Cohort, Parent-Offspring Trios	<a href="#">Link</a>	
<a href="#">phs02162.v1.p1</a> Kids First: Genetics of Kidney and Urinary Tract Malformations	Version 1: passed embargo	<a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">P</a>	147	Cohort, Parent-Offspring Trios	<a href="#">Link</a>	
<a href="#">phs02161.v1.p1</a> Kids First: Genomic Analysis of Esophageal Atresia and Tracheoesophageal Fistulas and Associated Congenital Anomalies	Version 1: passed embargo	<a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">P</a>	410	Cohort, Parent-Offspring Trios	<a href="#">Link</a>	
<a href="#">phs02130.v1.p1</a> Gabriela Miller Kids First Pediatric Research Program in Craniofacial Microsoma	Version 1: passed embargo	<a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">P</a>	278	Cohort, Parent-Offspring Trios	<a href="#">Link</a>	
<a href="#">phs01846.v1.p1</a> Kids First: The Intersection of Childhood Cancer and Birth Defects	Version 1:	<a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">P</a>	1805	Cohort, Parent-Offspring Trios	<a href="#">Link</a>	<a href="#">HDSeg K Test</a>

[List Top Level Studies](#)

# dbGaP – Access to Kids First data



## Gabriella Miller Kids First Pediatric Research Program in Genetics at the Intersection of Childhood Cancer and Birth Defects

dbGaP Study Accession: phs001846.v1.p1

[Request Access](#)

[Subject Sample Telemetry Report \(SSTR\)](#)

[Study version history](#)

[Study](#) | [Phenotype Datasets](#) | [Variables](#) | [Molecular Datasets](#) | [Analyses](#) | [Documents](#)

Jump to: [Authorized Access](#) | [Attribution](#) | [Authorized Requests](#)

### Study Description

The [Gabriella Miller Kids First Pediatric Research Program](#) (Kids First) is a trans-NIH effort initiated in response to the [2014 Gabriella Miller Kids First Research Act](#) and supported by the NIH Common Fund. This program focuses on gene discovery in pediatric cancers and structural birth defects and the development of the Gabriella Miller Kids First Pediatric Data Resource (Kids First Data Resource). All of the genomic and phenotypic data from this study are accessible through dbGaP. The data is also available at the [Kids First Portal](#), where other Kids First datasets can also be accessed in the cloud for data analysis, data visualization, collaboration and interoperability, open to all researchers and developers.

Birth defects and childhood cancer share biological pathways that are important for cell growth and division. We propose that sequencing pediatric patients suffering both conditions will allow us to discover the underlying genes and in turn advance our understanding of the causes of these devastating diseases.

- Study Weblinks:
  - [GDC](#)
- Study Design:
  - Family/Twin/Trios
- Study Type:
  - Cohort
  - Parent-Offspring Trios
- Total number of consented subjects: 1805
- [Subject Sample Telemetry Report \(SSTR\)](#)

### Important Links and Information

- Request access via [Authorized Access](#)
  - [Instructions for requestors](#)
  - [Data Use Certification \(DUC\) Agreement](#)
- [Talking Glossary of Genetic Terms](#)

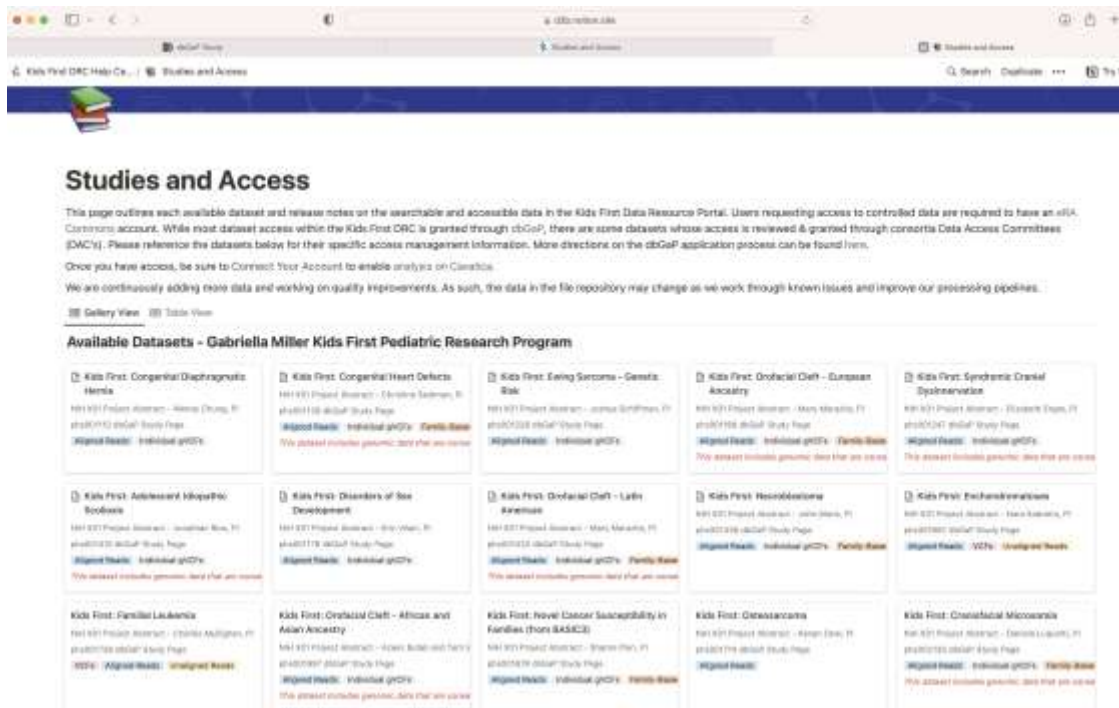
### Authorized Access

- **Data access provided by:** [dbGaP Authorized Access](#)
- **Release Date:** December 21, 2020
- **Embargo Release Date:** No embargo
- [Data Use Certification Requirements \(DUC\)](#)
- **Public Posting of Genomic Summary Results:** Not Applicable
- **Use Restrictions**

Consent group	Is IRB required?	Data Access Committee	Number of participants
General Research Use	No	Kids First DAC ( <a href="mailto:kidsfirstdac@mail.nih.gov">kidsfirstdac@mail.nih.gov</a> )	1805

- [List of components](#) downloadable from [Authorized Access](#)

# An auxiliary way to look at dbGaP projects



The screenshot shows a web browser window displaying the 'Studies and Access' page of the Kids First Data Resource Portal. The page title is 'Studies and Access' and it includes a search bar and navigation options. Below the header, there is a section titled 'Available Datasets - Gabriella Miller Kids First Pediatric Research Program'. This section contains a grid of 15 dataset cards, each with a title, a brief description, and a link to the 'Study Page'. The datasets listed include:

- Kids First: Congenital Diaphragmatic Hernia
- Kids First: Congenital Heart Defects
- Kids First: Ewing Sarcoma - Genetic Risk
- Kids First: Orofacial Cleft - European Ancestry
- Kids First: Syndromic Cranial Dysmaturaton
- Kids First: Attention Deficit/Hyperactivity Disorders
- Kids First: Disorders of Sex Development
- Kids First: Orofacial Cleft - Latin American
- Kids First: Neurofibromatosis
- Kids First: Exchondromatosis
- Kids First: Familial Leukemia
- Kids First: Orofacial Cleft - African and Asian Ancestry
- Kids First: Novel Cancer Susceptibility in Families (from BASK2)
- Kids First: Osteosarcoma
- Kids First: Osteosarcoma
- Kids First: Osteosarcoma

Each card also indicates the 'Request Status' (e.g., 'Individual gPCR', 'Family Based') and a note about whether the dataset includes genomic data that can be accessed.

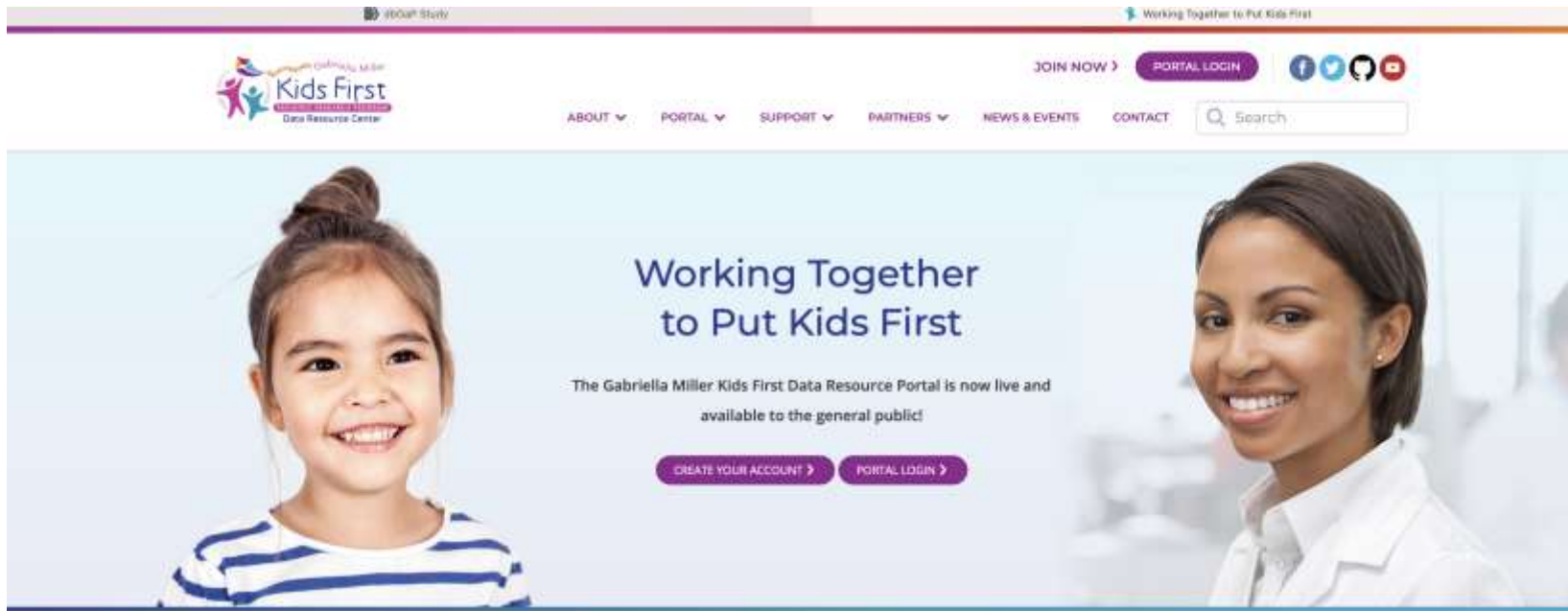
Also doesn't help see variable names, but does list file types

## dbGaP and access to metadata

- With the exception of seeing the number of participants, things like what types of experimental variables (targeted sequencing, whole exome, whole genome, RNAseq, scRNAseq, methylSeq, etc), clinical variables (diagnosis, stage, treatments, adverse events), demographic variables (age, race, ethnicity, urban vs rural, etc) are not available for many projects until you have access to the full dataset.
- This makes it hard to identify which specific datasets you might be interested in



# To understand Kids First, go to it First!



The screenshot shows the homepage of the Kids First Data Resource Center. At the top, there is a navigation bar with the text "Working Together to Put Kids First" on the right. The main header features the Kids First logo on the left, which includes the text "Gabriella Miller Kids First Data Resource Center". To the right of the logo are navigation links: "JOIN NOW >", "PORTAL LOGIN", and social media icons for Facebook, Twitter, LinkedIn, and YouTube. Below the navigation is a horizontal menu with links for "ABOUT", "PORTAL", "SUPPORT", "PARTNERS", "NEWS & EVENTS", and "CONTACT", followed by a search bar. The main content area features a large banner with a smiling young girl on the left and a smiling woman in a white lab coat on the right. The central text reads "Working Together to Put Kids First" and "The Gabriella Miller Kids First Data Resource Portal is now live and available to the general public!". Below this text are two buttons: "CREATE YOUR ACCOUNT >" and "PORTAL LOGIN >".

## Explore and Connect with Research Data Today!

The NIH Common Fund-supported Gabriella Miller Kids First Data Resource Center enables researchers, clinicians, and patients to work together to accelerate research and promote new discoveries for children affected with cancer and structural birth defects. Data from over 11,000 samples, including DNA and RNA, is available to empower your research today. Data collected from more than 30,000 samples are expected to be available in the next few years. Learn how to get started using the Data Resource Portal today!

[CLICK HERE TO GET STARTED >](#)

# Some stats about Kids First, even before creating a login

The screenshot shows the homepage of the Kids First Data Resource Center. At the top left is the logo for Kids First Data Resource Center. The navigation menu includes links for ABOUT, PORTAL, SUPPORT, PARTNERS, NEWS & EVENTS, and CONTACT. There are also buttons for JOIN NOW and DOWNLOAD, and social media icons for Facebook, Twitter, and YouTube. A search bar is located on the right side of the navigation bar.

## Explore and Connect with Research Data Today!

The NH Common Fund-supported Gabriella Miller Kids First Data Resource Center enables researchers, clinicians, and patients to work together to accelerate research and promote new discoveries for children affected with cancer and structural birth defects. Data from over 11,000 samples, including DNA and RNA, is available to empower your research today. Data collected from more than 90,000 samples are expected to be available in the next few years. Learn how to get started using the Data Resource Portal today!

[Click here to get started!](#)





### Available Data

Note: Due to technical limitations, these counts currently exclude TADCT data, & Sequencing data, which may also be only included in the portal. The genomic data shown are available on the GDC.

36 Studies	33,402 Participants	30,910 Families	98,833 Samples	199,161 Files	1.8 PB Size
------------	---------------------	-----------------	----------------	---------------	-------------

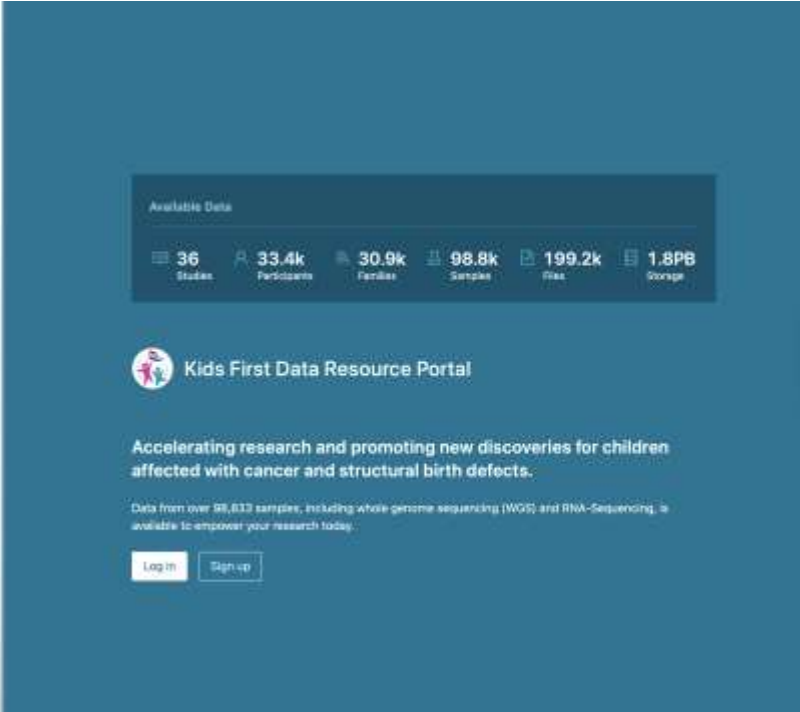

[More About The Datasets!](#)

### Our Partners

 <b>Researchers</b> Search, view, analyze, and	 <b>Healthcare Professionals</b>	 <b>Patients/Family Members</b>	 <b>Community Members</b>
---------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------


<https://kidsfirstdrc.org>

# Kids First login or sign up



Available Data

36 Studies	33.4k Participants	30.9k Families	98.8k Samples	199.2k Files	1.8PB Storage
---------------	-----------------------	-------------------	------------------	-----------------	------------------

 Kids First Data Resource Portal

Accelerating research and promoting new discoveries for children affected with cancer and structural birth defects.

Data from over 98,833 samples, including whole genome sequencing (WGS) and RNA-Sequencing, is available to empower your research today.

[Log in](#) [Sign up](#)

# You do have to accept the terms & conditions



## Terms & Conditions

### Terms & Conditions

Last Update Date: 11/22/2021

As a user of the Services you agree that you are 13 years of age or older and furthermore agree to the Terms and Conditions of Services defined herein and where applicable the terms defined by the [NIH Genetic Data User Code of Conduct](#). These terms include, but are not limited to:

1. You will request controlled-access datasets solely in connection with the research project described in an approved Data Access Request for each dataset;
2. You will make no attempt to identify or contact individual participants or groups from whom data were collected, or generate information that could allow participants' identities to be readily ascertained;
3. You will not distribute controlled-access datasets to any entity or individual beyond those specified in an approved Data Access Request;
4. You will adhere to computer security practices in compliance with [NIH Security Best Practices for Controlled-Access Data](#), such that only authorized individuals possess access to data files;
5. You acknowledge Intellectual Property Policies should they exist as specified in a dataset's associated Data Use Certification; and,
6. You will report any inadvertent data release in accordance with the terms in the Data Use Certification, breach of data security, or other data management incidents contrary to the terms of data access.

Decline

✓ Accept

# You can see the studies

The screenshot shows the 'Studies' page in the Kids First portal. The left sidebar contains filters for Search Studies, Domain, Program, Family Data, and Data Categories. The main content area shows a search bar and a table of 36 studies. The table columns include Code, Name, Program, Domain, dbGap, Participants, Families, and Available participants per Data Category (Seq, Str, Crv, Exp, Sv, Pat, Rad, Other, Files).

Code	Name	Program	Domain	dbGap	Participants	Available participants per Data Category											
						Families	Seq	Str	Crv	Exp	Sv	Pat	Rad	Other	Files		
TARGET:AML	TARGET: Acute Myeloid Leukemia	TARGET	Cancer	dbG00985	341	341	351	203									1152
KE-CHARGE	Kids First: CHARGE	Kids First	Birth Defect	dbG022982	617	217	617	617									386 1288
KE-DEGMA	Kids First: Orofacial Cleft, Alveolar and Alveolar Anomaly	Kids First	Birth Defect	dbG021987	759	253	715	715									714 2661
KE-NBL	Kids First: Neuroblastoma	Kids First	Cancer	dbG021428	1681	609	1612	1687	336	208							12 16726
KE-MMC	Kids First: Myeloid Malignancies	Kids First	Cancer	dbG021587	456	8	408	408		390							4362
KE-EATE	Kids First: Esophageal Atresia and Tracheoesophageal Fistulae	Kids First	Birth Defect	dbG021381	834	128	934	833									747
KE-DSD	Kids First: Disorders of Sex Development	Kids First	Birth Defect	dbG021328	183	60	183	182									602
KE-EASD	Kids First: Fetal Alcohol Spectrum Disorders	Kids First	Birth Defect	dbG022986	90	34	00	60									82 332
TARGET:NB	TARGET: Neuroblastoma	TARGET	Cancer	dbG020487	277	277	277	216									2338

# Drill down on the childhood cancer studies

**Search Studies**

**Domain**

Select All / None

- Cancer 18
- Birth Defect 77
- COVID-19 8

**Program**

Select All / None

- Kids First 70
- Pediatric Brain Tumor Atlas 3
- TARGET 2
- ICB 4

**Family Data**

- True 16
- None 8

**Data Categories**

Select All / None

- Sequencing Reads 18
- Simple Nucleotide Variation 14
- Other 8

**Studies**

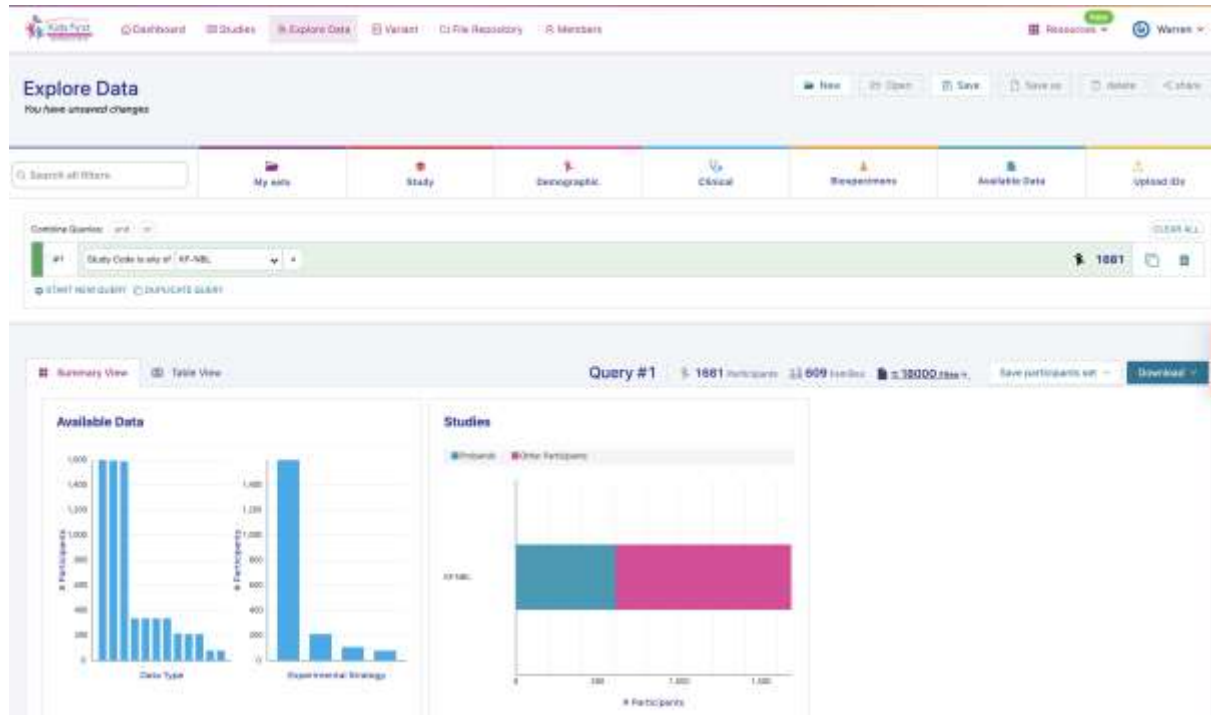
Domain: Cancer

New query

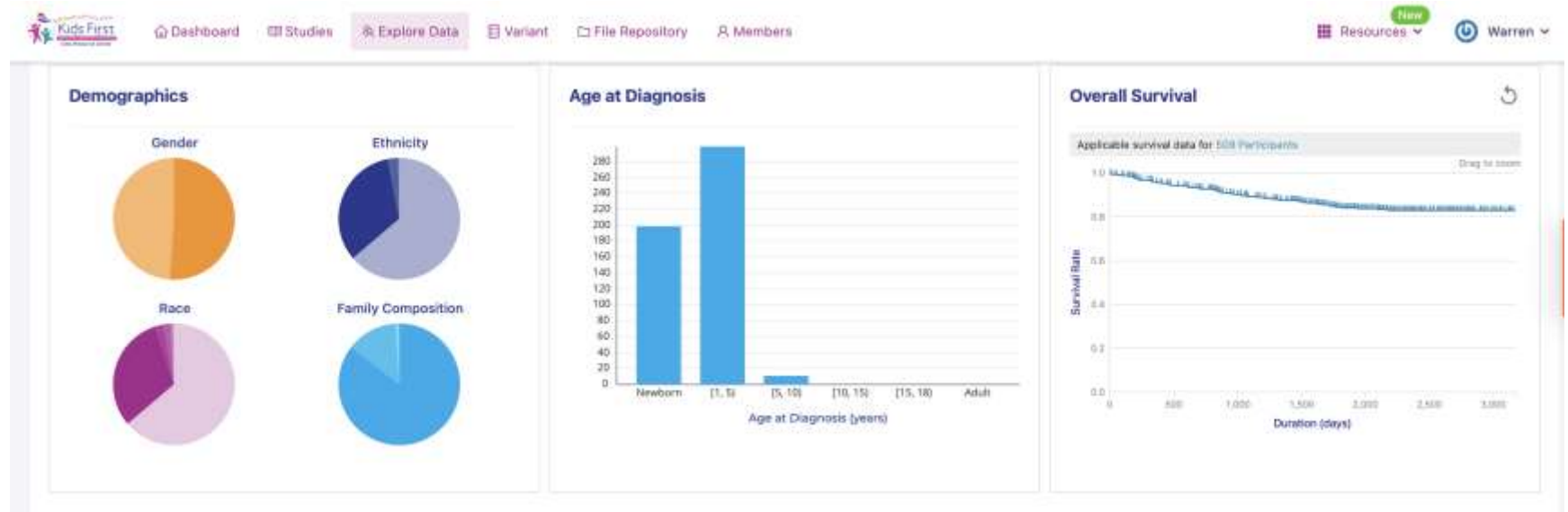
Showing 15 studies

Code	Name	Program	Domain	dbGap	Participants	Available participants per Data Category										
						Families	Seq	Sm	Crv	Exp	Sv	Pal	Rad	Other	Files	
TARGET-AML	TARGET: Acute Myeloid Leukemia	TARGET	Cancer	db002488	341	341	321	200								1102
KF-DBL	Kids First: Neuroblastoma	Kids First	Cancer	db001838	1681	608	1812	1507	336	208				12		18726
KF-SMC	Kids First: Myeloid Malignancies	Kids First	Cancer	db002187	408	0	408	408		300						4362
TARGET-NB	TARGET: Neuroblastoma	TARGET	Cancer	db000482	277	277	277	219								2388
KF-CHDALL	Kids First: Leukemia & Heart Defects in Down Syndrome	Kids First	Cancer, Birth Defect	db002230	2129	1703	2037	1896	207	400				492		11149
KF-NCFP	Kids First: Novel Cancer Susceptibility in Families (From NACFS)	Kids First	Cancer	db001002	718	246	276	276						174		2275
KF-ED	Kids First: Enchondromatosis	Kids First	Cancer	db001982	211	76	211	211						187		1018
PITA-PMOC	Pediatric Brain Tumor Atlas: PMOC	Pediatric Brain Tumor Atlas	Cancer		83	0	82	88		30				32		18344
KF-TALL	Kids First: T Cell ALL	Kids First	Cancer	db002228	1358	0	1358	1358	875	1043						38976

# Select Neuroblastoma – 1681 participants



# Quick visual breakdown, including survival analysis





# Looking at variants, more than 280 million of them!

The screenshot displays the 'Variants Exploration' interface. At the top, there is a navigation bar with 'Variant' selected. Below this, a search bar contains the text 'Use the filters to build a query' and a '281M' indicator. A '+ New query' button is located below the search bar. The main content area shows a table of variants with the following columns: Variant, Type, dbSnp, Consequences, CLINVAR, Studies, Part., Freq., ALT, and Hom. The table lists several variants, including deletions and SNVs on chromosome 10.

Variant	Type	dbSnp	Consequences	CLINVAR	Studies	Part.	Freq.	ALT	Hom
chr10:g.100048883del	deletion	—	▲ frameshift_variant CPN1 Y576X	—	1	1/11030	9.07e-5	1	0
chr10:g.100063863G>T	SNV	—	▲ stop_gained CPN1 Y274*	—	1	2/11030	1.81e-4	2	0
chr10:g.100063879C>T	SNV	—	▲ stop_gained CPN1 W269*	—	1	2/11030	1.81e-4	2	0
chr10:g.100065182C>T	SNV	rs728917122	▲ splice_donor_variant CPN1	—	1	1/11030	9.07e-5	1	0
chr10:g.100075058del	deletion	—	▲ frameshift_variant CPN1 I125X	—	1	1/11030	9.07e-5	1	0
chr10:g.100075972del	deletion	—	▲ frameshift_variant CPN1 D120X	—	1	1/11030	9.07e-5	1	0
chr10:g.100073876G>A	SNV	rs764408494	▲ stop_gained CPN1 Q119*	—	2	3/11030	2.72e-4	3	0
chr10:g.100075889del	deletion	—	▲ frameshift_variant CPN1 I154X	—	1	1/11030	9.07e-5	1	0
chr10:g.100081511G>A	SNV	rs772979803	▲ stop_gained CPN1 Q39*	—	2	3/11030	2.72e-4	3	0

# Select gene, then enter TP53

Search by Gene

TP53

- OR TP53 ENSG00000141510
- OR TP53AIP1 ENSG00000120471
- OR TP53BP1 ENSG00000067369
- OR TP53BP2 ENSG00000143014
- OR TP53BP2P1

OMIM

> ODD

### Variants Exploration

Variant Queries

Gene Symbol = TP53 3,453

+ New query

Showing 1 - 20 out of 3,453

Variant	Type	dbSNP	Consequences	CLINVAR	Status	Part.	Freq.
chr17:g.7867426>G	SNV	rs375161249	▲ splice_donor_variant TP53	—	3	8 / 11030	7.25e-4
chr17:g.7867265del	deletion	—	▲ frameshift_variant TP53 Y387X	—	2	3 / 11030	2.73e-4
chr17:g.7867261A>T	SNV	rs23979530	▲ stop_gained TP53 L389*	—	15	880 / 11030	7.98e-2
chr17:g.7867273del	deletion	—	▲ frameshift_variant TP53 L383X	—	15	2282 / 11030	8.71e-1
chr17:g.7867426C>T	SNV	rs1011101821	▲ splice_acceptor_variant TP53	—	1	1 / 11030	8.07e-5
chr17:g.7872611_78737	deletion	—	▲ splice_donor_variant TP53	—	1	1 / 11030	9.07e-5
chr17:g.7873785del	deletion	—	▲ frameshift_variant TP53 P276X	—	1	1 / 11030	9.07e-5

# Now we see the variants for TP53

The screenshot displays the Variant Explorer interface. The top navigation bar includes 'Fish First', 'Dashboard', 'Studies', 'Explore Data', 'Variant', 'File Repository', and 'Members'. On the right, there are 'Resources' and 'Warren' icons. The left sidebar contains a menu with 'Variant', 'Gene', 'Pathogenicity', 'Frequency', and 'Occurrence'. The main content area is titled 'Variants Exploration' and shows a search for 'TP53'. A filter for 'Gene Symbol' is set to 'TP53', resulting in 3,453 variants. A table lists the first 20 variants, including their dbSNP IDs, types, consequences, and associated studies.

Variant	Type	dbSNP	Consequences	CLNVAR	Studies	Part.	Freq.
chr17:g.7566900A>G	SNV	rs375181789	splice_donor_variant TP53	--	5	8 J 11030 4	7.25e
chr17:g.7567258del	deletion	--	frameshift_variant TP53 Y387X	--	2	3 J 11030 4	2.72e
chr17:g.7567261A>T	SNV	rs73979530	stop_gained TP53 L385*	--	15	88 J 11030 2	7.66e
chr17:g.7567272del	deletion	--	frameshift_variant TP53 L383X	--	15	238 J 11030 1	6.71e
chr17:g.7567426G>T	SNV	rs1013101821	splice_acceptor_variant TP53	--	1	1 J 11030 5	9.07e
chr17:g.7573611_75737--	deletion	--	splice_donor_variant TP53	--	1	1 J 11030 5	9.07e
chr17:g.7573788del	deletion	--	frameshift_variant TP53 P278X	--	1	1 J 11030 5	9.07e

# Now select for pathogenic variants – 6 of them

The screenshot shows the Variant Explorer interface. The top navigation bar includes 'Home', 'Dashboard', 'Studies', 'Explore Data', 'Variant', 'File Repository', and 'Members'. The user is logged in as 'Warren'. The main content area is titled 'Variants Exploration' and shows a query: 'Gene Symbol = TP53 and COSMIC = ClinVar = Pathogenic'. The results table shows 6 pathogenic variants in the TP53 gene.

Variant	Type	dbSnp	Consequences	CLINVAR	Studies	Part.	Freq.	ALT	Home
chr17:p.7670790G>A	SNV	rs582782528	missense_variant TP53 R337C	Pathogenic, Likely_pathogenic	1	3 / 11030	2.72e-4	3	0
chr17:p.7674230C>T	SNV	rs11542652	missense_variant TP53 R248Q	Pathogenic	1	1 / 11030	9.07e-5	1	0
chr17:p.7674230C>T	SNV	rs28834375	missense_variant TP53 G245S	Pathogenic	1	1 / 11030	9.07e-5	1	0
chr17:p.7675139C>T	SNV	rs582782144	missense_variant TP53 R158H	Pathogenic, Likely_pathogenic	2	3 / 11030	2.72e-4	3	0
chr17:p.7675895G>C	SNV	rs286201057	missense_variant TP53 T125R	Pathogenic, Likely_pathogenic	1	1 / 11030	9.07e-5	1	0
chr17:p.7678040C>A	SNV	rs11542654	missense_variant TP53 R110L	Pathogenic, Likely_pathogenic	1	1 / 11030	9.07e-5	1	0

# Add likely pathogenic – now we have 8 variants

The screenshot displays the Variant Explorer interface. On the left, a sidebar shows navigation options: Variant, Gene, Pathogenicity, Frequency, and Occurrence. The Pathogenicity section is expanded, showing a list of filters with counts:

- Likely Pathogenic: 8
- Pathogenic: 8
- Likely Benign: 76
- Benign: 69
- Uncertain Significance: 36
- Conflicting Interpretations Of Pathogenicity: 18
- Not Provided: 2
- Drug Response: 14

Below the filters are sections for VEP, SIFT, Polyphen2 HVAR, FAT1MM, CADD, DANM, and LRT. The main panel, titled 'Variants Exploration', shows a query bar with filters: 'Bicus Syndrome', 'TP53', 'd[3]', 'COSMIC', 'Glossa', 'd[2]', 'Clinvar', 'Pathogenic', and 'Likely Pathogenic'. A '+ New Query' button is visible. Below the query bar, a table displays 8 variants, all with 'Likely Pathogenic' annotations:

Variant	Type	dbSnp	Consequences	CUNVAR	Studies	Part.	Freq.	ALT	Hom
chr17:g.7670097C>T	SNV	--	↑ stop_gained TP53 W91*	Likely_pathogenic	1	1/11030	9.07e-5	1	0
chr17:g.7670700G>A	SNV	rs567781529	★ missense_variant TP53 R137C	Pathogenic, Likely_pathogenic	1	3/11030	2.72e-4	3	0
chr17:g.7674120C>T	SNV	rs11540862	★ missense_variant TP53 R249Q	Pathogenic	1	1/11030	9.07e-5	1	0
chr17:g.7674135C>T	SNV	rs28924525	★ missense_variant TP53 G145S	Pathogenic	1	1/11030	9.07e-5	1	0
chr17:g.7675139C>T	SNV	rs567781144	★ missense_variant TP53 R158H	Pathogenic, Likely_pathogenic	2	3/11030	2.72e-4	3	0
chr17:g.7675995G>C	SNV	rs788201047	★ missense_variant TP53 T125R	Pathogenic, Likely_pathogenic	1	1/11030	9.07e-5	1	0
chr17:g.7676040C>A	SNV	rs11540864	★ missense_variant TP53 R110I	Pathogenic, Likely_pathogenic	1	1/11030	9.07e-5	1	0
chr17:g.7679940C>G	SNV	rs5883838	▼ splice_region_variant TP53	Likely_pathogenic	1	1/11030	9.07e-5	1	0

TP53 is not druggable: Lets look at something clinically actionable



We are now going to start looking at EGFR.

EGFR T790M is a common resistance mutation for patients with NSLC getting Tyrosine Kinase Inhibitors (TKIs) like gefitinib and erlotinib, which are small molecule targeted therapies.

## NM\_005228.5(EGFR):c.2369C>T (p.Thr790Met)

Cite this record

**Interpretation:** drug response

**Review status:** ★★☆☆ reviewed by expert panel

**Submissions:** 13

**First in ClinVar:** Jan 31, 2015

**Most recent Submission:** Feb 7, 2023

**Last evaluated:** Mar 24, 2021

**Accession:** VCV000016613.24

**Variation ID:** 16613

**Description:** single nucleotide variant

### Variant details

Conditions

Gene(s)

### NM\_005228.5(EGFR):c.2369C>T (p.Thr790Met)

**Allele ID:** 31652

**Variant type:** single nucleotide variant

**Variant length:** 1 bp

**Cytogenetic location:** 7p11.2

**Genomic location:** 7: 55181378 (GRCh38) [GRCh38](#) [UCSC](#)

7: 55249071 (GRCh37) [GRCh37](#) [UCSC](#)

**HGVS:**

Nucleotide	Protein	Molecular consequence
NM_005228.5:c.2369C>T <a href="#">MANE SELECT</a>	NP_005219.2:p.Thr790Met	missense
NM_001346897.2:c.2234C>T	NP_001333826.1:p.Thr745Met	missense
NM_001346898.2:c.2369C>T	NP_001333827.1:p.Thr790Met	missense

... more HGVS

**Protein change:** [T790M](#), [T745M](#), [T523M](#), [T737M](#)

**Protein change:** T790M, T745M, T523M, T737M  
**Other names:** NP\_005219.2:p.Thr790Met  
**Canonical SPDI:** NC\_000007.14:55181377:C:T  
**Functional consequence:** -  
**Global minor allele frequency (GMAF):** -  
**Allele frequency:** The Genome Aggregation Database (gnomAD) 0.00006  
 The Genome Aggregation Database (gnomAD) 0.00010  
 The Genome Aggregation Database (gnomAD), exomes 0.00003  
 Trans-Omics for Precision Medicine (TOPMed) 0.00003  
 Exome Aggregation Consortium (ExAC) 0.00004  
 Trans-Omics for Precision Medicine (TOPMed) 0.00002  
**Links:** PharmGKB Clinical Annotation: 981475450  
 dbSNP: rs121434569  
 ClinGen: CA090928  
 Genetic Testing Registry (GTR): GTR000575663  
 UniProtKB: P00533#VAR\_026098  
 OMIM: 131550.0006  
 VarSome

## Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	More information
drug response (Mar 24, 2021)	reviewed by expert panel (Pharmacogenomics knowledge for personalized medicine) Method: curation	gefitinib response - Efficacy Drug used for Carcinoma, Non-Small-Cell Lung, and Drug Resistance Affected status: yes Allele origin: germline	PharmGKB Accession: SCV002031219.1 First in ClinVar: Dec 12, 2021 Last updated: Dec 12, 2021 Comment: Drug is not necessarily used to treat response condition	Publications: PubMed (16) Other databases <a href="https://www.pharmgkb.org/variant...">https://www.pharmgkb.org/variant...</a> <a href="https://www.pharmgkb.org/clinical...">https://www.pharmgkb.org/clinical...</a> Comment: PharmGKB Level of Evidence 2B: Variants in Level 2B clinical annotations are not in PharmGKB's Tier 1 VIPs. These clinical annotations describe variant-drug combinations with ... (more)
drug response (Mar 24, 2021)	reviewed by expert panel (Pharmacogenomics knowledge for personalized medicine) Method: curation	erlotinib response - Efficacy Drug used for Adenocarcinoma, Carcinoma, Non-Small-Cell Lung, Drug Resistance, and Lung Neoplasms Affected status: yes Allele origin: germline	PharmGKB Accession: SCV000268172.4 First in ClinVar: May 22, 2016 Last updated: Dec 12, 2021 Comment: Drug is not necessarily used to treat response condition	Publications: PubMed (11) Other databases <a href="https://www.pharmgkb.org/variant...">https://www.pharmgkb.org/variant...</a> <a href="https://www.pharmgkb.org/clinical...">https://www.pharmgkb.org/clinical...</a> Comment: PharmGKB Level of Evidence 2B: Variants in Level 2B clinical annotations

Hard to interpret!



# The COSMIC Database

<https://cancer.sanger.ac.uk/cosmic>

Very different purpose – it is to catalog somatic mutations associated with cancer

# The COSMIC Database



The screenshot shows the COSMIC database homepage. At the top, there is a navigation bar with links for Projects, Data, Tools, News, Help, About, Genome version, a search bar, and a Login button. The main heading is "COSMIC v96, released 31-MAY-22". Below this, a brief description states that COSMIC is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. A search bar is provided with the example text "eg BRCA1, COLO 429, Carcinomas, WWOX, BRCA1K, G536H1".

**Projects**

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- COSMIC**: The core of COSMIC, an expert-curated database of somatic mutations.
- Cell Lines Project**: Mutation profiles of over 1,000 cell lines used in cancer research.
- COSMIC-3D**: An interactive view of cancer mutations in the context of 3D structures.
- Cancer Gene Census**: A catalogue of genes with mutations that are causally implicated in cancer.
- Cancer Mutation Census**: Classification of genetic variants driving cancer.
- Actionability**: Mutations actionable in protein oncology.

**COSMIC News**

- From diagnosis to remission- Haematological cancer research to look out for in v97**: Here's a sneak peek into three papers curated as part of our v97 update, each highlighting unique areas of research improving the outcomes for haematological cancer patients. [Here...](#)
- Copy Number Signatures: A scalable research and clinical tool?**: Copy Number Signatures are the latest addition to COSMIC Mutational Signatures, we caught up with one of the leads, Dr Ludmil Alexandrov, to discuss the key findings, utility of the data, and hopes for the future. [Here...](#)
- What are the emerging trends in cancer research? Our five key takeaways from AACR-2022**: Read about the five emerging trends we took away from our time at AACR-2022 [Here...](#)

**Tools**

- Cancer Browser** - browse COSMIC data by tissue type and histology
- Genome Browser** - browse the human genome with COSMIC annotations
- GA4GH Beacon** - access COSMIC data through the GA4GH Beacon Project
- COSMIC in BioQuery** - search COSMIC via the ICR Cancer Genomics Cloud

<https://cancer.sanger.ac.uk/cosmic>

# COSMIC resistance mutations



## Drug Resistance

COSMIC has started to annotate mutations identified in the literature as resistance mutations, including those conferring acquired resistance (after treatment) and intrinsic resistance (before treatment). This is a work in progress which we aim to expand, including the curation of 2 or more drug treatments and responses.

Resistance to targeted drug treatment occurs in some patients following an initial drug response. This can be caused by the development of resistance mutations, such as those in the drug target preventing drug binding. Acquired resistance develops gradually within the tumour where subpopulations of cells may acquire or already have the mutations enabling them to emerge under selective drug pressure. Patients who initially responded to treatment relapse as a result of the emergence of the dominant resistant clone. Screening patients for mutations at tumour recurrence identifies these new mutations which were not present (at detectable levels) in the primary pre-treatment tumour. Functional studies may confirm the role of these secondary mutations in resistance.

Alternative transcripts are also displayed here for genes where reported resistant mutations are not located on the canonical transcript but are on the alternative, and also where reported resistant mutations are located at the same genomic position on both the canonical and alternative transcripts or on overlapping genes and/or fusions and share a COSM id.

Table to view drug and gene mutation frequency.

Drug	Genes	Unique Resistant Samples	Unique Resistant Mutations
Imatinib	ABL1, ABL1_ENST00000318560, KIT, PDGFRA	1370	189
Tyrosine kinase inhibitor - NS	ABL1, ABL1_ENST00000318560, EGFR, EGFR_ENST00000454757 (GRCh38), EGFR_ENST00000455089, EGFR_ENST00000442591 (GRCh37)	386	71
Gefitinib	EGFR, EGFR_ENST00000454757 (GRCh38), EGFR_ENST00000442591 (GRCh37), EGFR_ENST00000455089	238	9
Erlotinib	EGFR, EGFR_ENST00000454757 (GRCh38), EGFR_ENST00000442591 (GRCh37)	84	5
Crizotinib	ALK, ALK_ENST00000431873 (GRCh37), ALK_ENST00000618119 (GRCh38), MET, MET_ENST00000397752, MET_ENST00000539704	73	50
Endocrine therapy	ESR1, ESR1_ENST00000206249, ESR1_ENST00000338799, ESR1_ENST00000406595, ESR1_ENST00000427531, ESR1_ENST00000443427, ESR1_ENST00000456483	71	54
Dasatinib	ABL1, ABL1_ENST00000318560	66	11
Purine Analogue	NT5C2, NT5C2_ENST00000404739, NT5C2_ENST00000423468, NT5C2_ENST00000470299	57	81
Ibrutinib	BTX, BTX_ENST00000372880 (GRCh37), BTX_ENST00000308731 (GRCh38)	51	18
Afatinib	EGFR, EGFR_ENST00000454757 (GRCh38), EGFR_ENST00000442591 (GRCh37), EGFR_ENST00000455089	48	6
Docetaxel	EGFR, EGFR_ENST00000454757 (GRCh38), EGFR_ENST00000442591 (GRCh37), EGFR_ENST00000455089	48	62
Vandetanib	SMO	42	19
Vemurafenib	BRAF, BRAF_ENST00000288602 (GRCh38)	24	7

[https://cancer.sanger.ac.uk/cosmic/drug\\_resistance](https://cancer.sanger.ac.uk/cosmic/drug_resistance)

# COSMIC resistance mutations

**Gene**  
EGFR

- Gene view
- Overview
- External links
- Drug resistance
- Tissue distribution
- Genome browser
- Mutation distribution
- Variants
- References

Reset case

**Search**

Search COSMIC

**Filters**  
Show advanced filters

**Range** Show input fields  
1 1211

**Coordinate system**  
 Amino-acid  
 cDNA

Apply filters Reset filters

**Drug resistance**

This section shows the drugs associated with **EGFR** resistance mutations. In the tabs below you can see any other genes that have resistance mutations to the same drug(s), and the distribution of mutations that occur in those genes.

Alternative transcripts are also displayed here for genes where reported resistant mutations are not located on the canonical transcript but are on the alternative, and also where reported resistant mutations are located at the same genomic position on both the canonical and alternative transcripts or on overlapping genes and/or fusions and share a COSM ID.

You can change the list of drugs that are used to filter data in the panels below; click the name of a drug to toggle it on or off, then click "Update drugs".

Afatinib Erlotinib Gefitinib Osimertinib Tyrosine Kinase Inhibitor - NS

Update drugs

Mutations Genes

For each of the selected genes and selected drugs, this histogram shows the number of samples with a particular resistant mutation. You can change the list of genes that are shown in the histogram below; click the name of the gene to toggle it on or off, then click "Update genes".

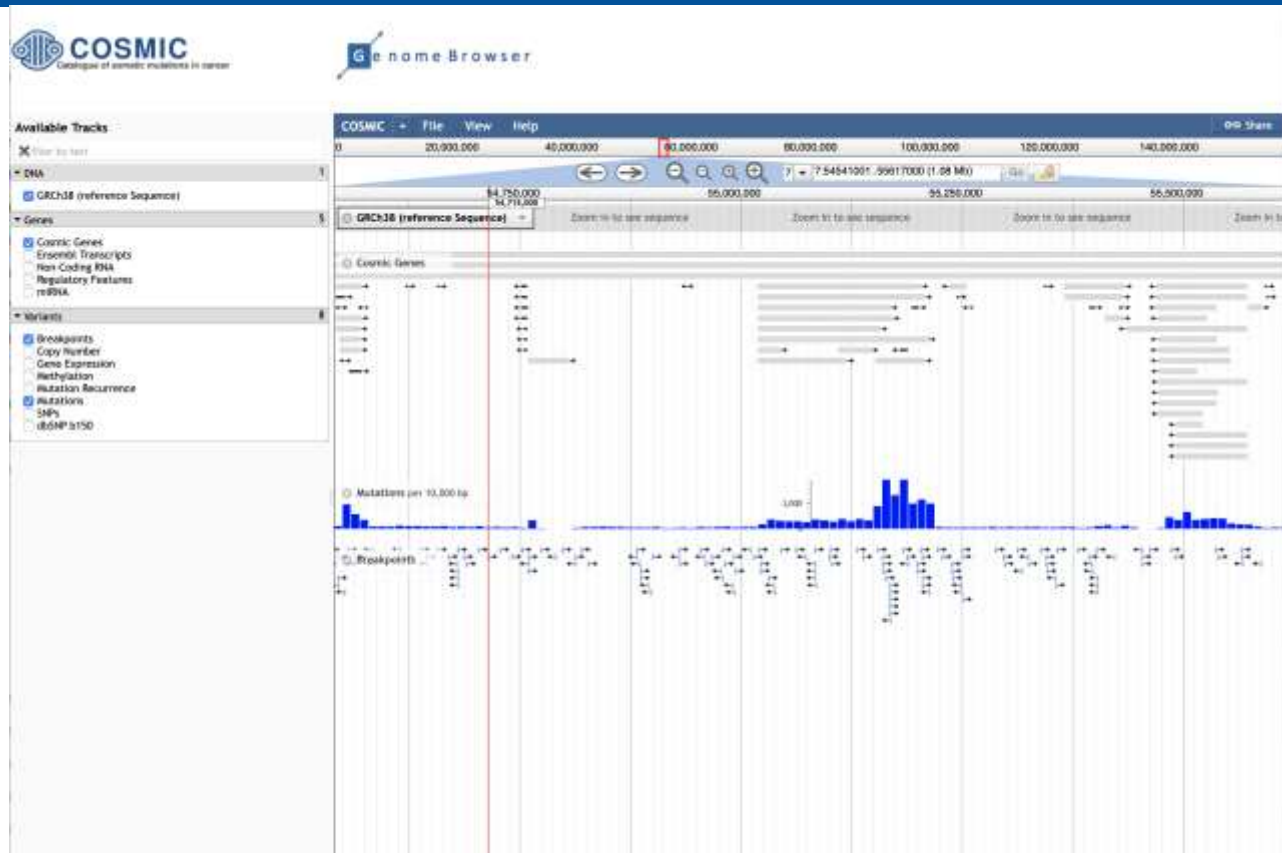
ABL1 ABL1\_ENST00000318560 ASXL1 BRAF EGFR EGFR\_ENST00000454757 NF1 NF1\_ENST000003356175 NF2 NF2\_ENST00000334961

Update genes

**Number of Samples**

[https://cancer.sanger.ac.uk/cosmic/drug\\_resistance](https://cancer.sanger.ac.uk/cosmic/drug_resistance)

# COSMIC view of mutations in and around EGFR



# COSMIC view of mutations around EGFR



**Available Tracks**

- Chrom by track
- DNA**
  - GRCh38 (reference Sequence)
- Genes**
  - Cosmic Genes
  - Ensembl Transcripts
  - Non-Coding RNA
  - Regulatory Features
  - miRNA
- Variants**
  - Breakpoints
  - Copy Number
  - Gene Expression
  - Methylation
  - Mutation Recurrence
  - Mutations
  - SAs
  - dGAP p150

**COSMIC - File View Help**

0 20,000,000 40,000,000 60,000,000 80,000,000 100,000,000 120,000,000 140,000,000

55,150,000 55,175,000 55,199,504 55,200,000 55,225,000

7 = 7,591,26001...20296800 (197.8 kb)

GRCh38 (reference Sequence)    Look it to see sequence    Look it to see sequence    Look it to see sequence    Look it to see sequence

Cosmic Genes

RF01333
RF00372
RF00375
EGFR_ENST00000455389
EGFR_ENST00000454757
EGFR_ENST0000042916
EGFR_ENST0000044576
EGFR_ENST0000042316
EGFR
EGFR_ENST00000618463
EGFR_ENST0000060046

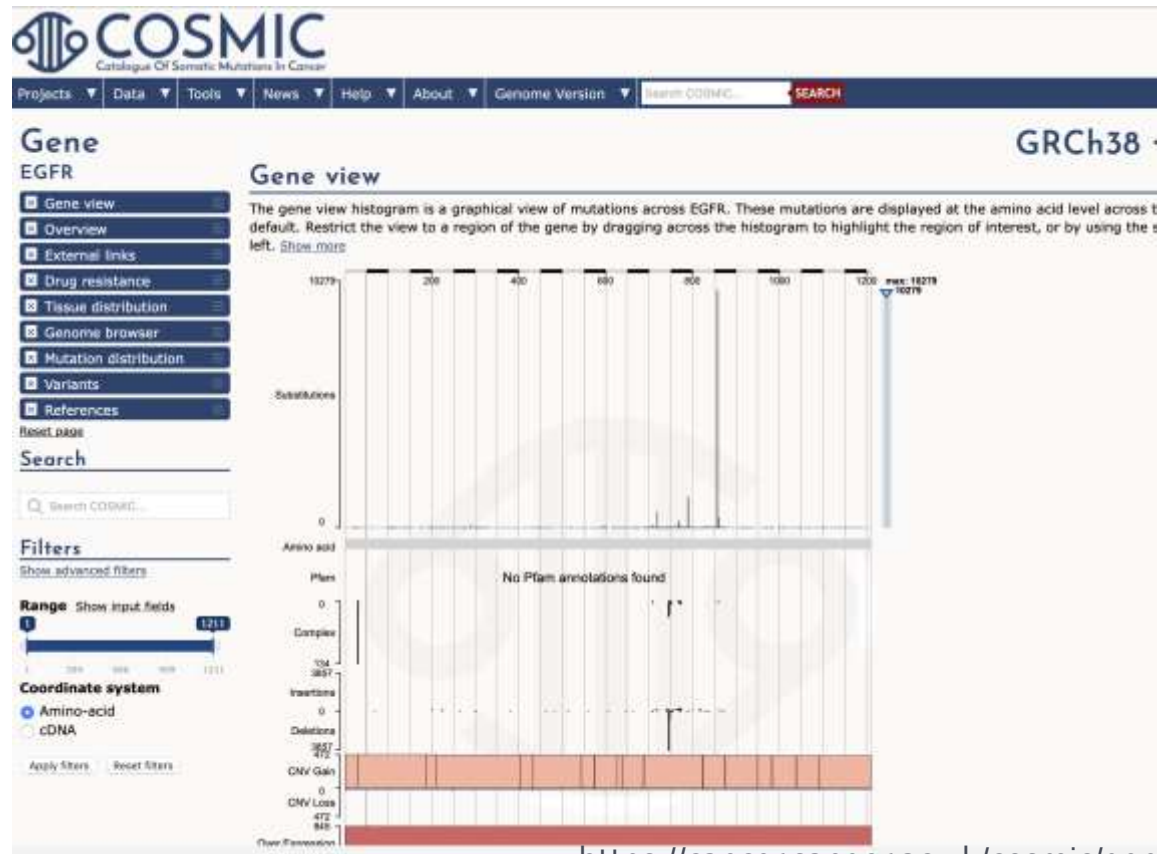
EGFR-AS1    EGFR\_ENST00000485503

Metastroms per 1,000 bp

Breakpoints

COS1175768	COS1175764	COS1175765	COS1175766	COS1175767	COS1175768	COS1175769	COS1175770	COS1175771	COS1175772	COS1175773	COS1175774	COS1175775	COS1175776	COS1175777	COS1175778	COS1175779	COS1175780	COS1175781	COS1175782	COS1175783	COS1175784	COS1175785	COS1175786	COS1175787	COS1175788	COS1175789	COS1175790	COS1175791	COS1175792	COS1175793	COS1175794	COS1175795	COS1175796	COS1175797	COS1175798	COS1175799	COS1178000	COS1178001	COS1178002	COS1178003	COS1178004	COS1178005	COS1178006	COS1178007	COS1178008	COS1178009	COS1178010	COS1178011	COS1178012	COS1178013	COS1178014	COS1178015	COS1178016	COS1178017	COS1178018	COS1178019	COS1178020	COS1178021	COS1178022	COS1178023	COS1178024	COS1178025	COS1178026	COS1178027	COS1178028	COS1178029	COS1178030	COS1178031	COS1178032	COS1178033	COS1178034	COS1178035	COS1178036	COS1178037	COS1178038	COS1178039	COS1178040	COS1178041	COS1178042	COS1178043	COS1178044	COS1178045	COS1178046	COS1178047	COS1178048	COS1178049	COS1178050	COS1178051	COS1178052	COS1178053	COS1178054	COS1178055	COS1178056	COS1178057	COS1178058	COS1178059	COS1178060	COS1178061	COS1178062	COS1178063	COS1178064	COS1178065	COS1178066	COS1178067	COS1178068	COS1178069	COS1178070	COS1178071	COS1178072	COS1178073	COS1178074	COS1178075	COS1178076	COS1178077	COS1178078	COS1178079	COS1178080	COS1178081	COS1178082	COS1178083	COS1178084	COS1178085	COS1178086	COS1178087	COS1178088	COS1178089	COS1178090	COS1178091	COS1178092	COS1178093	COS1178094	COS1178095	COS1178096	COS1178097	COS1178098	COS1178099	COS1178100	COS1178101	COS1178102	COS1178103	COS1178104	COS1178105	COS1178106	COS1178107	COS1178108	COS1178109	COS1178110	COS1178111	COS1178112	COS1178113	COS1178114	COS1178115	COS1178116	COS1178117	COS1178118	COS1178119	COS1178120	COS1178121	COS1178122	COS1178123	COS1178124	COS1178125	COS1178126	COS1178127	COS1178128	COS1178129	COS1178130	COS1178131	COS1178132	COS1178133	COS1178134	COS1178135	COS1178136	COS1178137	COS1178138	COS1178139	COS1178140	COS1178141	COS1178142	COS1178143	COS1178144	COS1178145	COS1178146	COS1178147	COS1178148	COS1178149	COS1178150	COS1178151	COS1178152	COS1178153	COS1178154	COS1178155	COS1178156	COS1178157	COS1178158	COS1178159	COS1178160	COS1178161	COS1178162	COS1178163	COS1178164	COS1178165	COS1178166	COS1178167	COS1178168	COS1178169	COS1178170	COS1178171	COS1178172	COS1178173	COS1178174	COS1178175	COS1178176	COS1178177	COS1178178	COS1178179	COS1178180	COS1178181	COS1178182	COS1178183	COS1178184	COS1178185	COS1178186	COS1178187	COS1178188	COS1178189	COS1178190	COS1178191	COS1178192	COS1178193	COS1178194	COS1178195	COS1178196	COS1178197	COS1178198	COS1178199	COS1178200	COS1178201	COS1178202	COS1178203	COS1178204	COS1178205	COS1178206	COS1178207	COS1178208	COS1178209	COS1178210	COS1178211	COS1178212	COS1178213	COS1178214	COS1178215	COS1178216	COS1178217	COS1178218	COS1178219	COS1178220	COS1178221	COS1178222	COS1178223	COS1178224	COS1178225	COS1178226	COS1178227	COS1178228	COS1178229	COS1178230	COS1178231	COS1178232	COS1178233	COS1178234	COS1178235	COS1178236	COS1178237	COS1178238	COS1178239	COS1178240	COS1178241	COS1178242	COS1178243	COS1178244	COS1178245	COS1178246	COS1178247	COS1178248	COS1178249	COS1178250	COS1178251	COS1178252	COS1178253	COS1178254	COS1178255	COS1178256	COS1178257	COS1178258	COS1178259	COS1178260	COS1178261	COS1178262	COS1178263	COS1178264	COS1178265	COS1178266	COS1178267	COS1178268	COS1178269	COS1178270	COS1178271	COS1178272	COS1178273	COS1178274	COS1178275	COS1178276	COS1178277	COS1178278	COS1178279	COS1178280	COS1178281	COS1178282	COS1178283	COS1178284	COS1178285	COS1178286	COS1178287	COS1178288	COS1178289	COS1178290	COS1178291	COS1178292	COS1178293	COS1178294	COS1178295	COS1178296	COS1178297	COS1178298	COS1178299	COS1178300
------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------

# COSMIC view of EGFR







MY CANCER GENOME®  
GENETICALLY INFORMED CANCER MEDICINE



Clinical Implications of Molecular Biomarkers



# EGFR in MyCancerGenome



MY CANCER GENOME  
GENETICALLY INFORMED CANCER MEDICINE

Clinical Trials

Diseases

Biomarkers

Drugs

Pathways

## Results for EGFR (1901)

Show results for

- Biomarkers (259)
- Diseases (66)
- Drugs (278)
- Clinical Trials (1295)
- Pathways (3)

### Biomarkers (259)

#### EGFR

**Diseases:** Cancer: Pancreatic Adenocarci...  
**Pathways:** Receptor tyrosine kinase/growt...  
**Clinical Trials:** 353  
**Drugs:** 9

#### EGFR L858R

**Alteration Groups:** EGFR Activating Muta...  
**Diseases:** Non-Small Cell Lung Carcinom...  
**Clinical Trials:** 198  
**Drugs:** 8

#### EGFR L861Q

**Alteration Groups:** EGFR Activating Muta...  
**Diseases:** Pancreatic Carcinoma, Esopha...  
**Clinical Trials:** 143  
**Drugs:** 8

#### EGFR S768I

**Alteration Groups:** EGFR Activating Muta...  
**Diseases:** Pancreatic Carcinoma, Non-Sm...  
**Clinical Trials:** 136  
**Drugs:** 8

#### EGFR Exon 19 Deletion

**Alteration Groups:** EGFR Activating Muta...  
**Diseases:** Pancreatic Carcinoma, Non-Sm...  
**Clinical Trials:** 191  
**Drugs:** 8

#### EGFR T790M

**Alteration Groups:** EGFR Resistance Mut...  
**Diseases:** Non-Small Cell Lung Carcinom...  
**Clinical Trials:** 80  
**Drugs:** 8

# EGFR in MyCancerGenome

Biomarkers /

EGFR

[↶ Back to Biomarkers List](#)

- > **Associated Genetic Biomarkers**
- > **Associated Diseases**
- > **Associated Pathways**

## Overview

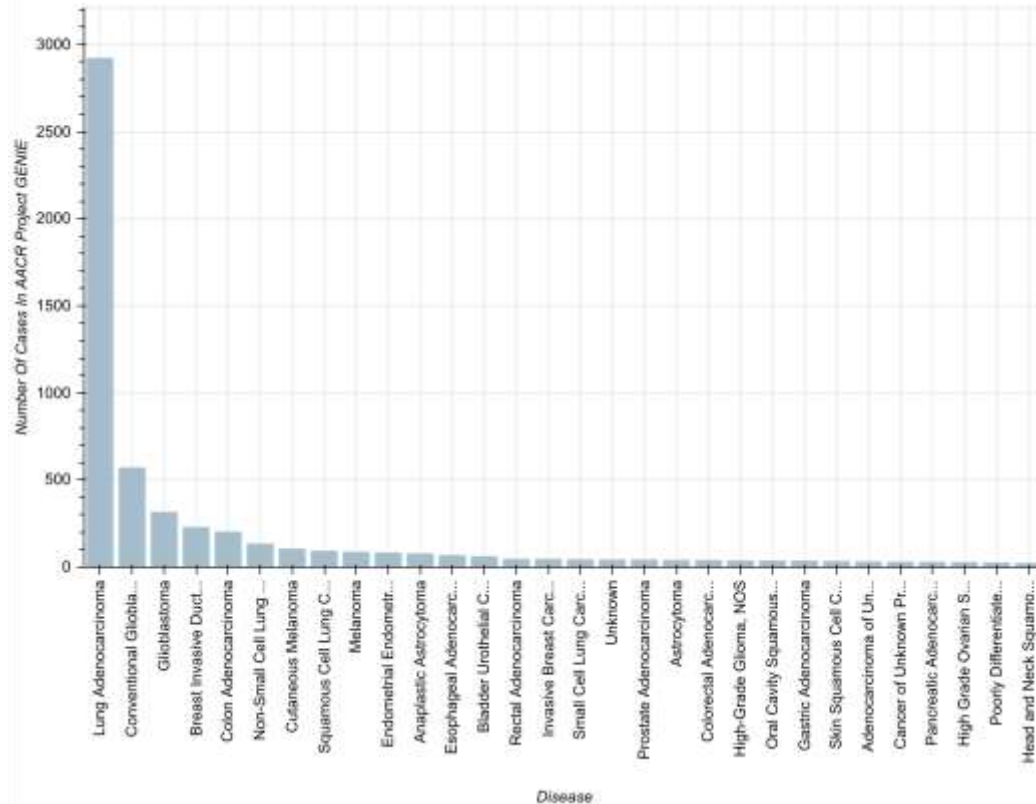
<a href="#">Location [1]</a>	7p11.2
<a href="#">Pathway</a>	Receptor tyrosine kinase/growth factor signaling
<a href="#">Protein [2]</a>	Epidermal growth factor receptor
<a href="#">Synonyms [1]</a>	mENA, HER1, NISBD2, ERBB, ERBB1, PIG61

EGFR (epidermal growth factor receptor, also known as ERBB1 and HER1) is a gene that encodes for the epidermal growth factor receptor protein. Missense mutations, deletions, and insertions are observed in cancers such as lung cancer and glioblastoma. Activating EGFR mutations increase the kinase activity of EGFR, leading to hyperactivation of downstream pro-survival signaling pathways (PMID: 15284455).

EGFR is altered in 6.83% of all cancers with lung adenocarcinoma, conventional glioblastoma multiforme, glioblastoma, breast invasive ductal carcinoma, and colon adenocarcinoma having the greatest prevalence of alterations [3].

**EGFR GENIE Cases - Top Diseases**

# Most common cancers with EGFR Mutations in GENIE



# Drugs in MyCancerGenome – selecting TKIs



MY CANCER GENOME<sup>®</sup>  
GENETICALLY INFORMED CANCER MEDICINE

Clinical Trials

Diseases

Biomarkers

Drugs

Pathways

## Drugs (101)

Tyrosine kinase inh....

Search by:

afatinib [x] [Q]

Refine by:

### Drug Categories

Search [x] [Clear]

- Tyrosine kinase inhibitors (101)
- Therapeutic antibodies (39)
- Serine/threonine kinase inhibitors (92)
- Immunotherapies (71)
- Antimetabolites (19)
- Alkylating agents (17)
- CDK inhibitors (17)
- PKC inhibitors (16)
- Histone deacetylase/HDAC inhibitors (12)
- Antimicrotubulin agents (11)

### Clinical Practice Guidelines

Search [x] [Clear]

- NCCN (32)
- FDA (30)
- NICE (12)
- SMC (12)
- BNF (10)

imatinib

**Diseases:** Melanoma, Myelodysplastic/...  
**Drug Categories:** ABL inhibitors, Tyrosi...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** STI-571, CGP57148, CGP5...

afatinib

**Diseases:** Non-Small Cell Lung Carcinoma  
**Drug Categories:** Tyrosine kinase inhibi...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** BIBW-2992, Giotin, Tomtov...

crizotinib

**Diseases:** Non-Small Cell Lung Carcino...  
**Drug Categories:** ALK inhibitors, ROS1...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** MET tyrosine kinase inhibito...

erlotinib

**Diseases:** Non-Small Cell Lung Carcinoma  
**Drug Categories:** Tyrosine kinase inhibi...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** OSI-774, CP-268,774, Tarce...

gefitinib

**Diseases:** Non-Small Cell Lung Carcinoma  
**Drug Categories:** Tyrosine kinase inhibi...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** ZD1839, Iressa, gefitinib, 4-...

osimertinib

**Diseases:** Non-Small Cell Lung Carcinoma  
**Drug Categories:** Tyrosine kinase inhibi...  
**Clinical Practice Guidelines:** NICE, SM...  
**Synonyms:** 1421373-65-0, Osimertinib (...)

# gefitinib in MyCancerGenome



MY CANCER GENOME  
GENETICALLY INFORMED CANCER MEDICINE

Clinical Trials

Diseases

Biomarkers

Drugs

Drugs /

gefitinib

↩ Back to Drugs List

> **Associated Genetic Biomarkers**

> **Associated Diseases**

## Overview

Generic Name(s):

gefitinib

Trade Name(s):

Iressa

NCI Definition (1):

An anilinoquinazoline with antineoplastic activity. Gefitinib inhibits the catalytic activity of numerous tyrosine kinases including the epidermal growth factor receptor (EGFR), which may result in inhibition of tyrosine kinase-dependent tumor growth. Specifically, this agent competes with the binding of ATP to the tyrosine kinase domain of EGFR, thereby inhibiting receptor autophosphorylation and resulting in inhibition of signal transduction. Gefitinib may also induce cell cycle arrest and inhibit angiogenesis. (NCI04)

## ▾ Biomarker-Directed Therapies

Gefitinib can be used in the treatment of non-small cell lung carcinoma. EGFR, EGFR A763\_Y764InsFQEA, and EGFR Exon 19 Deletion are the most frequent biomarker inclusion criteria for use of gefitinib [2].

Non-Small Cell Lung Carcinoma



[View Clinical Trials for gefitinib](#)

## ▾ Clinical Trials

Gefitinib has been investigated in 23 clinical trials, of which 20 are open and 3 are closed. Of the trials investigating gefitinib, 2 are phase 1 (1 open), 1 is phase 1/phase 2 (1 open), 14 are phase 2 (12 open), and 6 are phase 3 (6 open).

EGFR L858R, EGFR Exon 19 Deletion, and EGFR L861Q are the most frequent biomarker inclusion criteria for gefitinib clinical trials.

Non-small cell lung carcinoma, lung adenocarcinoma, and malignant solid tumor are the most common diseases being investigated in gefitinib clinical trials [2].

What about exclusion biomarkers like EGFR T790M?

# Going back to the NCI Genomic Data Commons

<https://portal.gdc.cancer.gov>





# Look for EGFR

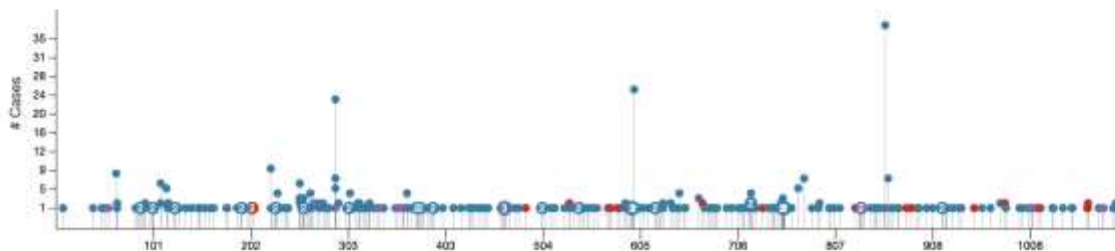
Project	Disease Type	Site	# SSM Affected Cases	# CNV Gains	# CNV Losses	# Mutations
<a href="#">TCGA-GBM</a>	→ 2 Disease Types	Brain	106 / 393 (26.97%)	276 / 596 (46.31%)	9 / 596 (1.51%)	74
<a href="#">TCGA-LUAD</a>	→ 3 Disease Types	Bronchus and lung	83 / 567 (14.64%)	54 / 513 (10.53%)	10 / 513 (1.95%)	51
<a href="#">TCGA-SKCM</a>	Nevi and Melanomas	Skin	60 / 469 (12.79%)	6 / 468 (1.28%)	9 / 468 (1.92%)	82
<a href="#">CPTAC-3</a>	→ 5 Disease Types	→ 6 Primary Sites	75 / 638 (11.91%)	0 / 0 (0.00%)	0 / 0 (0.00%)	56
<a href="#">TCGA-UCEC</a>	→ 4 Disease Types	→ 2 Primary Sites	57 / 530 (10.75%)	19 / 510 (3.73%)	5 / 510 (0.98%)	107
<a href="#">HCGI-CMDC</a>	→ 9 Disease Types	→ 14 Primary Sites	3 / 40 (7.50%)	0 / 0 (0.00%)	0 / 0 (0.00%)	3
<a href="#">TCGA-LGG</a>	Gliomas	Brain	35 / 510 (6.86%)	40 / 497 (8.05%)	1 / 497 (0.20%)	26
<a href="#">TCGA-COAD</a>	→ 4 Disease Types	→ 2 Primary Sites	24 / 400 (6.00%)	15 / 448 (3.35%)	1 / 448 (0.22%)	28
<a href="#">TCGA-STAD</a>	→ 2 Disease Types	Stomach	26 / 440 (5.91%)	33 / 432 (7.64%)	5 / 432 (1.16%)	29
<a href="#">TCGA-CESC</a>	→ 4 Disease Types	Cervix uteri	14 / 289 (4.84%)	11 / 294 (3.74%)	0 / 294 (0.00%)	17

Show 10 entries

1 2 3 4 5

## EGFR - Protein

Transcript: [ENST00000275483 \(1210 aa\)](#) [Reset](#) [Download](#)



Some overlapping domains are not shown by default. [Click here to show / hide them.](#)



EGFR T790M mutations arise as resistance mutations to 1<sup>st</sup> or 2<sup>nd</sup> gen TKI therapy  
So it is not common in patients/cancers who do not get these therapies

Viewing 323 / 323 Mutations

Consequence [\[v\]](#)

Select All [ Deselect All

- Missense: 289 / 289
- Stop Gained: 20 / 20
- Frameshift: 14 / 14



# Look for EGFR

## Most Frequent Somatic Mutations

[Open in Exploration](#)

Showing 1 - 10 of 577 somatic mutations

☰ JSON TSV Save/Edit Mutation Set

DNA Change	Type	Consequences	# Affected Cases in EGFR	# Affected Cases Across the GDC	Impact
<a href="#">chr7:g.55191822T&gt;G</a>	Substitution	<b>Missense</b> EGFR L858R	38 / 1,341  2.83%	38 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55165350G&gt;T</a>	Substitution	<b>Missense</b> EGFR G598V	25 / 1,341  1.86%	25 / 12,174	<span>MO</span> <span>DH</span> <span>PO</span>
<a href="#">chr7:g.55174772delGGAATTAA...</a>	Deletion	<b>Inframe Deletion</b> EGFR E746_A750del	25 / 1,341  1.86%	25 / 12,174	<span>MO</span> -- --
<a href="#">chr7:g.55154129C&gt;T</a>	Substitution	<b>Missense</b> EGFR A289V	23 / 1,341  1.72%	23 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55152581G&gt;T</a>	Substitution	<b>Missense</b> EGFR R222C	9 / 1,341  0.67%	9 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55142382T&gt;G</a>	Substitution	<b>Missense</b> EGFR L62R	8 / 1,341  0.60%	8 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55191831T&gt;A</a>	Substitution	<b>Missense</b> EGFR L861Q	7 / 1,341  0.52%	7 / 12,174	<span>MO</span> <span>DH</span> <span>PO</span>
<a href="#">chr7:g.55154129C&gt;A</a>	Substitution	<b>Missense</b> EGFR A289D	7 / 1,341  0.52%	7 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55181329G&gt;A</a>	Substitution	<b>Missense</b> EGFR V774M	7 / 1,341  0.52%	7 / 12,174	<span>MO</span> <span>DH</span> <span>PR</span>
<a href="#">chr7:g.55154017C&gt;T</a>	Substitution	<b>Missense</b> EGFR R252C	6 / 1,341  0.45%	6 / 12,174	<span>MO</span> <span>DH</span> <span>PO</span>

Show  entries◀ ▶ 1 2 3 4 5 6 7 8 9 10 ⋮

# CIViC DB

About Participate Community Help FAQ [Sign In/Sign Up](#)

Go to Genes & Variants

 Discover supported clinical interpretations of mutations related to cancer.

 Participate with colleagues to add variants and support for cancer-related mutations.

**The Precision Medicine Revolution**  
Precision medicine refers to the use of prevention and treatment strategies that are tailored to the unique features of each individual and their disease. In the context of cancer this might involve the identification of specific mutations shown to predict response to a targeted therapy. The biomedical literature describing these associations is large and growing rapidly. Currently these interpretations exist largely in private or unnumbered databases resulting in extensive repetition of effort.

**CIViC's Role in Precision Medicine**  
Realizing precision medicine will require this information to be centralized, debated and integrated for application in the clinic. CIViC is an open access, open source, community-driven web resource for Clinical Interpretation of Variants in Cancer. Our goal is to enable precision medicine by providing an educational forum for dissemination of knowledge and active discussion of the clinical significance of cancer genome alterations. For more details refer to the 2017 CIViC publication in Nature Genetics.

**BECOME AN EDITOR**  
and help moderate updates to CIViC!

[Click Here](#)  
to learn more and apply.

**Announcements**

June 26th, 2020  
 Explore CIViC Variants in ProtonPain: St. Jude's ProtonPain now incorporates a CIViC variant track, displaying CIViC curated data along with a variety of additional sources, and providing one-click

May 20th, 2019  
 New YouTube Tutorial: We've updated the CIViC Help player with a variety of new video tutorials, helping you to get started collaborating with us by suggesting sources, adding and editing evidence, and more. Also, we've produced videos that provide details about our collaborations with ClinGen Working Groups and a new introduction to CIViC collaboration opportunities. Please watch the associated introductory documentation, and let us know in

Similar to **ClinVar**,  
**ClinGen**,  
**MyCancerGenome**

CIViC

[About](#) [Participate](#) [Community](#) [Help](#) [FAQ](#) [Log In/Sign Up](#)

BROWSE
SEARCH
ACTIVITY
ADD

G
EGFR

[Gene Summary](#)
[Gene Talk](#)

Last Modified by [PainfulKiss](#) Last Reviewed by [Miyazaki](#)

EGFR is widely recognized for its importance in cancer. Amplification and mutations have been shown to be driving events in many cancer types. Its role in non-small cell lung cancer, glioblastoma and basal like breast cancers has opened many research and drug development efforts. Tyrosine kinase inhibitors have shown efficacy in EGFR amplified tumors, most notably gefitinib and erlotinib. Mutations in EGFR have been shown to confer resistance to these drugs, particularly the variant T790M, which has been functionally characterized as a resistance marker for both of these drugs. The later generation TKIs have seen some success in treating these resistant cases, and targeted sequencing of the EGFR locus has become a common practice in treatment of non-small cell lung cancer. Overproduction of ligands is another possible mechanism of activation of EGFR. ERBB ligands include EGF, TGF- $\alpha$ , AREG, EPIG, BTC, HD-EGF, EPR and NRG1-4 (for detailed information please refer to the respective ligand section).

**Sources:**

- [Tomida et al., 2013, \*Bioinformatics\*](#)
- [Charney et al., in \*Who\*](#)
- [Ahsan et al., 2014, \*Cancer Cell\*](#)

**Name:** epidermal growth factor receptor

**Entrez Symbol:** EGFR **Entrez ID:** 1358 **UniProtKB ID:** P02283

**Aliases:** ERBB, ERBB1, HER1, NS5B02, PIG61, mDNA

**Chromosome:** 7 **Start:** 55366714 **End:** 55324313 **Strand:** 1 (3' to 5')

**Protein Domains:** [Furin-like cysteine-rich domain](#), [Furin-like repeat](#), [Growth factor receptor cysteine-rich domain](#), [Growth factor receptor domain 4](#), [Leucine-rich repeat domain](#), [I, domain-like](#), (3 items)

**Pathways:** [Alpha6beta4integrin](#), [AndrogenReceptor](#), [EGFR1](#), [Gastrin](#), [Leptin](#), (18 items)

[View MyGene Info Details](#)

**EGFR Variants & Variant Groups**

Filter by name ▼ Display Options

A289V A702G A750T A763\_V759delPGEA A767\_V759delAGV A883T A864T AMPLIFICATION C797G C797S C797Y

COPY NUMBER VARIATION D757H D761Y D770\_N771insDL D770\_N771insGT D770\_N771insGY D770\_N771insHPG D770\_N771insHVD

D770delinsGY E6L74T\_P753insS E709L710-Q E709A and G719C E709K and G715A E709Q E734Q E746\_A750+P

E746\_A750del E746\_S753+H E746\_S752+D E746\_L751+I E746L775I delinsA E746L775I delinsVA E746Q E740V E858S

E884K EGFR RAD51 Ex13 del L858R Exon 18 deletion EXON 18 OVEREXPRESSION EXON 19 DELETION EXON 20 INSERTION

EXON 4 DELETION EXPRESSION F404I F404V G12V G465E G465R G465V G586V G719 G719A G719Q

G719E G719K G724E G810S Gain of function H773\_V774insH H773\_V774insHPH I462E I462R I491K I491R

I744\_K745insEPVNI K467N K467T KARNE K489Q K745\_E746delKSLRE K757H K805E L718Q L716V

L718V and L718Q L747\_A750+P L747\_P753-Q L747\_P753delnsS L747\_S752del L747\_S752delnsQ L747\_T751+P L747\_T751+Q

# CIViC DB – EGFR T790M

Mutation effect in the description is clear

Mutation effect on these drugs is not clear

**EGFR T790M**

**MP Expression**  
EGFR T790M

**Description**  
EGFR T790M was one of the very first mutations recognized to confer resistance to targeted therapies in non-small cell lung cancer. While successful in amplified EGFR, the efficacy of the first and second generation TKI's (erlotinib, gefitinib, neratinib) in treating patients harboring this mutation before treatment is notably lower. This lack of efficacy can likely be to blame for the poorer prognosis for patients with this mutation as compared to patients with wildtype EGFR or other types of EGFR mutations. Approximately half of EGFR mutant tumors with acquired resistance to TKI inhibitors have been shown to harbor this mutation, implicating it as a mechanism of acquired therapy resistance. A third generation TKI (osimertinib) has been approved for the treatment of EGFR T790M mutant NSCLC. Patients positive for T790M in a plasma-based test have similar outcomes like those with tumor biopsy testing.

**Molecular Profile Score**  
408.25

**MP Variants**

Variant	Gene
T790M	EGFR
THR760MET	Misense variant
R3121A34G99	166T5
CA20092E	BRAF V600E AND EGFR...
OpenCRAWT	Variant Report

**Evidence** 31 of 40 displayed

OID	Disease	Therapies	IT	DESC	EL	ET	EO	S	VO	R
EBD238	Lung Non-small Cell...	Erwinib	NA	A	A	0	0	0	0	5★
EBD1502	Lung Non-small Cell...	Osimertinib	NA	A	A	0	0	0	0	5★
EBD1867	Lung Non-small Cell...	Osimertinib	NA	A	A	0	0	0	0	5★
EBD140	Lung Non-small Cell...	Docetaxel	NA	B	A	0	0	0	0	4★

# CIViC DB – erlotinib

**Erlotinib** # C02030

**MyChemInfo**

**ChEMBL Definition:** A quinazoline compound having a 3-(4-ethylphenyl)amino group at the 4-position and two 2-methoxyethoxy groups at the 6- and 7-positions.

**Indications:** Adenocarcinoma of Lung, Diffuse Intrinsic Pontine Glioma, Astrocytoma, Biliary Tract Neoplasms, Urinary Bladder Neoplasms, Brain Neoplasms, Breast Neoplasms, Carcinoma, Adenocarcinoma, Bronchiolo-Alveolar, Carcinoma, Non-Small-Cell Lung, Carcinoma, Serous Cell, Carcinoma, Squamous Cell, Colorectal Neoplasms, Digestive System Diseases, Epidermolysis, Esophageal Neoplasms, Fallopian Tube Neoplasms, Glioblastoma, Glioma, Head and Neck Neoplasms, Carcinoma, Hepatocellular, Kidney Neoplasms, Leukemia, Liver Neoplasms, Lung Neoplasms, Melanoma, Mesothelioma, Mouth Neoplasms, Multiple Myeloma, Myelodysplastic Syndromes, Nasopharyngeal Neoplasms, Neoplasm Metastasis, Neoplasms, Neoplasms, Germ Cell and Embryonal, Oligodendroglioma, Ovarian Neoplasms, Pancreatic Neoplasms, Pharyngeal Neoplasms, Polychemia Vera, Adenomatous Polyposis Coli, Prostate, Rectal Neoplasms, Skin Neoplasms, Thyroid, Colorectal Neoplasms, Leukemia, Myeloid, Acute, Central Nervous System Neoplasms, Carcinoma, Lung Cell, Oligosarcoma, Neuroendocrine Tumors, Hemorrhagic Fever, Ebola, Hematologic Neoplasms, Hepatitis C, Chronic, Prostatic Neoplasms, Castration-Resistant

**ChEMBL Identifiers:** [ChEMBL:114785](#) [ChEMBL:553](#) [PubChem:176070](#) [PubChem:15134887924](#) [PubMed:337925](#) [DrugBank:DB00530](#)

**SMILES:** AAKEHROO7KAMQ-04FFFDYSA-N

**Aliases:** None specified

**Evidence associated with Erlotinib** 35 of 204 displayed

EID	Molecular Profile	Disease	Therapy	DESC	EL	EF	ED	S	VO	A
E0238	EGFR T790M	Lung Non-small Cell...	Erlotinib							
E02994	EGFR L858R	Lung Non-small Cell...	Erlotinib							
E02986	EGFR Exon 19 Deletion	Lung Non-small Cell...	Erlotinib							
E017240	EGFR L858R OR EGFR Exon 19 Deletion	Lung Non-small Cell...	Erlotinib, Ramucicicab							
E03861	KRAS G12D	Lung Cancer	Erlotinib							
E00982	KRAS G12D	Lung Cancer	Erlotinib							

Mutation effect on erlotinib is not clear

# In conclusion



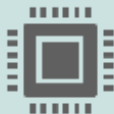
Many different metaphors for building a data repository



Many different methods for making data FAIR (Findable, Accessible, Interoperable, Reusable)



Data wrangling – organizing, formatting, harmonizing, semantically annotating – is hard work and different resources take different approaches



Usability is dependent on the use case, the tools, the problem domain

# What is the right data resource to use?



What is the question you want to answer?



Are you exploring how to answer the problem or ready to analyze?



Do you have an analysis plan and workflow defined?



Are your analysis tools appropriate for the resource you want to use?



Do you need to explore a dataset to understand if the data is available? Does the resource support the kind of exploration you want to do?

# What is the right data resource to use?

Does it have the right kind of data? (clinical trial data, scRNA-seq, specific cell line, disease area, model system)

What are the access policies? (unrestricted, controlled access)

Level of security required (none, limited, regulated [think HIPAA], very sensitive)

Ability to peruse the metadata?

Can you download the data?

If it is an enclave/closed ecosystem, what tools are supported?

Can you bring your own tools?





## **There are many resources out there to use**

Just ask yourself a few pertinent questions as you use them!

Thank you!



**Questions?**

# What is Real World Data?

Collected in the context of patient care. Real World Data was called out as part of the 21<sup>st</sup> Century Cures Act



21<sup>st</sup> Century Cures Act: <https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act>

Graphic from HealthCatalyst: <https://www.healthcatalyst.com/insights/real-world-data-chief-driver-drug-development>