

# Emerging Approaches for Tumor Analyses In Epidemiological Studies

---

---

## Session 2: Public Databases

November 9, 2022

---

---

# Session Overview

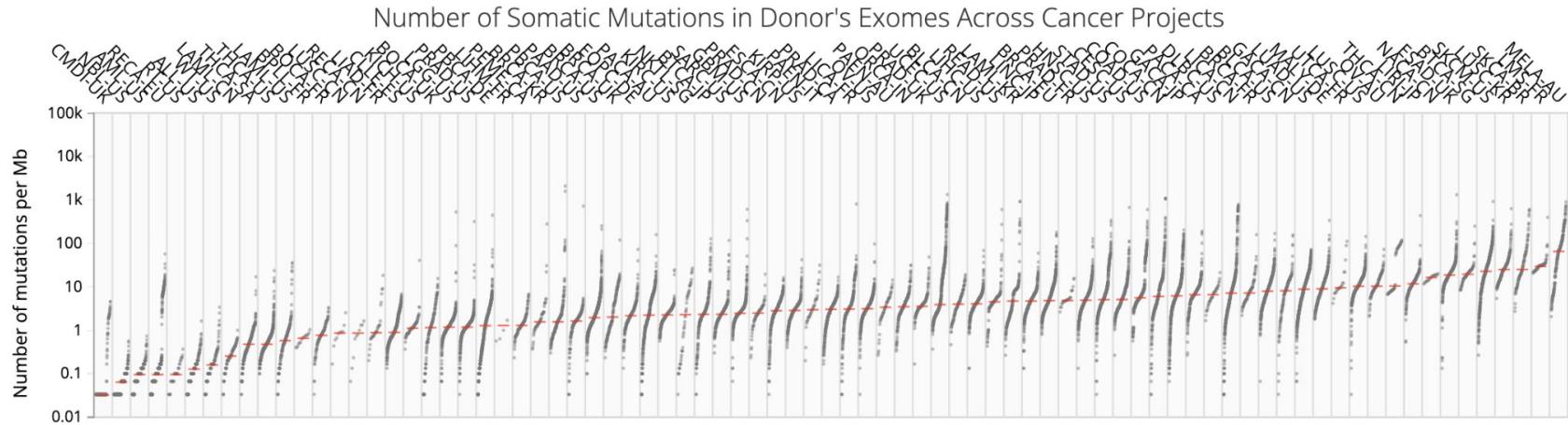
- Major Cancer Genomic Studies with Data Portal Available
- Common Genetics Data Resources
- Specialized Databases for Genomic Analyses
- Analytical Programming Packages for Cancer Genomic Datasets
- Cloud Resources with Available Cancer Genomic Datasets
- “Awesome” Bioinformatics Resources

# Major Cancer Genomic Studies with Data Portal Available

WGS → WES → TS



# WGS: International Cancer Genome Consortium (ICGC)



- Large proportion of TCGA samples included
- Controlled (sensitive genomic and clinical data) vs uncontrolled datasets
- Instant analysis with cloud computing
- Pan-Cancer Analysis of Whole Genomes (PCAWG) (~2700 donors)
- ICGC ARGO Data Platform

234,022 Files    16,307 Donors    1.73 PB

**Data Release 28**    March 27th, 2019

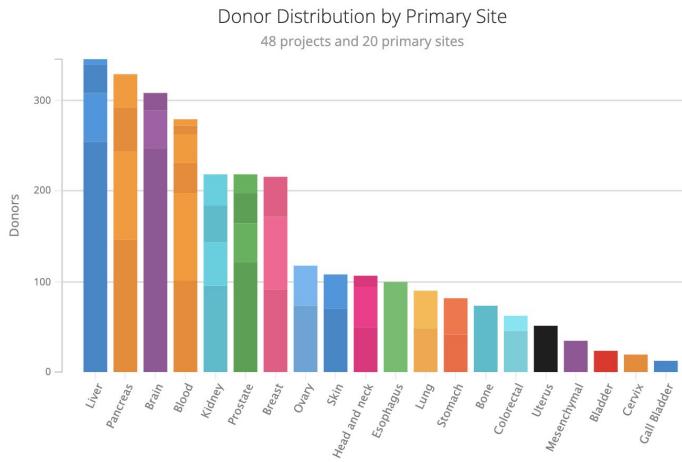
Cancer projects	86
Cancer primary sites	22
Donor with molecular data in DCC	22,330
Total Donors	24,289
Simple somatic mutations	81,782,588

**Summary and Projects**

DCC: ICGC Data Coordination Center

ARGO: Accelerating Research in Genomic Oncology

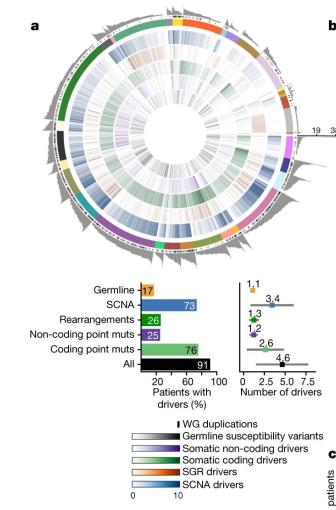
# WGS:Pan-Cancer Analysis of Whole Genomes (PCAWG)



Data Type	# Donors	# Files	Format	Size
SGV	2,715	8,505	VCF	517.27 GB
StGV	2,715	5,668	VCF	7.29 GB
Aligned Reads	2,793	12,168	BAM	794.32 TB
Unaligned Reads	1	1	BAM	104.20 GB
Simple Somatic Mutations	2,715	25,501	VCF	189.99 GB
Copy Number Somatic Mutations	2,715	5,671	VCF	132.62 MB
Structural Somatic Mutations	2,715	14,195	VCF	1.61 GB

Available data as of Oct 25, 2021

2658 cancer whole genomes  
with matched normal tissues  
across 38 tumor types



## Major Publications

1. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. **Pan-cancer analysis of whole genomes**. Nature (2020).
2. Rheinbay, E. et al. **Analyses of non-coding somatic drivers in 2,693 cancer whole genomes**. Nature (2020).
3. PCAWG Transcriptome Core Group et al. **Genomic basis of RNA alterations in cancer**. Nature (2020).
4. Li, Y. et al. **Patterns of somatic structural variation in human cancer genomes**. Nature (2020).
5. Gerstung, M. et al. **The evolutionary history of 2,658 cancers**. Nature (2020).
6. Alexandrov, L. B. et al. **The Repertoire of Mutational Signatures in Human Cancer**. Nature (2020).
7. Phillips, M. et al. **Of Clouds and Genomic Data Protection**. Nature (2020).

# ICGC Data Portal

ICGC Data Portal

Cancer genomics data sets visualization, analysis and download.

Quick Search  Search

e.g. BRAF, KRAS G12D, D035100, MU7870, Ff998, apoptosis, Cancer Gene Census, imatinib, G0016049

Advanced Search

By donors By genes By mutations

Data Release 28 March 27th, 2019

Cancer projects 86

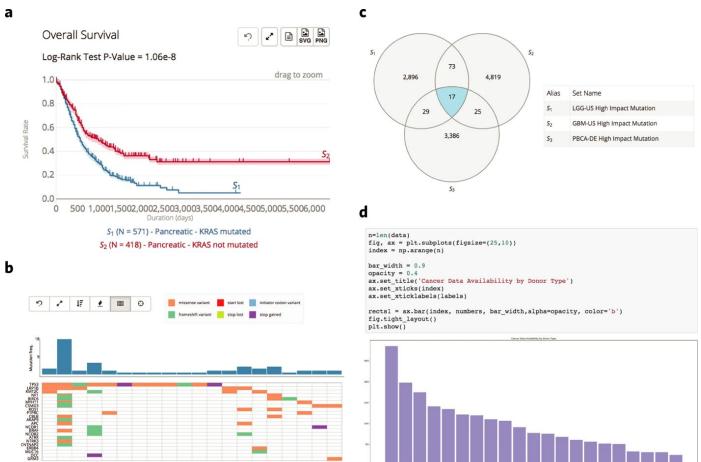
Cancer primary sites 22

Donor with molecular data in DCC 22,330

Total Donors 24,289

Simple somatic mutations 81,782,588

Download Release



## Using the ICGC Data Portal

Repository

- Collaboratory ... 121,467
- GDC - Chicago 90,099
- Azure - Toronto 77,486
- AWS - Virginia 39,193
- PDC - Chicago 22,200
- EGA - Hinxton 5,091

Select all  less

Data Type

- Unaligned Reads 69,131
- SSM 61,591
- Aligned Reads 51,538
- StSM 14,195
- Biospecimen Data 8,881
- Clinical Data 8,841
- SGV 8,505
- CNSM 5,671
- StGV 5,668
- No Data 1

Select all  less

Analysis Software

- DKFZ/EMBL varia... 19,850
- Broad variant call... 19,809
- BWA with Mark D... 17,988
- Sanger variant cal... 11,340
- STAR 2-Pass 9,253
- BWA-align 9,046
- SomaticSniper An... 9,045
- MuTect2 Annotati... 9,037
- MuSE Annotation 9,028
- VarScan2 Annotat... 8,980
- BWA MEM 5,809
- MUSE variant call ... 2,835
- PCAWG SNV-MNV... 2,778
- PCAWG InDel call... 2,778
- STAR 1,465
- TopHat2 1,465
- Pilot50 150
- Silver bullet 1
- No Data 93,365

Select all  less

Experimental Strategy

- WXS 110,698
- WGS 83,530
- RNA-Seq 12,840
- miRNA-Seq 9,046
- Validation 100
- Bisulfite-Seq 86
- No Data 17,722

Select all  less

Only Files in Study

- PCAWG 71,709
- None 162,313

Select all

File Format

- VCF 95,631
- FASTQ 63,543
- BAM 57,103
- BCR XML 17,715
- TGZ 23
- XLSX 7

Select all  less

Access

- Controlled 216,300
- Open 17,722

Select all

# WGS:100,000 Genomes Project | Genomics England

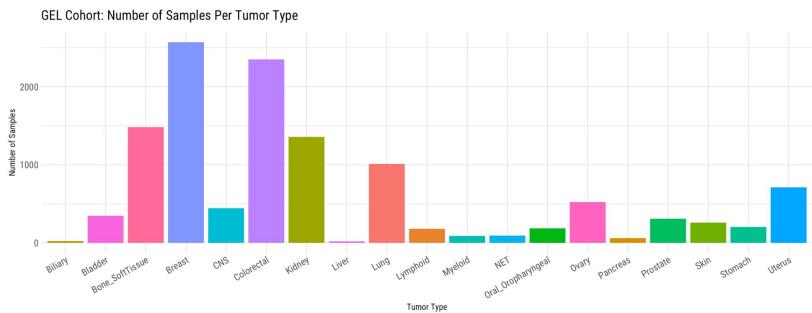
Genomics England's very first initiative – sequencing 100,000 genomes from around 85,000 NHS patients affected by rare disease or cancer – is leading to groundbreaking insights and continued findings into the role genomics can play in healthcare.

## **Highlighted Pan-Cancer findings:**

Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature*, 2022

Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science*, 2022

Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* 2022.



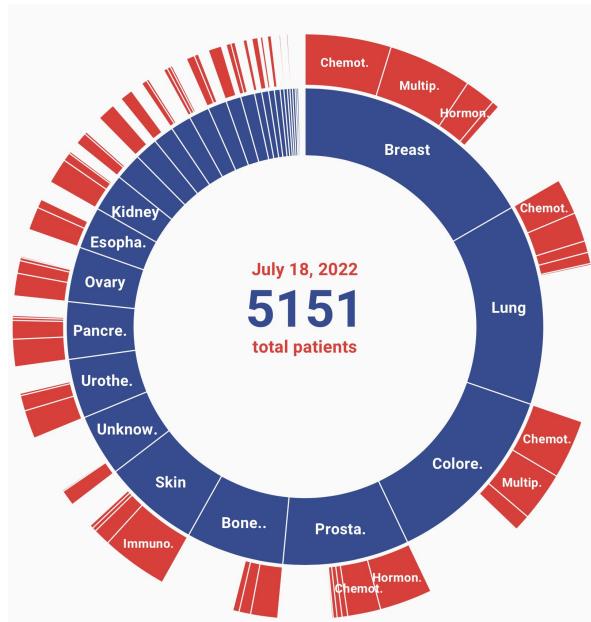
## **The GEL cohort:**

**15, 838** tumor-normal sample pairs

High-quality data derived from flash-frozen material, involving 12,222 GEL tumor samples from 11,585 individuals (several participants had synchronous or metachronous tumors).

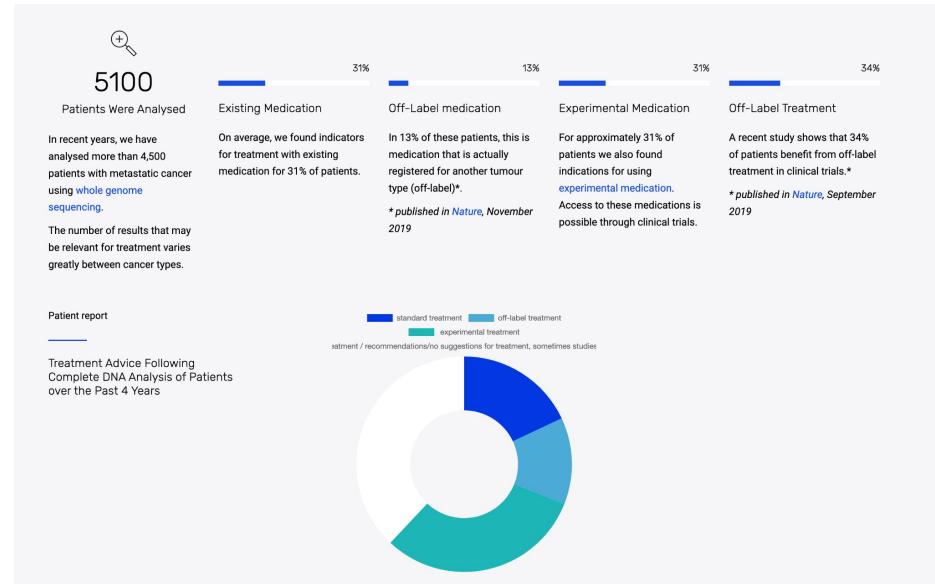
**19 tumor types:** skin, lung, stomach, colorectal, bladder, liver, uterus, ovary, biliary, kidney, pancreas, breast, prostate, bone and soft tissue, central nervous system (CNS), lymphoid, oropharyngeal, neuroendocrine tumors (NETs), and myeloid

# WGS: Hartwig Medical Foundation



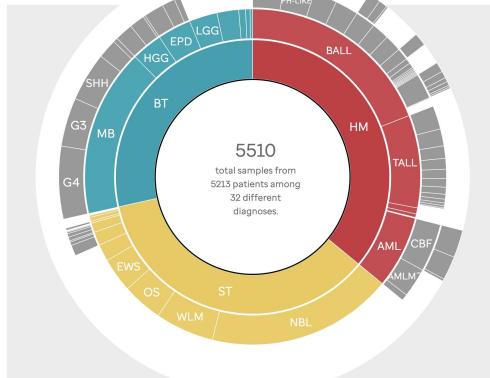
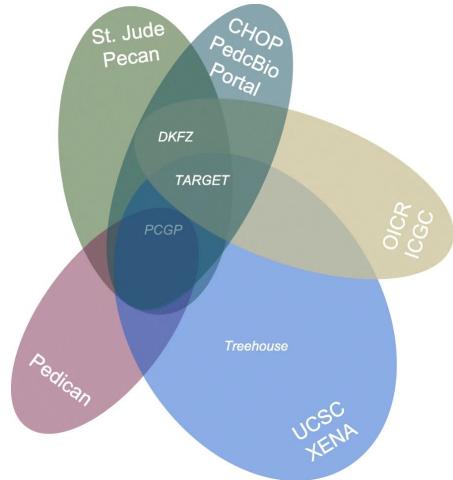
This is the largest database of **metastatic tumor data** obtained with WGS in the world. The uniqueness of the database consists of combining genetic data with treatment and treatment outcome data.

<https://database.hartwigmedicalfoundation.nl>



# WGS: Pediatric Cancer Genome Project (PCGP) and Pediatric Cancer Databases (PeCan)

- The pediatric cancer genome project (PCGP) is a collaborative project created by St. Jude Children's Research Hospital and Washington University School of Medicine. The originally provided data portal "PCGP explore" was based on whole genome sequencing of pediatric tumors with the aim to cover the full spectrum of mutations in pediatric cancers [1]. PCGP is now part of St. Jude PeCan data portal.
- The Pan-Cancer Study of Childhood Cancers (PedPanCan) by the DKFZ includes various sources like ICGC Pedbrain Tumor, PCGP and from Heidelberg and others, and has been integrated into St. Jude PeCan.



5,510 Samples
5,213 Patients
32 Diagnoses
20,537 Genes
126,414 Mutations

# St. Jude Cloud Genomics Platform

St. Jude Cloud Genomics Platform Sign in ☰

**Select Data**

**Diagnoses**  
Samples grouped by primary diagnosis

**Publications**  
Samples grouped by publication

**Studies**  
Datasets curated by St. Jude

**Samples**  
All samples curated by St. Jude

**Filter Selections**

**SEQUENCING TYPES**  
Multiple RNA-Seq  
WES WGS

**FILE TYPES**  
BAM CNV  
Feature Counts gVCF  
Somatic VCF

**SAMPLE TYPES**  
Autopsy Cell Line  
Diagnosis Germline  
Metastasis Relapse  
Xenograft

**Share Selection**

Tumor	Paired Tumor-Normal	Germline	Search	Search Exact Phrase	
<input type="checkbox"/> Primary Diagnosis	Sequencing Types	File Types	Samples	Total Files	Total File Size
<input type="checkbox"/> Acoustic Neuroma	WGS WES RNA-Seq	BAM gVCF Feature Counts	3	23	188.48 GB
<input type="checkbox"/> Acute Leukemias of Ambiguous Lineage	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF Feature Counts	12	125	1.96 TB
<input type="checkbox"/> Acute Lymphoblastic Leukemia, NOS	WGS WES RNA-Seq	BAM gVCF Feature Counts	2	14	127.06 GB
<input type="checkbox"/> Acute Megakaryoblastic Leukemia	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF CNV Feature Counts	134	744	4.97 TB
<input type="checkbox"/> Acute Megakaryoblastic Leukemia, KMT2A rearrangement	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF Feature Counts	18	105	823.19 GB
<input type="checkbox"/> Acute Monoblastic/Monocytic Leukemia, KMT2A rearrangement	WGS WES RNA-Seq	BAM gVCF Feature Counts	2	22	352.16 GB
<input type="checkbox"/> Acute Myeloid Leukemia	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF Feature Counts	260	2,245	31.53 TB
<input type="checkbox"/> Acute Myeloid Leukemia, CEBPA alteration	WGS WES RNA-Seq	BAM gVCF Feature Counts	3	33	771.16 GB
<input type="checkbox"/> Acute Myeloid Leukemia, Core Binding Factor	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF CNV Feature Counts	249	1,444	11.76 TB
<input type="checkbox"/> Acute Myeloid Leukemia, KMT2A rearrangement	WGS WES RNA-Seq Multiple	BAM gVCF Somatic VCF Feature Counts	80	587	8.58 TB

Page 1 of 29 < >

# St. Jude Cloud Visualization Community

ProteinPaint Gene, position, or SNP human hg38 hg38 genome browser Apps Help Code updated: Thu Jun 30 2022; server launched: Thu Jun 30 2022 11:36:22 Running BLAT Copyright: St. Jude Children's Research Hospital Our Team Google Group Licensing

Use Cases  
find workflows and processes for specific needs

DNAxenix File Viewer  
Instructions on launching ProteinPaint from DNAxenix

Featured Datasets

Pediatric Cancer Mutation Pediatric2 GenomePaint NCI GDC CIVIC Survivorship Study

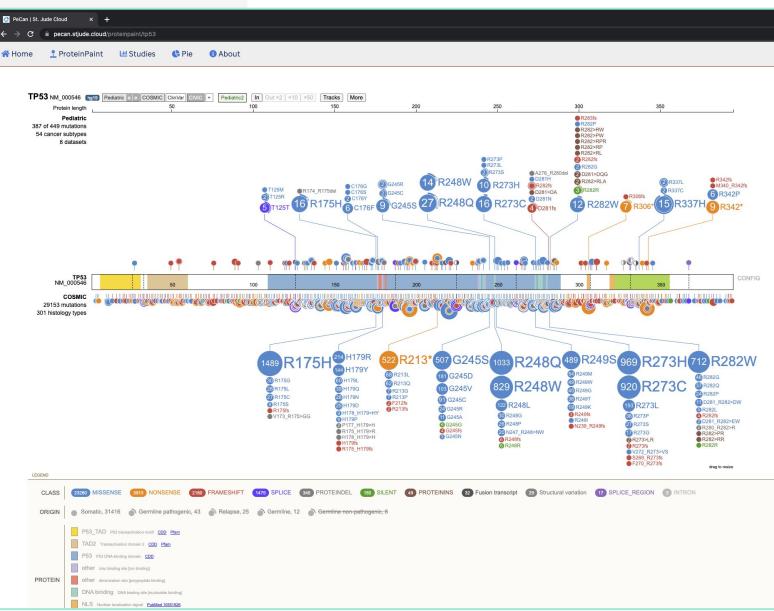
Launch Apps

Genome Browser Tracks

GenomePaint Explore coding and non-coding variants in cancer	BigWig Quantitative data at genomic positions	JSON BED Genomic feature annotation
Allelic Imbalance From tumor and germline DNA	Profile gene value Profiling results & gene-level values	Expression rank Against all samples from a cohort
Hi-C Chromatin interaction at a locus	Lollipop Coding mutations and gene fusions	Splice Junction Single sample and cohort
Track list & facet table	Arc Track Pairwise chromatin interactions	ASE Allele-specific expression analysis
BAM Sequence reads alignment	GDC BAM slicing Visualize sequencing reads in NCI GDC	Additional Track Features ProteinPaint arguments applicable to any track

TP53 MA\_00004 Protein high Pediatric COADREAD Colon - Pediatric 387 of 449 mutations 34 cancer types 6 datasets

# ProteinPaint



# WES: The Cancer Genome Atlas (TCGA)

Analysis platforms supporting TCGA dataset:

- Query based platforms: cBioPortal, FireBrowse, GEPIA, Genomic Data Commons, UCSC, TCGA Batch Effects Viewer, Tumor Map, TANRIC, SurvNet, Regulome Explorer, Xena
- Cloud-based platforms: Seven Bridges, Terra, DNAnexus etc.
- Programming-based packages: TCGAbiolinks, RTCGA etc.

NATIONAL CANCER INSTITUTE  
THE CANCER GENOME ATLAS

## TCGA BY THE NUMBERS

TCGA produced over

**2.5**  
PETABYTES  
of data

To put this into perspective, 1 petabyte of data is equal to

**212,000**  
DVDs



TCGA data describes

**33**  
DIFFERENT  
TUMOR TYPES  
...including  
**10**  
RARE  
CANCERS

...based on paired tumor and normal tissue sets collected from

**11,000**  
PATIENTS  
...using

**7**  
DIFFERENT  
DATA TYPES  


## TCGA RESULTS & FINDINGS



MOLECULAR  
BASIS OF  
CANCER

Improved our  
understanding of  
the genomic underpinnings  
of cancer



TUMOR  
SUBTYPES

Revolutionized how  
cancer is classified



THERAPEUTIC  
TARGETS

Identified genomic  
characteristics of tumors  
that can be targeted with  
currently available  
therapies or used to help  
with drug development

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the same subtype found in other tissues at the molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.\*

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

## THE TEAM



**20**  
COLLABORATING  
INSTITUTIONS

across the United States  
and Canada

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



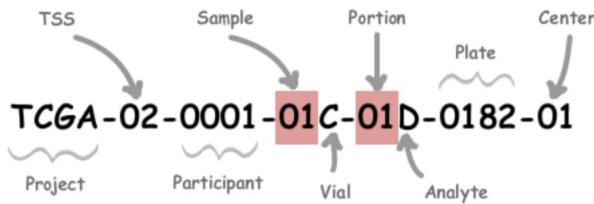
## WHAT'S NEXT?

\*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

# TCGA Cancers



## TCGA Barcode



Cancer Type Studied	# Cases Characterized	Publication	Study Abbreviation
<a href="#">Acute Myeloid Leukemia</a>	200 (200)	NEJM 2013	LAML
<a href="#">Adrenocortical Carcinoma</a>	92 (91)	Cancer Cell 2016	ACC
<a href="#">Bladder Urothelial Carcinoma</a>	412 (131)	Nature 2014, Cell 2017	BLCA
<a href="#">Breast Ductal Carcinoma</a>	778 (430)	Nature 2012	BRCA/DCIS
<a href="#">Breast Lobular Carcinoma</a>	201 (127)	Cell 2015	BRCA/LCIS
<a href="#">Cervical Carcinoma</a>	307 (228)	Nature 2017	CESC
<a href="#">Cholangiocarcinoma</a>	51 (38)	Cell Reports 2017	CHOL
<a href="#">Colorectal Adenocarcinoma</a>	633 (276)	Nature 2012	COAD
<a href="#">Esophageal Carcinoma</a>	185 (164)	Nature 2017	ESCA
<a href="#">Gastric Adenocarcinoma</a>	443 (295)	Nature 2014	STAD
<a href="#">Glioblastoma Multiforme</a>	617 (206)	Nature 2008, Cell 2013	GBM
<a href="#">Head and Neck Squamous Cell Carcinoma</a>	528 (279)	Nature 2015	HNSC
<a href="#">Hepatocellular Carcinoma</a>	377 (363)	Cell 2017	LIHC
<a href="#">Kidney Chromophobe Carcinoma</a>	113 (66)	Cancer Cell 2014	KICH
<a href="#">Kidney Clear Cell Carcinoma</a>	537 (446)	Nature 2013	KIRC
<a href="#">Kidney Papillary Cell Carcinoma</a>	291 (161)	NEJM 2016	KIRP
<a href="#">Lower Grade Gioma</a>	516 (293)	NEJM 2015	LGG
<a href="#">Lung Adenocarcinoma</a>	585 (230)	Nature 2014, Nature Genetics 2016	LUAD
<a href="#">Lung Squamous Cell Carcinoma</a>	504 (178)	Nature 2012, Nature Genetics 2016	LUSC
<a href="#">Mesothelioma</a>	74 (87)	Cancer Discovery 2018	MESO
<a href="#">Ovarian Serous Adenocarcinoma</a>	608 (489)	Nature 2011	OV
<a href="#">Pancreatic Ductal Adenocarcinoma</a>	185 (150)	Cancer Cell 2017	PAAD
<a href="#">Paraganglioma &amp; Pheochromocytoma</a>	179 (173)	Cancer Cell 2017	PCPG
<a href="#">Prostate Adenocarcinoma</a>	500 (333)	Cell 2015	PRAD
<a href="#">Sarcoma</a>	261 (206)	Cell 2017	SARC
<a href="#">Skin Cutaneous Melanoma</a>	470 (331)	Cell 2015	SKCM
<a href="#">Testicular Germ Cell Cancer</a>	150 (137)	Cell Reports 2018	TGCT
<a href="#">Thymoma</a>	124 (117)	Cancer Cell 2018	THYM
<a href="#">Thyroid Papillary Carcinoma</a>	507 (496)	Cell 2014	THCA
<a href="#">Uterine Carcinosarcoma</a>	57 (57)	Cancer Cell 2017	UCS
<a href="#">Uterine Corpus Endometrioid Carcinoma</a>	560 (373)	Nature 2013	UCEC
<a href="#">Uveal Melanoma</a>	80 (80)	Cancer Cell 2017	UVM

# TCGA Molecular Characterization Platforms

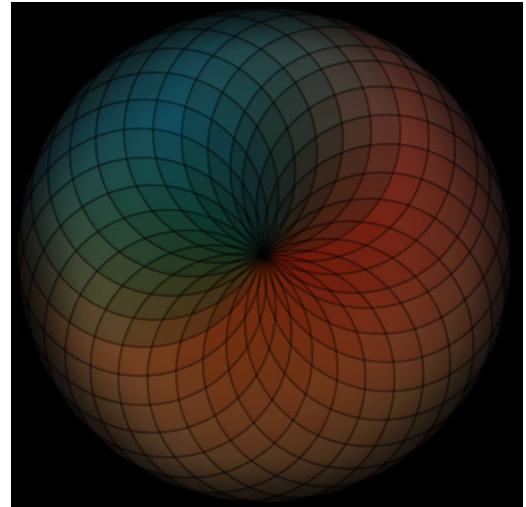
Center	TCGA Platform Code	DCC PlatformName	Instrument Support Materials	Sequence Download	TCGA ADF Download
Broad Institute of MIT and Harvard	ABI	Applied Biosystems Sequence Data	<a href="#">3730/3730xl DNA Analyzers</a>	<a href="#">Primers</a>	N/A
McDonnell Genome Institute at Washington University	ABI	Applied Biosystems Sequence Data	<a href="#">3730/3730xl DNA Analyzers</a>	<a href="#">Primers</a>	N/A
Human Genome Sequencing Center at Baylor College of Medicine	ABI	Applied Biosystems Sequence Data	<a href="#">3730/3730xl DNA Analyzers</a>	<a href="#">Primers</a>	N/A
University of North Carolina	AgilentG4502A_07_1	Agilent 244K Custom Gene Expression G4502A-07-1	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
University of North Carolina	AgilentG4502A_07_2	Agilent 244K Custom Gene Expression G4502A-07-2	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
University of North Carolina	AgilentG4502A_07_3	Agilent 244K Custom Gene Expression G4502A-07-3	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Memorial Sloan Kettering Cancer Center	CGH-1x1M_G4447A	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Broad Institute of MIT and Harvard	Genome_Wide_SNP_6	Affymetrix Genome-Wide Human SNP Array 6.0	<a href="#">Genome-Wide Human SNP Array 6.0</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
McDonnell Genome Institute at Washington University	Genome_Wide_SNP_6	Affymetrix Genome-Wide Human SNP Array 6.0	<a href="#">Genome-Wide Human SNP Array 6.0</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
University of North Carolina	H-miRNA_8x15K	Agilent 8 x 15K Human miRNA-specific microarray	<a href="#">Human mouse and rat miRNA Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
University of North Carolina	H-miRNA_8x15Kv2	Agilent Human miRNA Microarray Rel12.0	<a href="#">Human mouse and rat miRNA Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Memorial Sloan Kettering Cancer Center	HG-CGH-244A	Agilent Human Genome CGH Microarray 244A	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Harvard Medical School	HG-CGH-244A	Agilent Human Genome CGH Microarray 244A	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Harvard Medical School	HG-CGH-415K_G4124A	Agilent Human Genome CGH Custom Microarray 2x415K	<a href="#">SurePrint G3 CGH+SNP Microarray</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
McDonnell Genome Institute at Washington University	HG-U133_Plus_2	Affymetrix Human Genome U133 Plus 2.0 Array	<a href="#">Human Genome U133 Array</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Broad Institute of MIT and Harvard	HT_HG-U133A	Affymetrix HT Human Genome U133 Array Plate Set	<a href="#">Human Genome U133 Array</a>	<a href="#">TSV</a>	<a href="#">TCGA ADF</a>
Berkeley Lab	HuEx-1_0-st-v2	Affymetrix Human Exon 1.0 ST Array	<a href="#">Exon 1.0 ST Array</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
HudsonAlpha Institute for Biotechnology	Human1MDuo	Illumina Human1M-Duo BeadChip	<a href="#">Infinium HD BeadChip DNA</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
HudsonAlpha Institute for Biotechnology	HumanHap550	Illumina 550K Infinium HumanHap550 SNP Chip	<a href="#">Infinium BeadChip DNA</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Johns Hopkins-USC Epigenome Center	HumanMethylation27	Illumina Infinium Human DNA Methylation 27	<a href="#">Infinium Human DNA Methylation 27</a>	<a href="#">See ADF File</a>	<a href="#">TCGA ADF</a>
Johns Hopkins-USC Epigenome Center	HumanMethylation450	Illumina Infinium Human DNA Methylation 450	<a href="#">Infinium Human DNA Methylation 450</a>	<a href="#">See ADF File</a>	<a href="#">TCGA ADF</a>
Johns Hopkins-USC Epigenome Center	IlluminaDNAMethylation_OMA002_CPI	Illumina DNA Methylation OMA002 Cancer Panel I	<a href="#">GoldenGate DNA Methylation</a>	<a href="#">FASTA</a>	<a href="#">TCGA ADF</a>
Johns Hopkins-USC Epigenome Center	IlluminaDNAMethylation_OMA003_CPI	Illumina DNA Methylation OMA003 Cancer Panel I	<a href="#">GoldenGate DNA Methylation</a>	<a href="#">See ADF File</a>	<a href="#">TCGA ADF</a>
McDonnell Genome Institute at Washington University	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
Broad Institute of MIT and Harvard	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
Human Genome Sequencing Center at Baylor College of Medicine	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
University of California Santa Cruz	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
University of North Carolina	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
BC Cancer Agency	IlluminaGA_miRNASeq	Illumina Genome Analyzer miRNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
Harvard Medical School	IlluminaGA_mRNA_DGE	Illumina Genome Analyzer mRNA Digital Gene Expression	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
BC Cancer Agency	IlluminaGA_RNASEq	Illumina Genome Analyzer RNA Sequencing	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
University of North Carolina	IlluminaGA_RNASeqV2	Illumina Genome Analyzer RNA Sequencing Version 2 analysis	<a href="#">Genome Analyzer IIx</a>	N/A	N/A
Harvard Medical School	IlluminaHiSeq_DNASeq	Illumina HiSeq for Copy Number Variation	<a href="#">HiSeq 2000</a>	N/A	N/A
BC Cancer Agency	IlluminaHiSeq_miRNASeq	Illumina HiSeq 2000 miRNA Sequencing	<a href="#">HiSeq 2001</a>	N/A	N/A
University of North Carolina	IlluminaHiSeq_RNASeq	Illumina HiSeq 2000 RNA Sequencing	<a href="#">HiSeq 2002</a>	N/A	N/A
University of North Carolina	IlluminaHiSeq_RNASeqV2	Illumina HiSeq 2000 RNA Sequencing Version 2 analysis	<a href="#">HiSeq 2003</a>	N/A	N/A
University of North Carolina	IlluminaHiSeq_TotalRNASeqV2	Illumina HiSeq 2000 Total RNA Sequencing Version 2 analysis	<a href="#">HiSeq 2004</a>	N/A	N/A
Johns Hopkins-USC Epigenome Center	IlluminaHiSeq_WGBS	Illumina HiSeq 2000 Bisulfite-converted DNA Sequencing	<a href="#">HiSeq 2005</a>	N/A	N/A
MD Anderson Cancer Center	MDA_RPPA_Core	M.D. Anderson Reverse Phase Protein Array Core	<a href="#">MD Anderson RPPA Core Facility</a>	<a href="#">N/A</a>	<a href="#">N/A</a>
Nationwide Children's Hospital	microsat_i	Microsatellite Instability Analysis	<a href="#">SOLID 3 Plus</a>	N/A	N/A
University of California Santa Cruz	SOLID_DNASeq	ABI SOLID DNA System Sequencing	<a href="#">SOLID 3 Plus</a>	N/A	N/A
Broad Institute of MIT and Harvard	SOLID_DNASeq	ABI SOLID DNA System Sequencing	<a href="#">SOLID 3 Plus</a>	N/A	N/A
Human Genome Sequencing Center at Baylor College of Medicine	SOLID_DNASeq	ABI SOLID DNA System Sequencing	<a href="#">SOLID 3 Plus</a>	N/A	N/A

# TCGA Findings

TCGA outcomes & impact

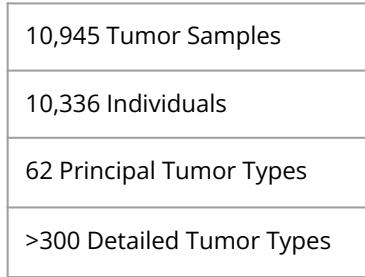
TCGA's Pan-Cancer Atlas

TCGA Research Network Publications

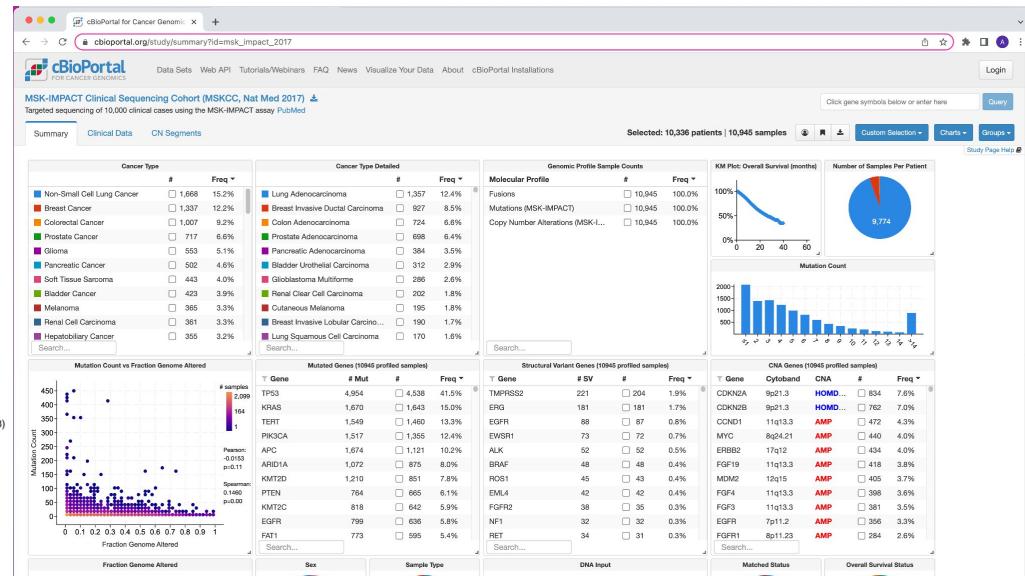
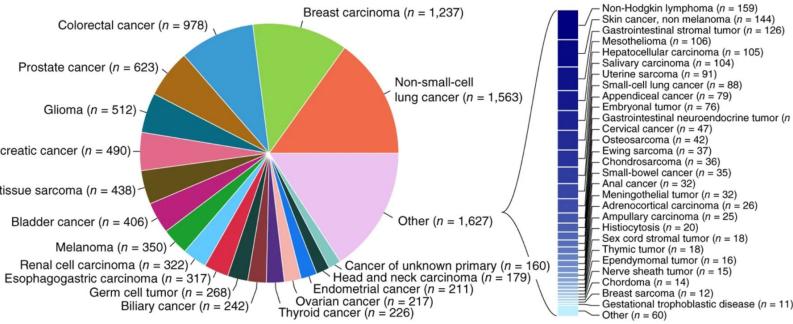


# TS:MSK-IMPACT

- Hybridization capture-based NGS panel capable of detecting protein-coding mutations, copy number alterations (CNAs), and selected promoter mutations and structural rearrangements in **341 (and, more recently, 410) cancer-associated genes**
- Full data set publicly available through cBioPortal for Cancer Genomics: <http://cbioportal.org/msk-impact>



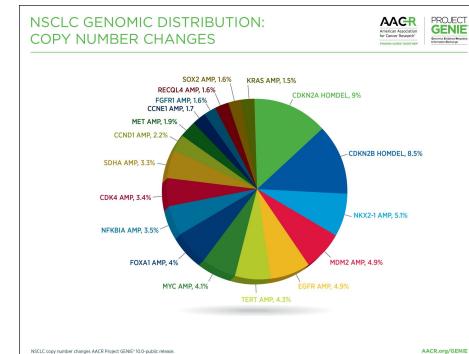
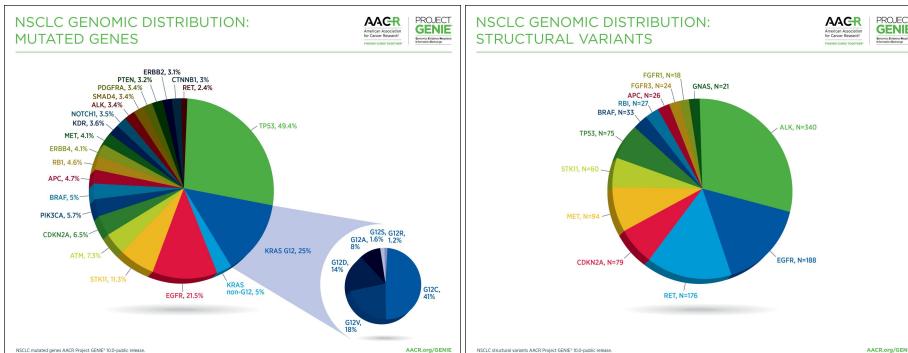
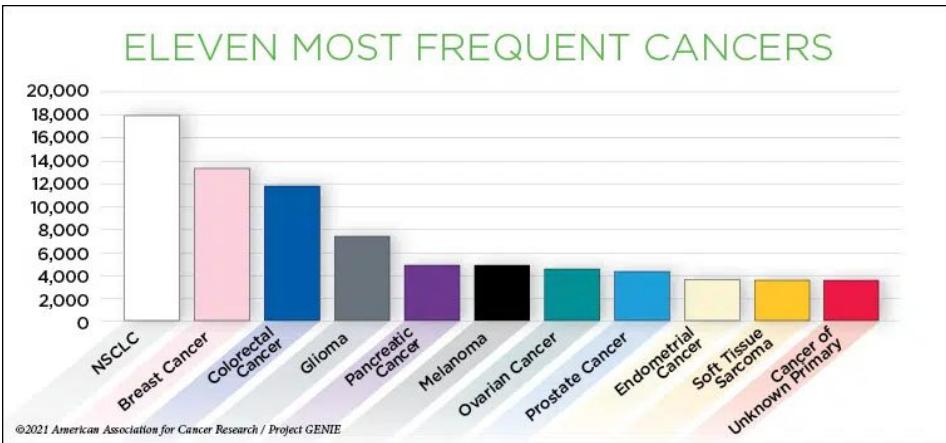
a



# TS: Genomics Evidence Neoplasia Information Exchange (GENIE)

The AACR Project GENIE is an international data-sharing consortium focused on generating an evidence base for precision cancer medicine by integrating clinical-grade cancer genomic data with clinical outcome data for tens of thousands of cancer patients treated at multiple institutions worldwide.

The first public data release was available to the global community in January 2017; our current release, **GENIE 11.0-public**, was released in January 2022. The registry now contains over **136,000 sequenced samples** from over **121,000 patients**, making the AACR Project GENIE registry among the largest fully public cancer genomic data sets released to date.



Data can be access via [cBioPortal](#) or download data directly from [Sage Bionetworks](#).

# Major Cancer Genomic Studies with Data Portal Available

## Other Data Portals



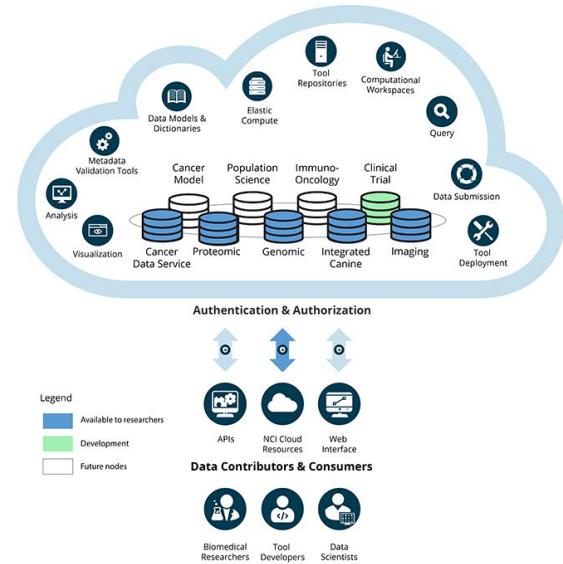
# Cancer Research Data Commons (CRDC) at NCI

## Key Datasets

Dataset Name	Description	Available Resources
The Cancer Genome Atlas (TCGA)	A collaboration between NCI and the National Human Genome Research Institute (NHGRI) that has characterized tumor and normal tissues from 11,000 patients, covering 33 cancer types	GDC, Broad, SB, ISB, IDC
Therapeutically Applicable Research to Generate Effective Treatments (TARGET)	A consortium of extramural and NCI investigators working to characterize and understand hard-to-treat childhood cancers and translate findings into the clinic.	GDC, Broad, SB, ISB
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	A national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics.	GDC, PDC, Broad, ISB, SB, IDC
Human Cancer Model Initiative (HCFI)	An international consortium that is generating novel, next-generation, tumor-derived culture models complete with genomic and clinical data.	GDC, SB, ISB
Cancer Genome Characterization Initiatives (CGCI)	An initiative examining genomes, exomes, and transcriptomes of various types of adult and pediatric cancers.	GDC, SB, ISB
Foundation Medicine (FM) ↗	Targeted sequencing data from ~18,000 adult patients generated by the Foundation Medicine Inc., molecular information company seeking to match patients with personalized treatment plans.	GDC, SB, ISB
Multiple Myeloma Research Foundation (MMRF) ↗	Data from nearly 1,000 patients with extensive molecular and clinical data, including longitudinal information collected over the course of disease for many patients.	GDC, SB, ISB
Genomics Evidence Neoplasia Information Exchange (GENIE) ↗	Over 44,000 cases from the international pan-cancer registry continuing to be collected by the American Association for Cancer Research (AACR) initiative.	GDC, SB, ISB
International Cancer Proteogenomic Consortium (ICPC)	An international consortium that brings together more than a dozen countries to study the application of proteogenomic analysis in predicting cancer treatment success and to share data and results with researchers worldwide, hastening progress for patients.	PDC, SB, ISB
Children's Brain Tumor Tissue Consortium (CBTTC) ↗	A collaborative research consortia focused on identifying therapies for children with brain tumors	PDC, SB, ISB
Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) ↗	A research collaboration to detect colorectal cancer susceptibility loci using genome-wide sequencing	CDS, SB
Comparative molecular life history of spontaneous canine and human gliomas (GLIOMA01) ↗	Characterization of the genomic and transcriptomic landscape of canine glioma to enable cross-species comparative genomic analysis of sporadic glioma	ICDC, SB

CRDC is a cloud-based data science infrastructure that connects data sets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data to drive scientific discovery.

## NCI Cancer Research Data Commons (CRDC)



## Genomic Data Commons (GDC)

## Proteomic Data Commons (PDC)

## Integrated Canine Data Commons (ICD)

## Imaging Data Commons (IDC)

## Cancer Data Commons (CDS)

## NCI Cloud Resources

# GDC Data Portal

NATIONAL CANCER INSTITUTE  
GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart GDC Apps

Harmonized Cancer Datasets  
Genomic Data Commons Data Portal

Get Started by Exploring:  
Projects Exploration Analysis Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 32.0 - March 29, 2022

PROJECTS 70 PRIMARY SITES 67 CASES 85,416 FILES 819,832 GENES 21,754 MUTATIONS 2,670,227

Click to go back, hold to see history

Cases by Major Primary Site

Primary Site	Cases (in thousands)
Adrenal Gland	0.1
Bile Duct	0.1
Bladder	0.2
Brain	0.3
Bone Marrow	0.4
Cervix	0.5
Esophagus	0.6
Eye	0.1
Head and Neck	0.2
Lung	1.1
Lymph Nodes	0.1
Nervous Tissue	0.1
Ovary	0.2
Pancreas	0.1
Prostate	0.2
Soft Tissue	0.1
Stomach	0.1
Thyroid	0.1
Uterus	0.1

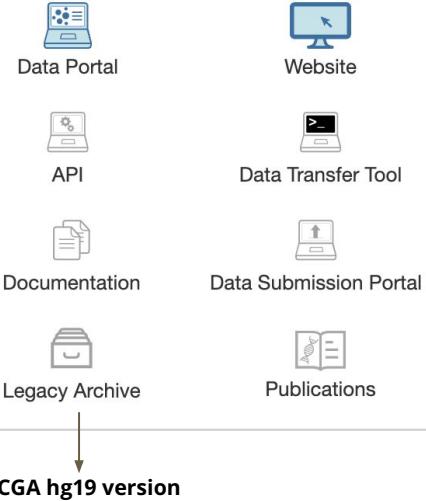
GDC Applications  
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Data Portal Website API Data Transfer Tool Documentation Data Submission Portal Legacy Archive Publications

Site Home Policies Accessibility FOIA Support  
U.S. Department of Health and Human Services National Institutes of Health National Cancer Institute USA.gov  
NIH... Turning Discovery Into Health ©  
UI v1.28.0 @ 07/9/2021 API v3.0.0 @ 07/9/2021 Data Release 32.0 - March 29, 2022

## Bioinformatics pipelines:

- Bioinformatics Pipeline: DNA-Seq Analysis
- Bioinformatics Pipeline: mRNA Analysis
- Bioinformatics Pipeline: miRNA Analysis
- Bioinformatics Pipeline: Copy Number Variation Analysis
- Bioinformatics Pipeline: Methylation Analysis Pipeline
- Bioinformatics Pipeline: Protein Expression



## Analysis:

Set Operations

Cohort Comparison

Clinical Data Analysis

## Visualization:

Cases/Genes/Mutations/Oncogrid

# cBioPortal for Cancer Genomics

## Major features:

- Host multi-omics and clinical data from >340 cancer genomic studies and >30 tissue sites.
- Support many genomic visualization and analyses, including mutational distribution, oncoplot, mutual exclusivity analysis, co-expression, group comparison analysis, survival analysis, integrative analysis, etc.
- All the data are harmonized in the same format and can be directly downloaded from the web or [Datahub](#).
- Provide R client for accessing the study datasets.
- Support OQL & Expression for query.
- Provide link to share the query.
- Support local installation.
- Web-based API (Application Programming Interface) available.

**Exclusively for tumor data !!**

# What data is in cBioPortal?

## Data sources



THE CANCER GENOME ATLAS

National Cancer Institute  
National Human Genome Research Institute



International  
Cancer Genome  
Consortium



CCLE Cancer Cell Line Encyclopedia



PROJECTGENIE  
Genomics Evidence Neoplasia Information Exchange

- Clinical Data:
- Treatments
  - Survival
  - Etc

- omic data:
- Mutations
  - Fusions
  - Copy number
  - mRNA expression
  - Protein levels
  - DNA Methylation

Background biological data  
(e.g. networks, 3D protein structure)

cBioPortal  
for Cancer Genomics

<https://genie.cbiportal.org/>

Curated effect & therapy implications



Precision Oncology Knowledge Base



MY CANCER GENOME®  
GENETICALLY INFORMED CANCER MEDICINE

Predicted functional effect

PolyPhen-2

mutationassessor.org  
functional impact of protein mutations

Variant recurrence



Catalogue of somatic mutations



# Download Datasets

luad\_tcga\_pan\_can\_atlas\_2018.tar.gz



Data Sets Web API R/MATLAB Tutorials/Webinars FAQ News Visualize Your Data About cBioPortal Installations

case_lists			
data_armlevel_cna.txt			
data_clinical_patient.txt			
data_clinical_sample.txt			
data_clinical_supp_hypoxia.txt			
data_cna_hg19.seg			
data_cna.txt			
data_fusions.txt			
data_log2_cna.txt			
data_microbiome.txt			
data_mrna_seq_v2_rsem_zscores_ref_all_samples.txt			
data_mrna_seq_v2_rsem_zscores_ref_diploid_samples.txt			
data_mrna_seq_v2_rsem_zscores_ref_normal_samples.txt			
data_mrna_seq_v2_rsem.txt			
data_mutations.txt			
data_normals_RNA_Seq_v2_mRNA_median_Zscores.txt			
data_normals_RNA_Seq_v2_mRNA_median.txt			
data_rppa_zscores.txt			
data_rppa.txt			
LICENSE			
meta_armlevel_cna.txt			
meta_clinical_patient.txt			
meta_clinical_sample.txt			
meta_cna_hg19_seg.txt			
meta_cna.txt			
meta_fusions.txt			
meta_log2_cna.txt			
meta_microbiome.txt			
meta_mrna_seq_v2_rsem_zscores_ref_all_samples.txt			
meta_mrna_seq_v2_rsem_zscores_ref_diploid_samples.txt			
meta_mrna_seq_v2_rsem_zscores_ref_normal_samples.txt			
meta_mrna_seq_v2_rsem.txt			
meta_mutations.txt			
meta_rppa_zscores.txt			
meta_rppa.txt			
meta_study.txt			
normals			
data_mrna_seq_v2_rsem_normal_samples_zscores_ref_normal_samples.txt			
data_mrna_seq_v2_rsem_normal_samples.txt			
meta_mrna_seq_v2_rsem_normal_samples_zscores_ref_normal_samples.txt			
meta_mrna_seq_v2_rsem_normal_samples.txt			

## Datasets

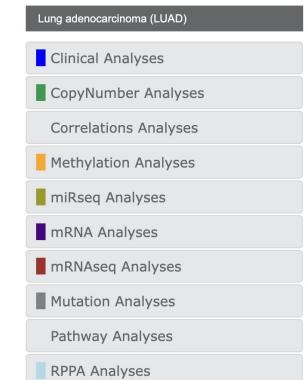
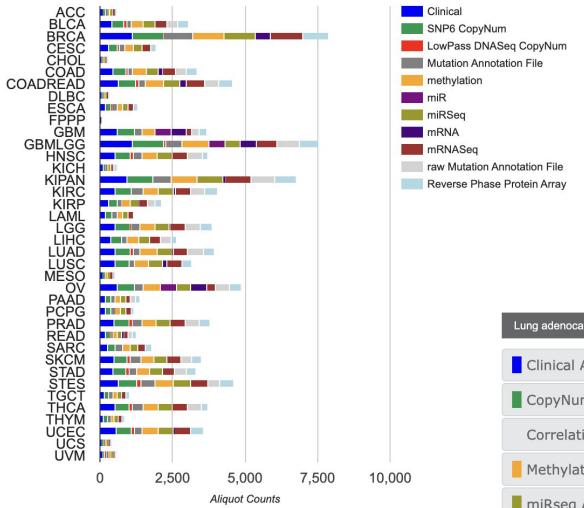
The table below lists the number of available samples per cancer study and data type. It also provides links to download the data for each study. For alternative ways of downloading, see the [Download Documentation](#).

Name	Reference	All	Mutations	CNA	RNA-Seq
Acinar Cell Carcinoma of the Pancreas (UHL, J Pathol 2014)	Jalil et al. J Pathol 2014	23	23	0	0
Acral Melanoma (TGEN, Genome Res 2017)	Liang et al. Genome Res 2017	38	38	38	36
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2015)	Anderson et al. Nat Genet 2015	93	93	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)	Zhang et al. Nat Genet 2016	73	73	0	0
Acute Myeloid Leukemia (OHsu, Nature 2018)	Tyner et al. Nature 2018	672	622	0	451
Acute Myeloid Leukemia (TGCA, Firehose Legacy)		200	197	191	173
Acute Myeloid Leukemia (TGCA, NEJM 2019)		200	200	191	173
Acute Myeloid Leukemia (TGCA, PanCancer Atlas)		200	200	191	173
Acute myeloid leukemia or myelodysplastic syndromes (WashU, J. 2017)		136	136	0	0
Adenoid Cystic Carcinoma (FMI, Am J Surg Pathl, 2014)	Ross et al. Am J Surg Pathl 2014	28	28	28	0
Adenoid Cystic Carcinoma (JHU, Cancer Prev Res 2016)	Retig et al. Cancer Prev Res 2016	25	25	0	0
Adenoid Cystic Carcinoma (MDA, Clin Cancer Res 2015)	Mitani et al. Clin Cancer Res 2015	102	65	0	0
Adenoid Cystic Carcinoma (MGH, Nat Gen 2016)	Drier et al. Nature Genetics 2016	10	10	0	0
Adenoid Cystic Carcinoma (MSKCC, Nat Genet 2013)	Ho et al. Nat Genet 2013	60	60	60	0
Adenoid Cystic Carcinoma (Sanger/MDA, JCI 2013)	Stephens et al. JCI 2013	24	24	0	0
Adenoid Cystic Carcinoma of the Breast (MSKCC, J. Pathol. 2015)	Martelotto et al. J Pathol 2015	12	12	12	0
Adenocarcinoma Project (J Clin Invest 2019)	Allen et al. J Clin Invest 2019	1049	1049	928	0
Adrenocortical Carcinoma (TGCA, Firehose Legacy)		92	90	90	79
Adrenocortical Carcinoma (TGCA, PanCancer Atlas)		92	91	89	78
Adult Soft Tissue Sarcoma (TGCA, Cell 2017)		206	206	206	206
Amputary Carcinoma (Baylor College of Medicine, Cell Reports 2018)	Gingras et al. Cell Rep 2016	160	160	0	0
Anaplastic Oligodendroglioma and Anaplastic Oligoastrocytoma (MSKCC, Neuro Oncol 2017)	Thomas et al. Neuro Oncol 2017	22	22	22	0
Basal Cell Carcinoma (UNIGE, Nat Genet 2016)	Bonilla et al. Nat Genet 2016	293	293	0	0
Bladder Cancer (MSKTCGA, 2020)		476	474	442	296
Bladder Cancer (MSKCC, Eur Urol 2014)	Kim et al. Eur Urol 2015	109	109	109	0
Bladder Cancer (MSKCC, J Clin Oncol 2013)	Iyer et al. J Clin Oncol 2013	97	97	97	0
Bladder Cancer (MSKCC, Nat Genet 2016)	Al-Ahmadi et al. Nat Genet 2016	34	34	33	0
Bladder Cancer (TGCA, Cell 2017)	Robertson et al. Cell 2017	413	412	408	408
Bladder Urothelial Carcinoma (BGI, Nat Genet 2013)	Guo et al. Nat Genet 2013	99	99	0	0
Bladder Urothelial Carcinoma (DFCI/MSKCC, Cancer Discov 2014)	Van Allen et al. Cancer Discov 2014	50	50	0	0
Bladder Urothelial Carcinoma (TGCA, Firehose Legacy)		413	130	408	408
Bladder Urothelial Carcinoma (TGCA, Nature 2014)		131	130	128	129
Bladder Urothelial Carcinoma (TGCA, PanCancer Atlas)		411	410	408	407
Brain Lower Grade Glioma (TGCA, Firehose Legacy)		530	286	513	530
Brain Lower Grade Glioma (TGCA, PanCancer Atlas)		514	514	511	514
Brain Tumor PDXs (Mayo Clinic, 2018)		97	83	83	66
Breast Cancer (HTAN, Cell Rep Med 2022)	Johnson et al. Cell Rep Med. 2022	5	5	5	0
Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)	Pereira et al. Nat Commun 2016, Rueda et al. Nature 2019, Curtis et al. Nature 2012	2509	2509	2173	1904
Breast Cancer (MSK, Cancer Cell 2018)	Razavi et al. Cancer Cell 2018	1918	1918	1918	0
Breast Cancer (MSK, Clinical Cancer Res 2020)		60	60	0	0
Breast Cancer (MSK, Cell Reports 2019)		444	444	444	444

# FIREBROWSE (Broad GDAC)

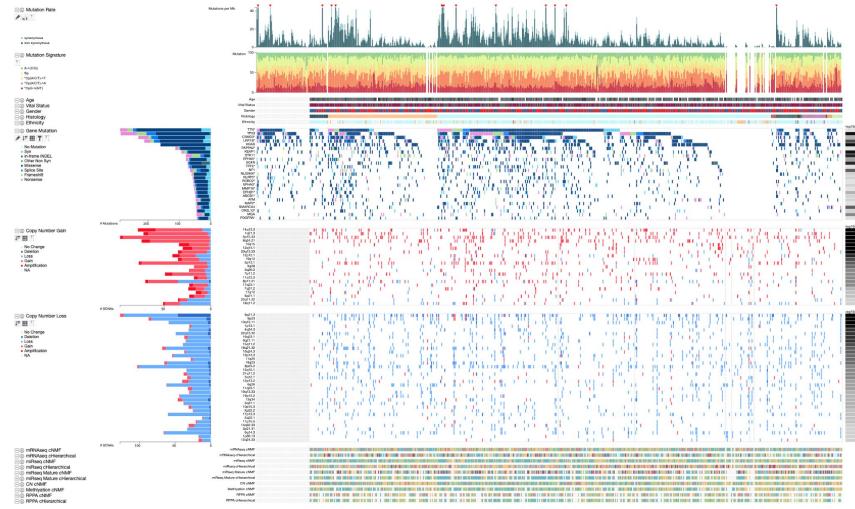
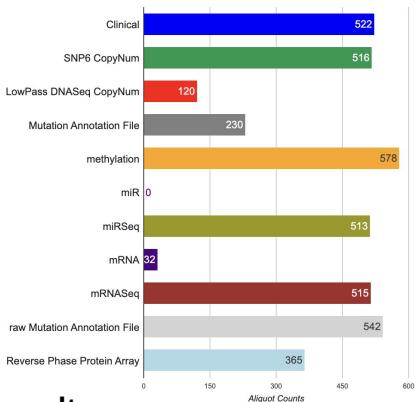
## Dataset: TCGA

TCGA data version 2016\_01\_28



## Major genomic analysis results

TCGA data version 2016\_01\_28 for LUAD



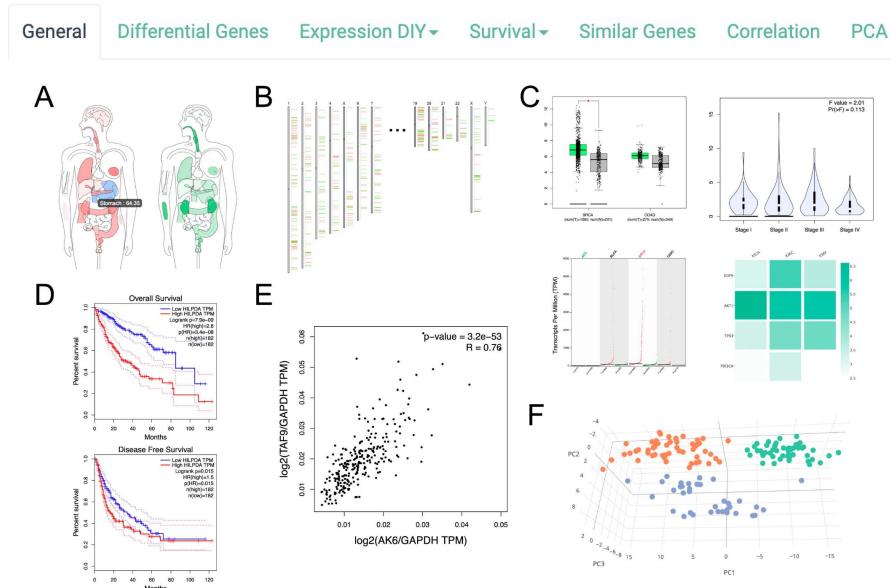
## iCoMut for FireBrowse



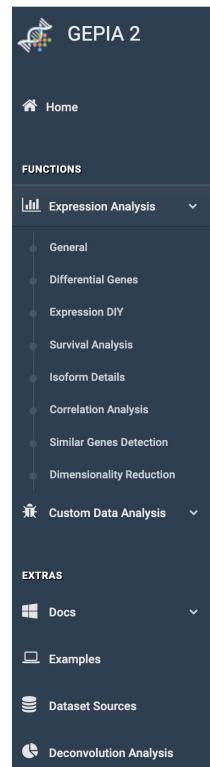
# Gene Expression Profiling Interactive Analysis (GEPIA)

Datasets: **TCGA** and **GTEX** data

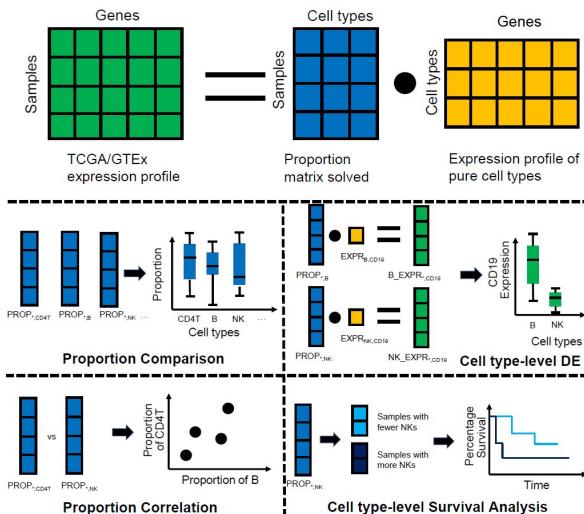
- Single Gene Analysis
- Cancer Type Analysis
- Multiple Gene Analysis



[GEPIA2](#)



[GEPIA2021](#)

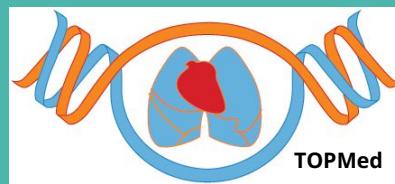
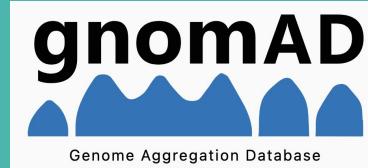


a standalone extension with multiple deconvolution based analysis for GEPIA. We deconvolute each sample tool in TCGA/GTEX with the bioinformatics tools **CIBERSORT**, **EPIC** and **quantiTseq**. Based on the inferred cell proportions in each bulk-RNA sample, we can then perform multiple downstream analysis:

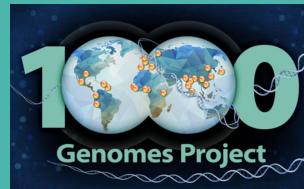
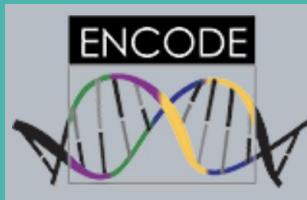
# Summary of Data Portals

Data Portals	Studies	Data type	Interactive analysis	Integrative analysis	Visualization	Access Control	Recommendation
<b>GEPIA</b>	TCGA; GTEx	Expression data	Yes	No	Limited	No	Tumor-Normal expression analysis
<b>GTEx Portal</b>	GTEx	Gene expression and genotyping data	Limited	Limited	Yes	Raw data and most analyzed data	Germline analysis
<b>ICGC Portal</b>	PCAWG and other ICGC Data	Multi-Omics Data	Limited	No	No	Raw data and most analyzed data	Data download
<b>GDC</b>	>20 Studies (e.g. TCGA, GENIE)	Multi-Omics Data	Limited	No	Limited	Raw data and most analyzed data	Data download
<b>Firebrowse</b>	TCGA	Multi-Omics Data	No	Yes	Limited	No	TCGA deep analysis
<b>St. Jude Cloud</b>	PeCan	Multi-Omics Data	No	No	Yes	Yes	Data download and visualization
<b>cBioPortal</b>	>340 Studies (e.g. TCGA, IMPACT, GENIE, PCAW, ICGC)	Multi-Omics Data	Yes	Yes	Yes	No	Integrative analysis

# Other Genetic Data Resources



Cell Model Passports  
A Hub for Preclinical Cancer Models



# Genotype-Tissue Expression (GTEx)

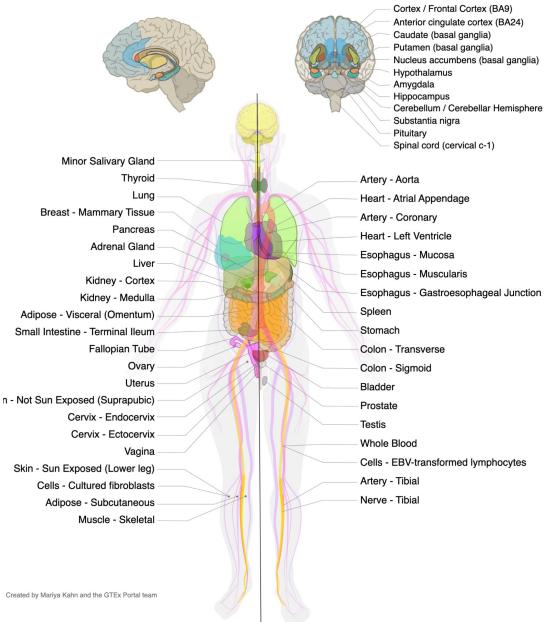
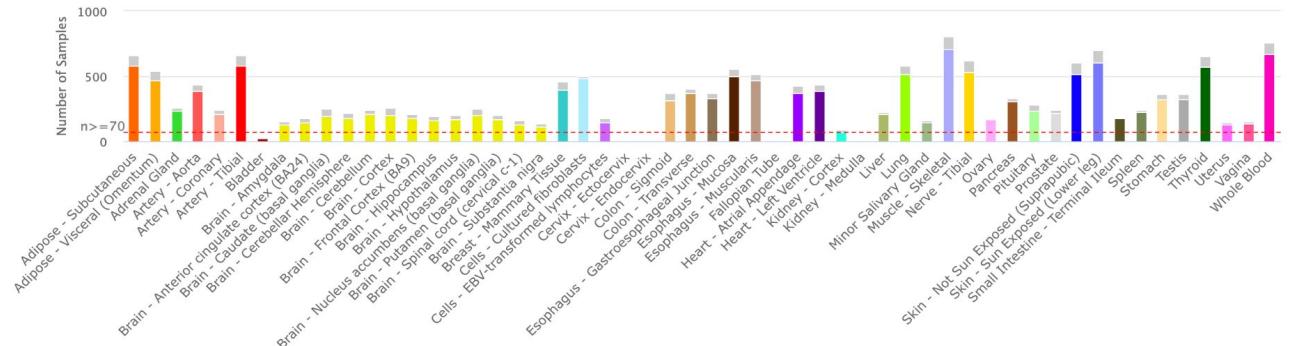
Public resource of tissue-specific gene expression

Samples collected from **54** non-diseased tissue sites across over 900 individuals

Datasets include SNP array, WGS, WES, bulk RNA-Seq and snRNA-Seq

Available protected access data include:

- BAM files for RNA-Seq, Whole Exome Seq, and Whole Genome Seq
- Genotype Calls (.vcf) for WES and WGS
- Allele Specific Expression (ASE) tables
- All expression matrices from the Portal, including samples that did not pass the Analysis Freeze QC
- Expanded sample attributes
- Expanded subject phenotypes, including age and ethnicity



Created by Marily Kahn and the GTEx Portal team

V8 Release	# Tissues	# Donors	# Samples
Total	54	948	17382
With Genotype	54	838	15253
Has eQTL Analysis*	49	838	15201

\* Number of samples with genotype >= 70

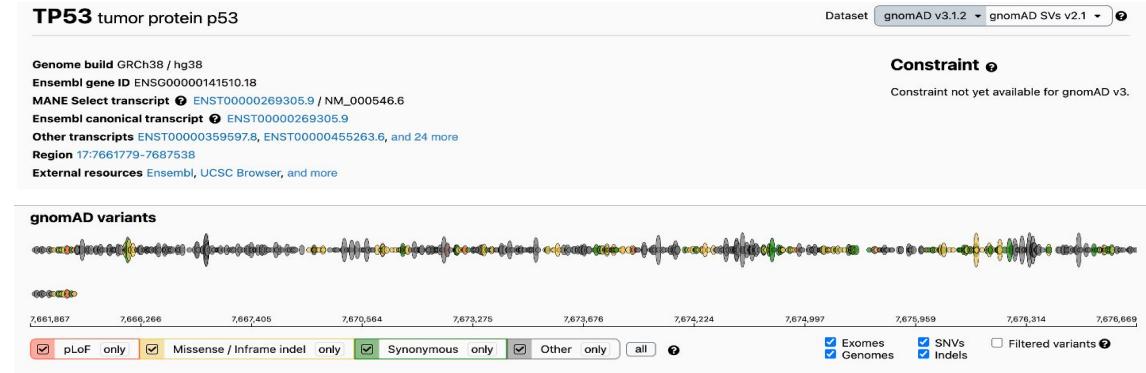
# gnomAD: Genome Aggregation Database



gnomAD is a coalition of investigators seeking to aggregate and harmonize exome and genome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

Available access data include:

- Variants & Coverage &Constraint
- Multi-nucleotide variants (MNVs)
- Proportion expressed across transcripts
- Structural variants
- Loss-of-function curation results
- Variant co-occurrence
- Linkage disequilibrium
- Ancestry classification



Different Version:

- gnomAD v2.1 data set contains data from **125,748 exomes** and **15,708 whole genomes**, all mapped to the **GRCh37/hg19** reference sequence.
- gnomAD v3.1 data set contains **76,156 whole genomes**, all mapped to the **GRCh38** reference sequence.
- gnomAD v3.1 contains a substantially larger number of African American samples than v2.1 and provides allele frequencies in the Amish population for the first time. gnomAD v3.1 also has a fully genotyped callset available from the Human Genome Diversity Project and 1000 Genomes Project, representing > 60 distinct populations.

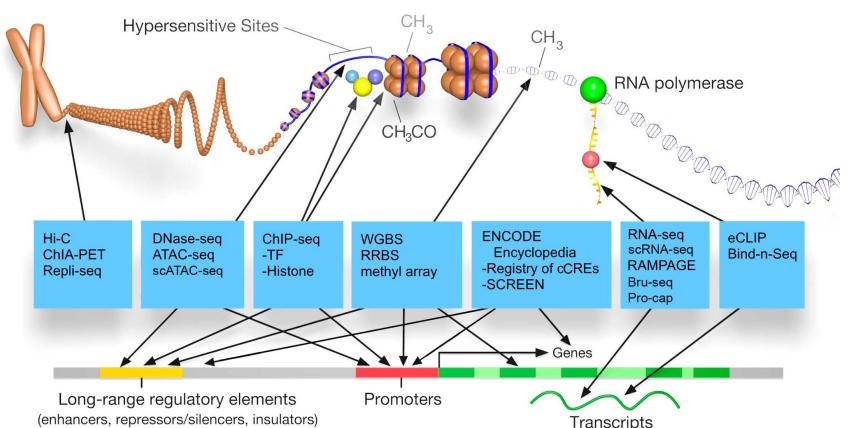
# EnCODE and Portal

ENCODE Encyclopedia Version 5:

The ENCODE Consortium not only produces high-quality data, but also analyzes the data in an integrative fashion. The ENCODE Encyclopedia organizes the most salient analysis products into annotations and provides tools to search and visualize them. The

Encyclopedia has two levels of annotations:

- Integrative-level annotations integrate multiple types of experimental data and ground level annotations.
  - Ground-level annotations are derived directly from the experimental data, typically produced by uniform processing pipelines.

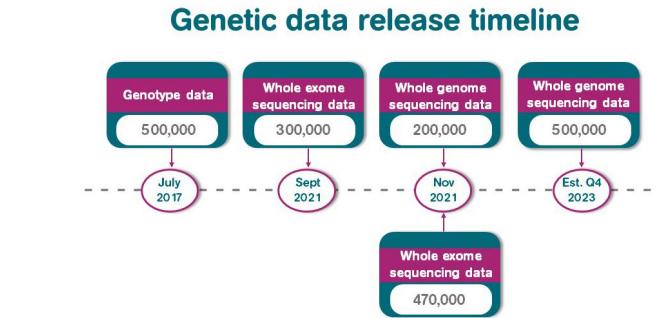


Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

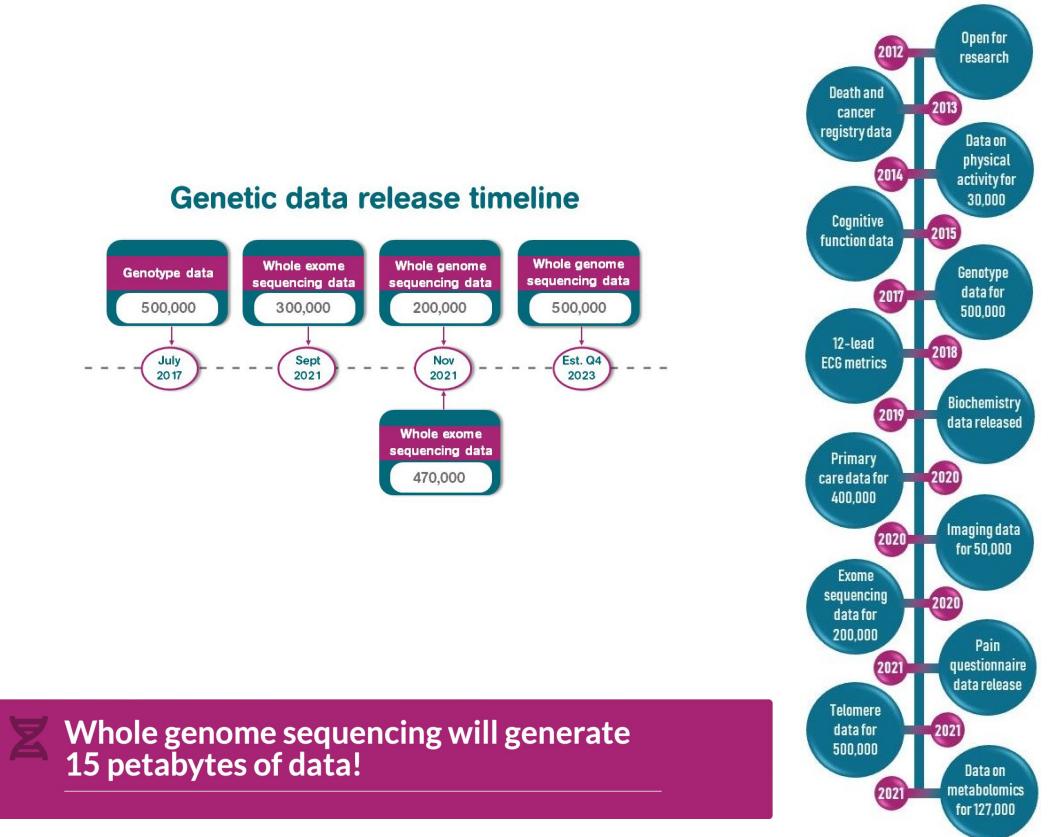
UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

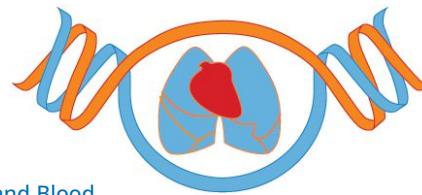
## Data available timeline

## Research Analysis Platform (RAP)



Whole genome sequencing will generate 15 petabytes of data!





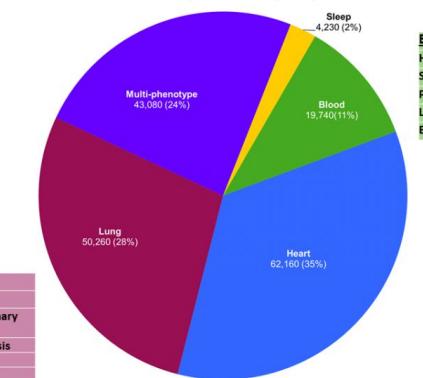
The **Trans-Omics for Precision Medicine** (TOPMed) program, sponsored by the **National Institutes of Health (NIH) National Heart, Lung and Blood Institute (NHLBI)**, is part of a broader **Precision Medicine Initiative**, which aims to provide disease treatments tailored to an individual's unique genes and environment. TOPMed contributes to this Initiative through the integration of whole-genome sequencing (WGS) and other omics (e.g., metabolic profiles, epigenomics, protein and RNA expression patterns) data with molecular, behavioral, imaging, environmental, and clinical data.

## TOPMed WGS and Omics Summary of Approved Data:

Short Name	Study/Cohort name	PI	Populations	dbGaP ID	WGS	RNA-seq	Methylation	Metabolomics	Proteomics
ATGC	Asthma Translational Genomics Collaborative	Burchard; Esteban; Williams, L.; Keoki;		ATGC dbGaP IDs	16,494	9,290			
MESA	Multi-Ethnic Study of Atherosclerosis	Rotter, Jerome; Rich, Stephen	Multi-ethnic populations	phs001416	7,107	8,903	2,086	12,800	14,200
HCHS_SOL	Hispanic Community Health Study - Study of Latinos	Kaplan, Robert; North, Kari		phs001395	7,834	7,733		12,226	
ARIC+VTE	Venous Thromboembolism project	Boerwinkle, Eric	20% African American	phs001211 phs001402 phs000993	10,531	6,111	16,524		
CARDIA	Cell Disease Whole Genome Sequence Analysis in Early Cerebral Small Vessel Disease	Fornage, Myriam; Hou Lifang		phs001612	3,472	6,000	9,480	9,000	8,000
MLOF	My Life, Our Future: Genotyping for Progress in Hemophilia	Konkle, Barbara; Johnsen, Jill		phs001515	5,670	4,500			
PVDOMICS	Pulmonary Vascular Disease Omics Analyses	Erzurum, Serpil; Barnard, John; Geraci, Marc; Beck, Gerald; Comhair, Suzy		phs002358	1,137	4,388	1,800		
SPIROMICS	SubPopulations and Intermediate Outcome Measures In COPD Study	Meyers, Deborah A		phs001927	2,711	3,980		3,417	
.....									
<b>TOTAL</b>					<b>205,092</b>	<b>68,219</b>	<b>56,987</b>	<b>60,769</b>	<b>25,748</b>

## Phenotype Focus

Phases 1-7 (~180K Participants)



**Blood:**  
Hemophilia  
Sickle Cell Disease  
Platelets  
Lipids  
Blood Cancers

**Heart:**  
Hypertension  
Myocardial Infarction  
Coronary Artery Disease  
Stroke  
Small Vessel Disease  
Venous Thromboembolism  
Congenital Heart Disease  
Atrial Fibrillation  
Coronary Artery Calcification  
Adiposity  
Congestive Heart Failure

**Lung:**  
Asthma  
Chronic Obstructive Pulmonary Disease  
Idiopathic Pulmonary Fibrosis  
Sarcoidosis  
Interstitial Lung Disease

A primary goal of the TOPMed program is to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders. TOPMed is providing deep WGS and other omics data to pre-existing '**parent**' studies having large samples of human subjects with rich phenotypic characterization and environmental exposure data.

# 1000 Genomes

The 1000 Genomes Project created a catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy. The reference data resources generated by the project remain heavily used by the biomedical science community.

## Overview of data collection

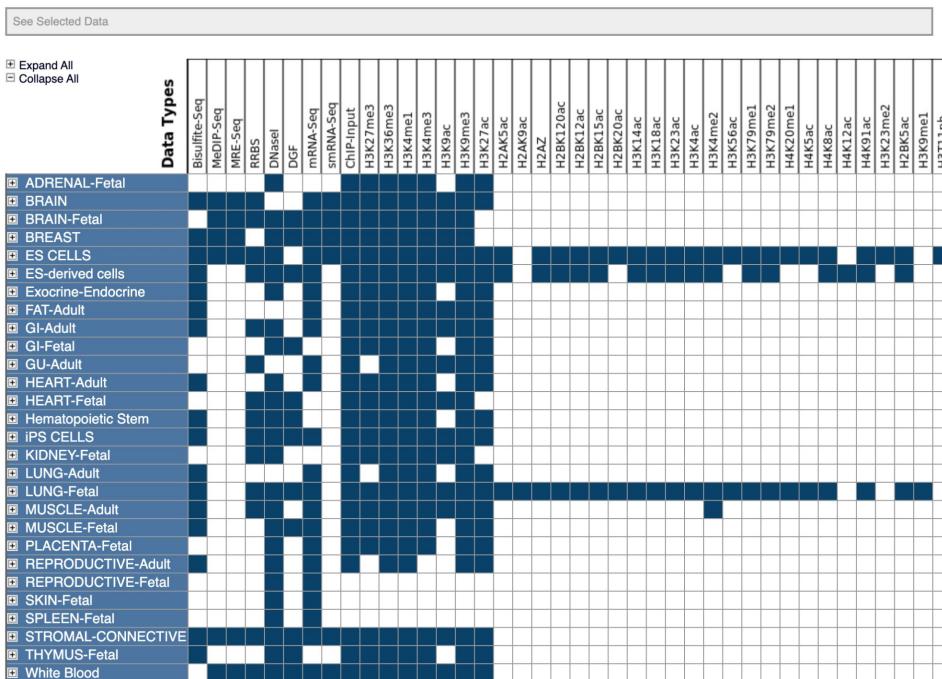
Samples	Populations	Publications	Website
<a href="#">1000 Genomes 30x on GRCh38</a>	3202	26	<a href="#">Byrska-Bishop et al., 2021</a>
<a href="#">Human Genome Structural Variation Consortium, Phase 2</a>	44	26	Ebert et al., 2021 <a href="#">Mark J P Chaisson et al., 2019</a>
<a href="#">1000 Genomes on GRCh38</a>	2709	26	Zheng-Bradley et al., 2017 <a href="#">Lowy-Gallego et al., 2019</a>
<a href="#">1000 Genomes phase 3 release</a>	3115	26	The 1000 Genomes Project Consortium, 2015 <a href="#">Sudmant et al., 2015</a>
<a href="#">1000 Genomes phase 1 release</a>	1182	14	<a href="#">The 1000 Genomes Project Consortium, 2012</a>
<a href="#">The Human Genome Structural Variation Consortium</a>	9	3	<a href="#">Chaisson et al., 2019</a>
<a href="#">Human Genome Diversity Project</a>	828	54	<a href="#">Bergström et al., 2020</a>
<a href="#">Simons Genome Diversity Project</a>	276	129	<a href="#">Mallick et al., 2016</a>
<a href="#">Gambian Genome Variation Project (GRCh38)</a>	518	4	<a href="#">Malaria Genomic Epidemiology Network et al., 2019</a>
<a href="#">Gambian Genome Variation Project (GRCh37)</a>	400	4	<a href="#">Malaria Genomic Epidemiology Network et al., 2019</a>
<a href="#">Geuvadis</a>	465	5	<a href="#">Lappalainen et al., 2013</a>
<a href="#">Illumina Platinum pedigree</a>	6	1	<a href="#">Eberle et al., 2017</a>
<a href="#">90 Han Chinese high coverage genomes</a>	90	2	<a href="#">Lan et al., 2017</a>
<a href="#">HGDP Transcriptome</a>	45	7	<a href="#">Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, et al., 2014</a>
<a href="#">Human Genome Structural Variation Consortium, Phase 3</a>	30	16	



# RoadMap

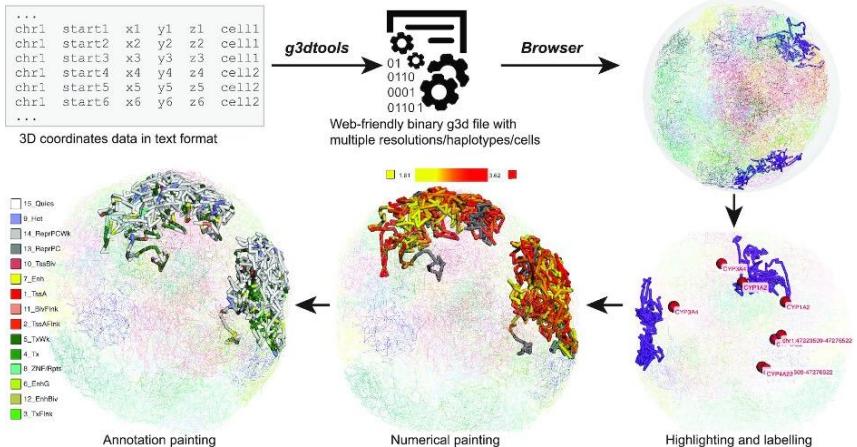
Roadmap Epigenomics Project, part of the NIH  
Roadmap Epigenomics Mapping Consortium  
Stem cells and primary ex vivo tissues selected to  
represent the normal counterparts of tissues and  
organ systems frequently involved in human disease  
Links to data download for Release 9 of the Human  
Epigenome Atlas 2,804 genome-wide datasets, which  
includes:

- 1,821 histone modification datasets
  - 360 DNase datasets
  - 277 DNA methylation datasets
  - 166 RNA-Seq datasets
  - subset of 1,936 datasets grouped into 111 reference epigenomes
  - 150.21 billion mapped sequencing reads corresponding to 3,174-fold coverage of the human genome



# WashU Epigenome Browser

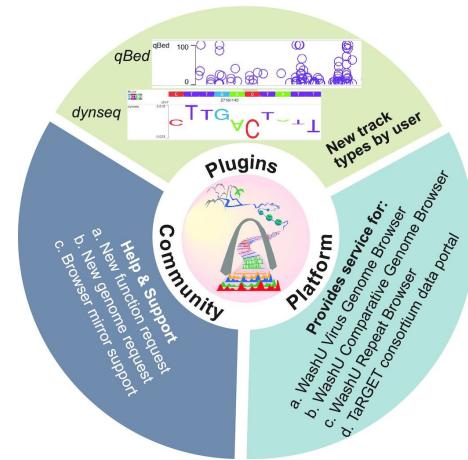
- WashU Epigenome Browser is a platform that hosts thousands of epigenome and transcriptome datasets for multiple cell types, tissues, individuals and species, to support multiple types of long-range genome interaction data.
- WashU Epigenome Browser enables investigators to explore epigenomic data in the context of higher-order chromosomal domains and to generate multiple types of intuitive, publication-quality figures of interactions.



**Table 1.** data statistics. Currently hosted track number for each species

Species	Genome assembly	Number of tracks	Number of 3D models	Number of images
Human	<i>hg38</i>	76 943	34	540 278
	<i>hg19</i>	134 484		544 814
Mouse	<i>mm10</i>	27 681	11 011	124
Chicken	<i>galGal5</i>	103		
Zebrafish	<i>danRer10</i>	66		
Fruit fly	<i>dm6</i>	6		
Plasmodium falciparum	<i>pfa3d7</i>	14	3	
Yeast	<i>sacCer3</i>	1	1	
Virus	SARS-CoV-2	2 642 765		2

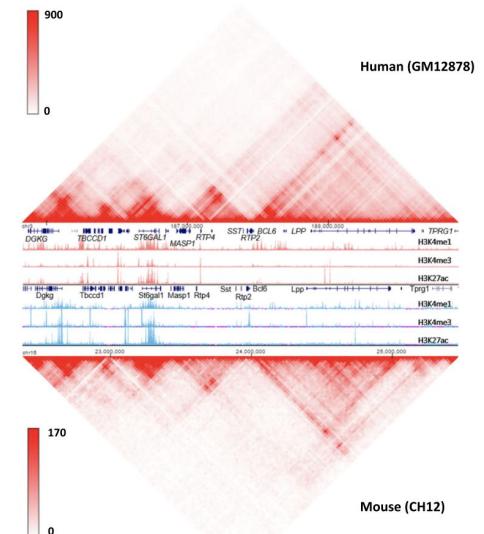
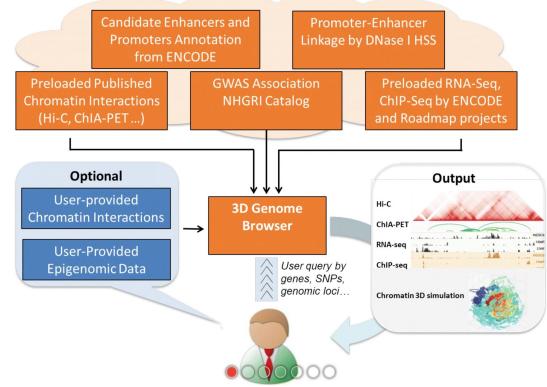
<https://epigenomegateway.wustl.edu/browser/>



# 3D Genome Browser

- 3D Genome Browser is a fast web-based browser that allows users to smoothly explore both published and their own chromatin interaction data.
- 3D Genome Browser features six distinct modes that allow users to explore interactome data tailored toward their own needs, from exploring organization of higher-order chromatin structures at domain level to investigating high-resolution enhancer-promoter interactions.

Data type	Samples and conditions	Total datasets
Hi-C	70	288
Virtual 4C, derived from Hi-C	Same as above	Same as above
ChIA-pet	14	14
Capture Hi-C	19	19
HiChIP	2	2
PLAC-Seq	3	3
GAM	1	1
DNase Hi-C	2	2
SPRITE	2	2
Total number	113	331



# UCSC Genome Browser

- The University of California Santa Cruz (UCSC) Genome Browser Database is an up to date source for genome sequence data integrated with a large collection of related annotations.
- UCSC is optimized to support fast interactive performance with the web-based UCSC Genome Browser, a tool built on top of the database for rapid visualization and querying of the data at many levels.

UNIVERSITY OF CALIFORNIA,  
SANTA CRUZ Genomics Institute

UCSC Genome Browser

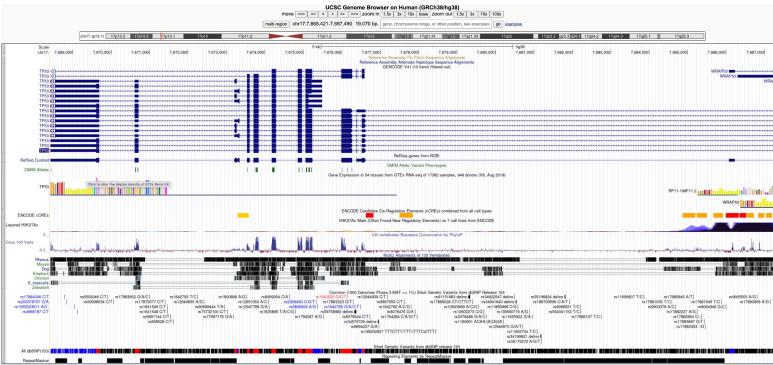
Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Human GRCh38/hg38  
Human GRCh37/hg19  
Human T2T-CHM13  
Mouse GRCm39/mm39  
Mouse GRCm38/mm10  
Genome Archive GenArk  
SARS-CoV-2 (COVID-19)  
Other

**Our tools**

- Genome Browser interactively visualize genomic data
- COVID-19 Research use the SARS-CoV-2 genome browser and explore coronavirus datasets
- BLAT rapidly align sequences to the genome
- Table Browser download data from the Genome Browser database
- Variant Annotation Integrator get functional effect predictions for variant calls
- Data Integrator combine data sources from the Genome Browser database
- Genome Browser in a Box (GBIB) run the Genome Browser on your laptop or server
- In-Silico PCR rapidly align PCR primer pairs to the genome
- LiftOver convert genome coordinates between assemblies
- Track Hub import and view external data tracks
- REST API returns data in JSON format

More tools...



## Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data a

### Select dataset

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)  
group: Genes and Gene Predictions track: GENCODE V41  
table: knownGene describe table schema

### Define region of interest

region:  genome  position chr17:7,668,421-7,687,490 lookup define regions  
identifiers (names/accessions):

### Optional: Subset, combine, compare with another track

filter:

intersection:

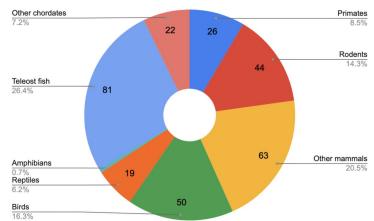
### Retrieve and display data

output format: all fields from selected table  Send output to  Galaxy  GREAT  
output filename:  (add .csv extension if opening in Excel, leave blank to keep output in browser)  
output field separator:  tsv (tab-separated)  csv (for excel)  
file type returned:  plain text  gzip compressed

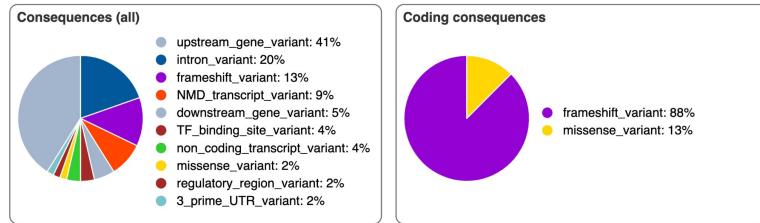
GIVE: Build your own genome browser

# Ensembl

- Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data.
- Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.



- The Ensembl Variant Effect Predictor (VEP) is a powerful toolset for the analysis, annotation, and prioritization of genomic variants in coding and non-coding regions.
- VEP provides access to an extensive collection of genomic annotation, with a variety of interfaces to suit different requirements, and simple options for configuring and extending analysis.



## Ensembl Variant Effect Predictor (VEP)

### VEP interfaces

#### Web interface

- Point-and-click interface
- Suits smaller volumes of data

Documentation

Launch VEP

#### Command line tool

- More options and flexibility
- For large volumes of data

Documentation

#### REST API

- Language-independent API
- Simple URL-based queries

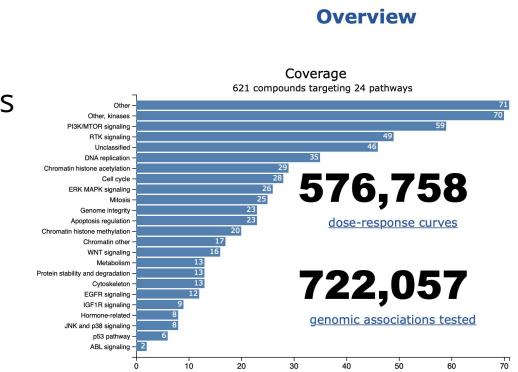
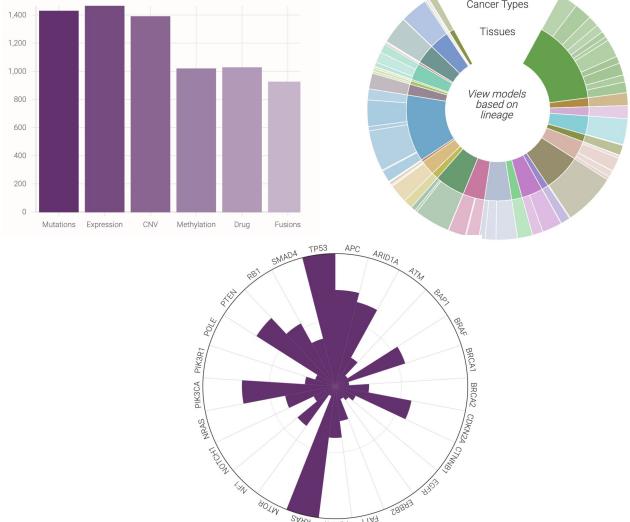
Documentation

VEP REST API

# Cell Line Specific Data Resources

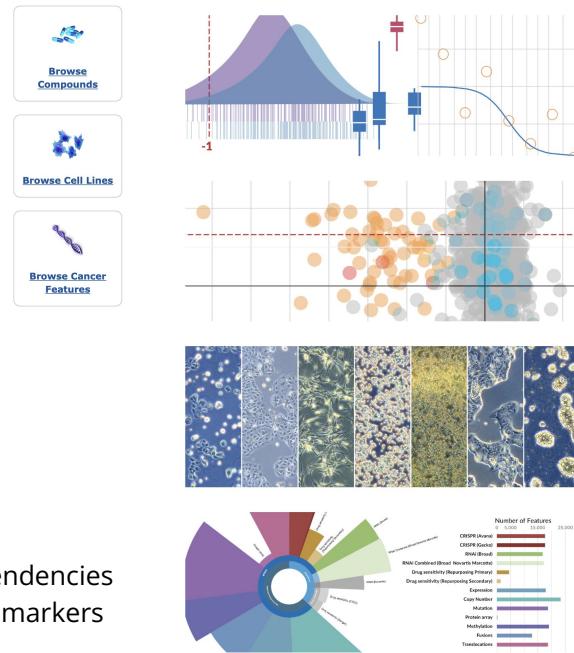
## Cell Model Passports

A Hub for Preclinical Cancer Models -  
Annotation, Genomics & Functional Datasets



## Genomics of Drug Sensitivity in Cancer

Association between drug response data and genomic markers of sensitivity



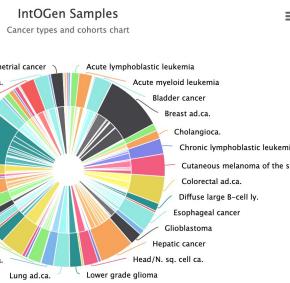
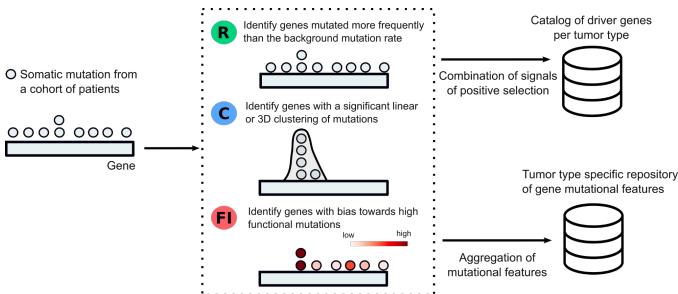
# **Specialized Databases for Genomic Analyses**

# Driver Gene Analysis

## Integrative Onco Genomics (IntOGen)

Collects and analyzes somatic mutations in tumor genomes to identify cancer driver genes

66 cancer types
221 cohorts
28,076 samples
>200 million mutations
568 drivers



## OncokB

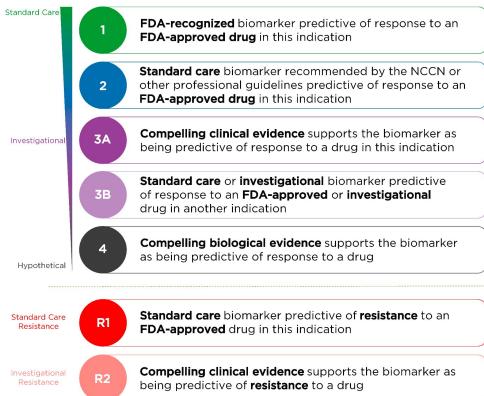
### MSK's Precision Oncology Knowledge Base

Annotation of biological consequences and clinical implications of genetic variants in cancer

Database contains: 688 genes, 5729 alterations, 133 cancer types, 111 drugs

Potential driver genes in different categories. Each category has a series of levels:

- Therapeutic levels
- Diagnostic levels
- Prognostic levels
- FDA levels



## Cosmic Cancer Gene Census

The Cancer Gene Census (CGC) is an ongoing effort to catalogue those genes which contain mutations that have been causally implicated in cancer and explain how dysfunction of these genes drives cancer.

## Cancer Hotspots

A resource for statistically significant mutations in cancer identified in large scale cancer genomics data

Single residue and in-frame indel mutation hotspots in 24,592 tumor samples

Can download hotspot results and mutation data (MAF)

Show/Hide † Mouse over Variants and Samples values for more information

Search:

Gene	Residue	Type	Variants †	Q-value	Samples †
NRAS	Q61	single residue	R K L	0	422
PIK3CA	E545	single residue	K	0	633
IDH1	R132	single residue	H C	0	766
PIK3CA	H1047	single residue	R L	0	647
BRAF	V600	single residue	E	0	897
EGFR	L858	single residue	R	0	144
TP53	R175	single residue	H	0	416
KRAS	Q61	single residue	H R L K	0	190
KRAS	G13	single residue	D C	0	264
TP53	R248	single residue	Q W	0	560
KRAS	G12	single residue	D V C R	0	2175
TP53	R273	single residue	C H L	0	609
PIK3CA	E542	single residue	K	0	372
AKT1	E17	single residue	K	1.15e-288	349
GNAS	R201	single residue	H C	7.47e-257	139
FGFR3	S249	single residue	C	3.49e-239	114
PIK3CA	N345	single residue	K I	3.87e-219	98
PTEN	R130	single residue	Q G *	1.92e-210	168
HRAS	Q61	single residue	R K L	8.15e-210	102
TP53	Y220	single residue	C	4.54e-208	152

Showing 1 to 20 of 1,165 mutations

Previous 1 2 3 4 5 ... 59

# FIREBROWSE (Broad GDAC) for Somatic Analysis

## CopyNumber Analyses

Aggregate AnalysisFeatures  
CopyNumber Clustering CNMF  
CopyNumber Clustering CNMF thresholded

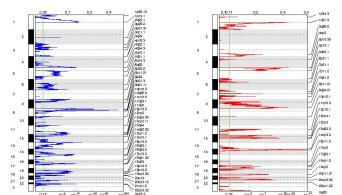
**CopyNumber Gistic2**  
CopyNumberLowPass Gistic2

Correlate Clinical vs CopyNumber Arm  
Correlate Clinical vs CopyNumber Focal

Correlate CopyNumber vs mRNA  
Correlate CopyNumber vs mRNASeq

Correlate molecularSubtype vs CopyNumber Arm  
Correlate molecularSubtype vs CopyNumber Focal

Pathway Paradigm mRNA And Copy Number  
Pathway Paradigm RNASeq And Copy Number



## Mutation Analyses

Aggregate AnalysisFeatures  
Correlate Clinical vs Mutation  
Correlate Clinical vs Mutation APOBEC Categorical  
Correlate Clinical vs Mutation APOBEC Continuous  
Correlate Clinical vs MutationRate  
Correlate molecularSubtype vs Mutation  
Correlate mRNASeq vs Mutation APOBEC

**Mutation APOBEC**

Mutation Assessor

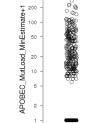
Mutation CHASM

MutSig2.0

MutSig2CV

MutSigCV

Pathway Overlaps MSigDB MutSig2CV



## mRNAseq Analyses

Aggregate AnalysisFeatures  
Correlate Clinical vs mRNASeq  
Correlate CopyNumber vs mRNASeq  
Correlate mRNASeq vs Mutation APOBEC

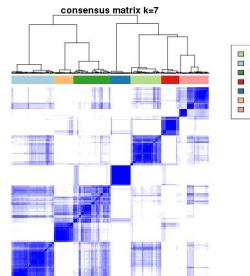
miRseq FindDirectTargets

mRNASeq Clustering CNMF

**mRNASeq Clustering Consensus Plus**

Pathway Paradigm RNASeq

Pathway Paradigm RNASeq And Copy Number



## Correlations Analyses

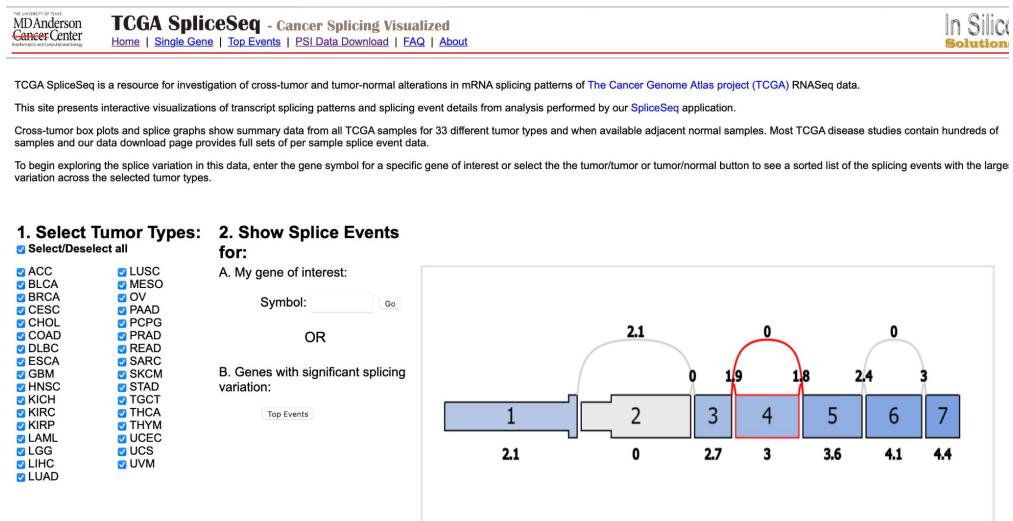
Correlate Clinical vs CopyNumber Arm  
Correlate Clinical vs CopyNumber Focal  
Correlate Clinical vs Methylation  
Correlate Clinical vs miRseq  
Correlate Clinical vs Molecular Subtypes  
Correlate Clinical vs mRNA  
Correlate Clinical vs mRNASeq  
Correlate Clinical vs Mutation  
Correlate Clinical vs Mutation APOBEC Categorical  
Correlate Clinical vs Mutation APOBEC Continuous  
Correlate Clinical vs MutationRate  
Correlate Clinical vs RPPA  
Correlate CopyNumber vs mRNA  
Correlate CopyNumber vs mRNASeq  
Correlate Methylation vs mRNA  
Correlate molecularSubtype vs CopyNumber Arm  
Correlate molecularSubtype vs CopyNumber Focal  
Correlate molecularSubtype vs Mutation  
Correlate mRNASeq vs Mutation APOBEC

And more...

# Alternative Splicing

## Oncosplicing

Oncosplicing is a database to systematically study clinically relevant alternative splicing in 33 TCGA cancers and 31 GTEx tissues.



## TCGA SpliceSeq

TCGA SpliceSeq is a resource for investigation of cross-tumor and tumor-normal alterations in mRNA splicing patterns of **The Cancer Genome Atlas project (TCGA)** RNASeq data.

# PCAWG Data Portal for WGS Based Somatic Analysis

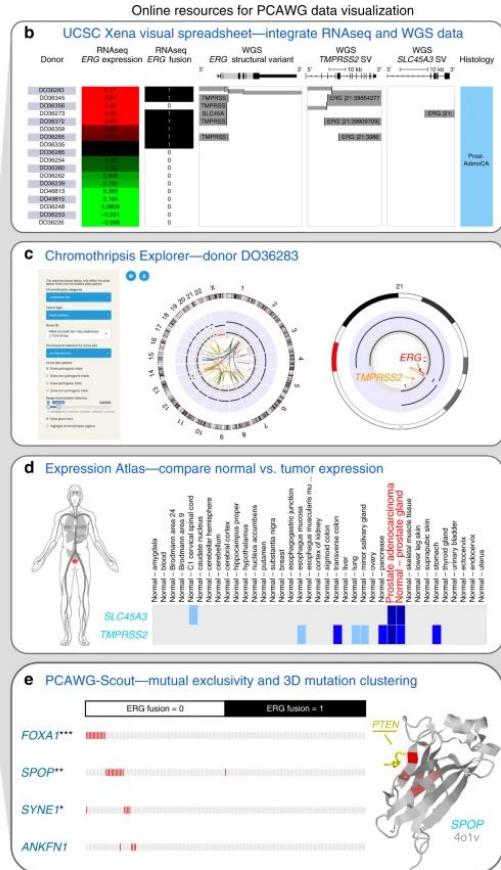
a

Search and download > 70,000 primary BAMs and VCFs

ICGC data repository

PCAWG AWGs primary results (mutation, CNV, expression, etc.)

Search and download primary results files



## PCAWG Data Portal

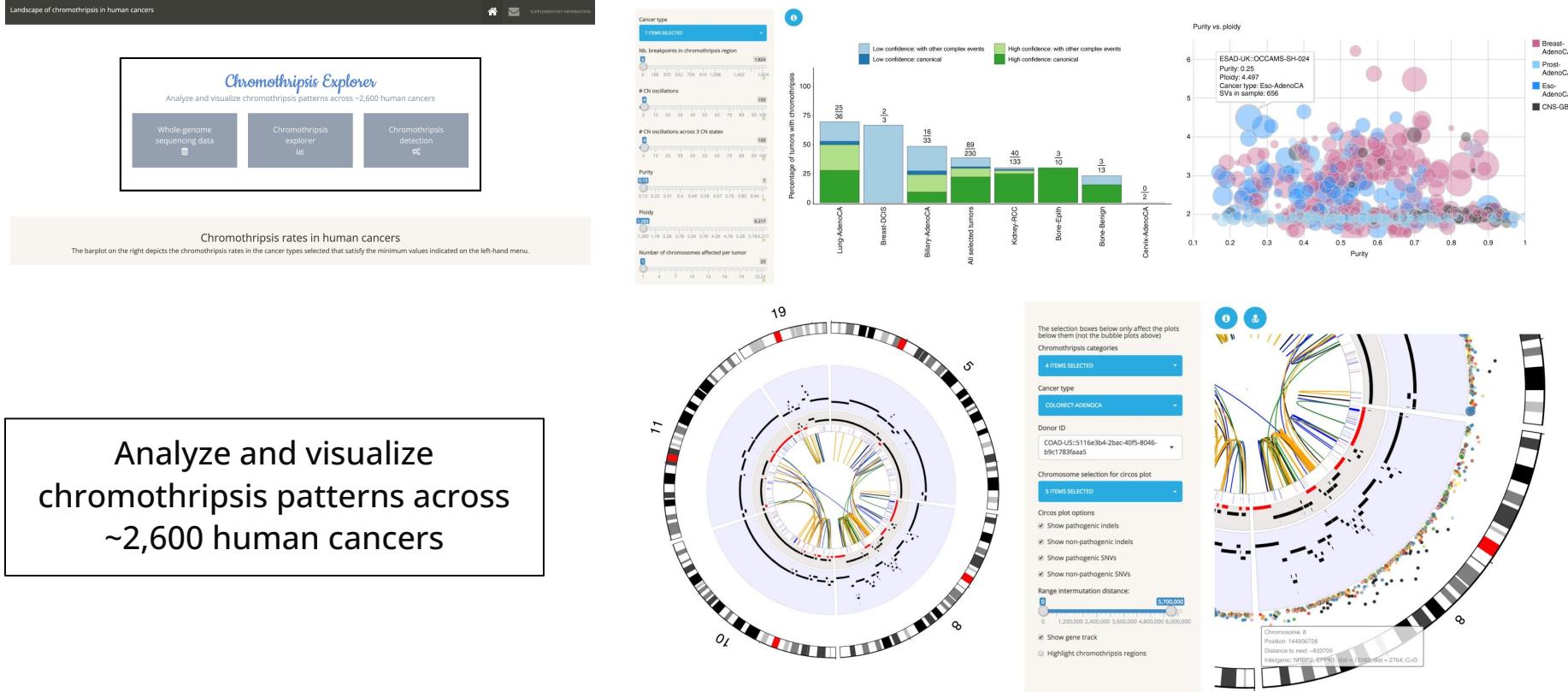
DCC / PCAWG / Filter by file name...

Name

- README.md
- APOBEC\_mutagenesis
- benchmarking\_data
- cell\_lines
- clinical\_and\_histology
- consensus\_cnv
- consensus\_snv\_indel
- consensus\_sv
- data\_releases
- donors\_and\_biospecimens
- driver\_mutations
- drivers
- evolution\_and\_heterogeneity
- germline\_variations
- Hartwig
- hla\_and\_neoantigen
- msi
- mutational\_signatures
- networks
- pathogen\_analysis

A user guide for the online exploration  
and visualization of PCAWG data

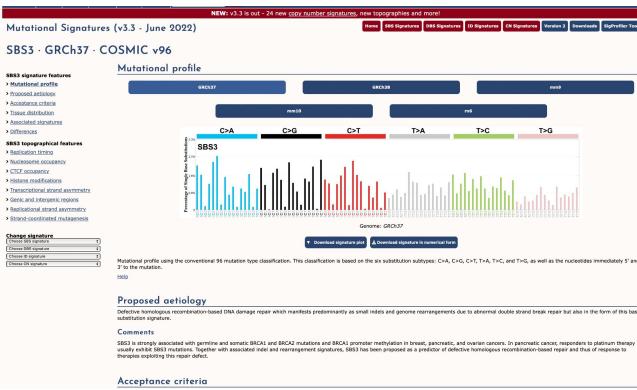
# Chromothripsis Explorer



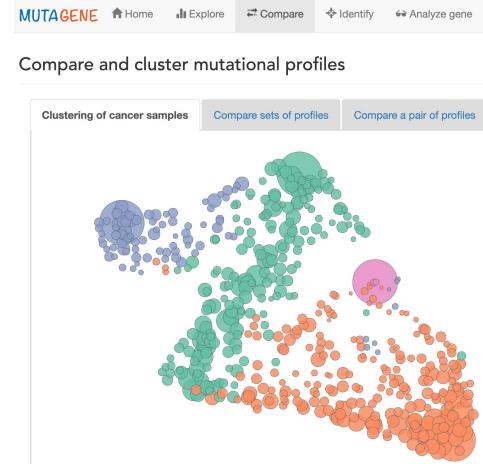
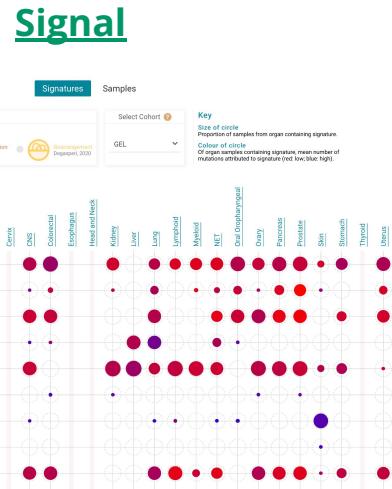
# Mutational Signature Analysis

## MUTAGENE

## COSMIC Mutational Signatures



- Support different mutational profiles (SBS/ID/DBS/CNV)
  - Detail annotations (proposed aetiology and acceptance criteria, tissue distribution, signature associations, version difference)
  - Topographies associations (Replication timing, Nucleosome occupancy, CTCF occupancy, Histone modifications, Transcriptional strand asymmetry, Genic and intergenic regions, Replication strand asymmetry, strand-coordinated mutagenesis)



- Explore context-dependent mutational profiles and signatures
  - Compare and cluster mutational profiles
  - Identify mutational processes based on NMF algorithm
  - Compares observed mutational frequencies to expected background mutability to identify potential drivers

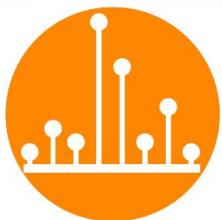
# mSigPortal: Integrative mutational signature portal for cancer genomic studies

## MSIGPORTAL

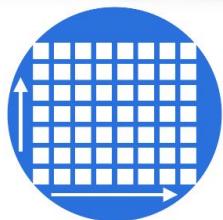
*Integrative mutational signature portal for cancer genomic studies*



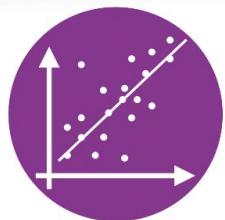
Signature Catalog



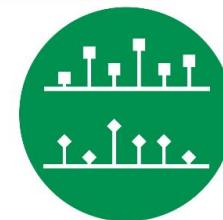
Signature Visualization



Signature Exploration



Signature Association



Signature Extraction

Coming soon!

All existing human and mouse signatures based on different genome builds and algorithm versions

Allows identification of signature features at sample level and discovery of new signatures

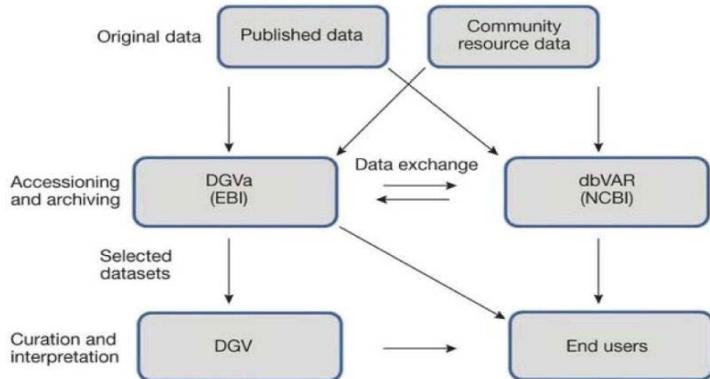
Explore etiological factors associated with signature at sample level

Analyze signature association with other genomic features and clinical data

Extract and compare mutational signatures using state-of-the-art algorithms

# Database of Genomic Variants (DGV): SV + CNV

The objective of the Database of Genomic Variants is to provide a comprehensive summary of structural variation in the human genome. We define structural variation as genomic alterations that involve segments of DNA that are **larger than 50bp**. The content of the database is only representing structural variation identified in healthy control samples. The Database of Genomic Variants provides a useful catalog of control data for studies aiming to correlate genomic variation with phenotypic data.



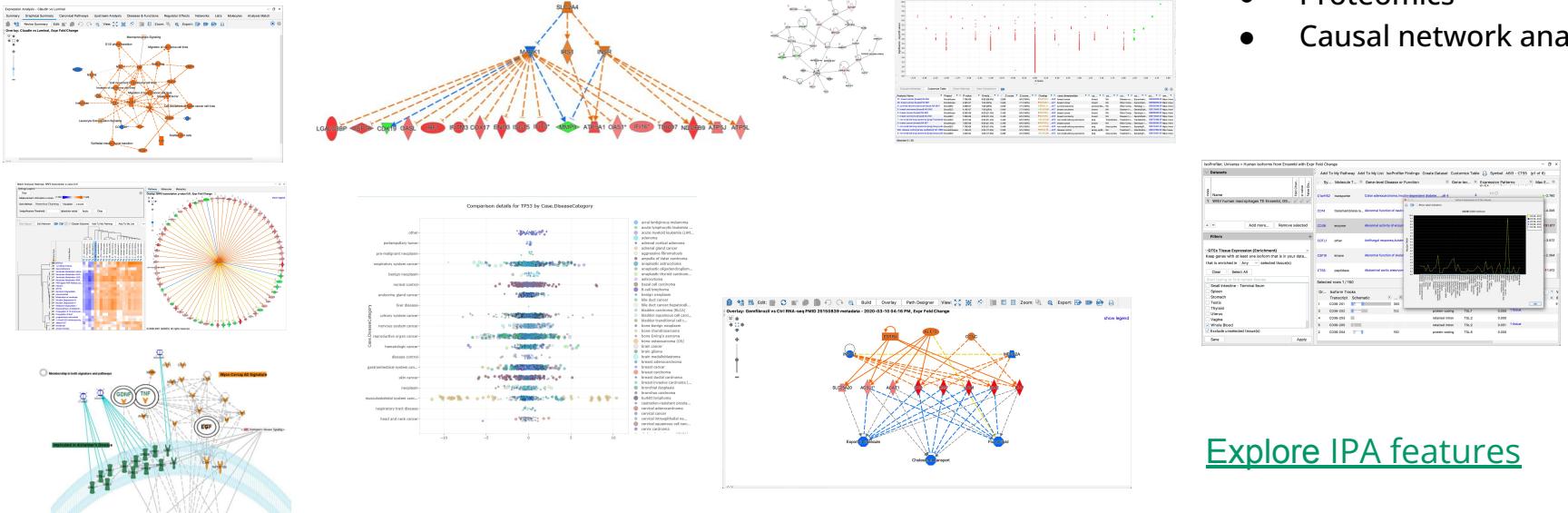
## Other sites displaying data from DGV:

- DECIPHER
- Ensembl
- UCSC
- HapMap
- GeneCards

**D**atabase of *G*enomic *V*ariants  
A curated catalogue of human genomic structural variation

# Ingenuity Pathway Analysis (IPA)

- Pathway enrichment
- Identifying key regulators and activity to explain expression patterns
- Predicting downstream effects on biological and disease processes
- Providing targeted data on genes, proteins, chemicals, and drugs
- Building interactive models of experimental systems



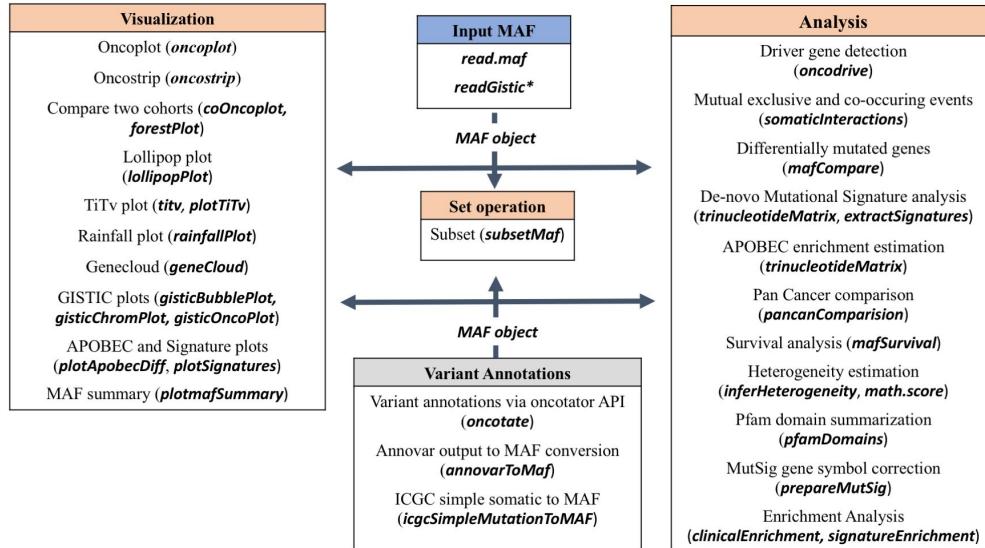
**Applications:**  
IPA helps to uncover the discovery behind the data in:

- Transcriptomics
- Biomarker discovery
- miRNA research
- Toxicogenomics
- Metabolomics
- Drug repositioning
- Proteomics
- Causal network analysis

[Explore IPA features](#)

# Analytical Programming Packages for Cancer Genomic Datasets

# MAFtools: Summarize, Analyze and Visualize MAF Files



[\*\*TCGAmutations\*\*](#) - An R data package for TCGA somatic mutations

**Data source:**  
[Broad firehose](#)  
[TCGA MC3](#)

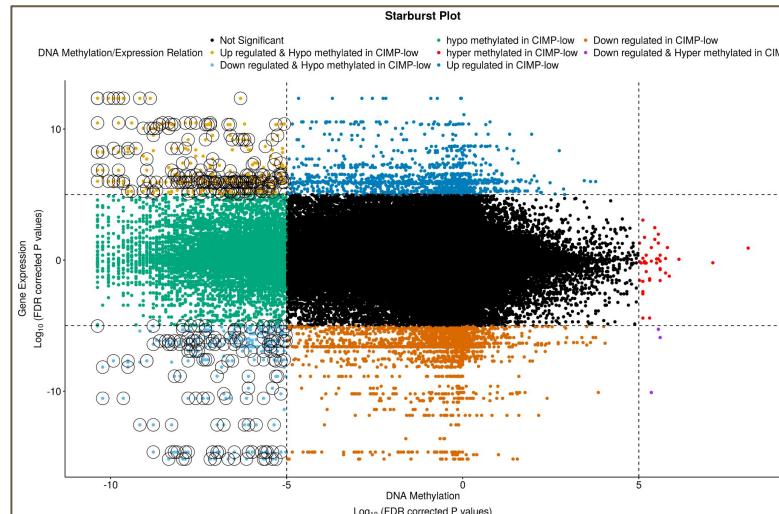
(Multi-Center Mutation Calling in Multiple Cancers)

Most of MAFs from different cancer genomic studies can also be download from NCI GDC

# TCGAbiolinks

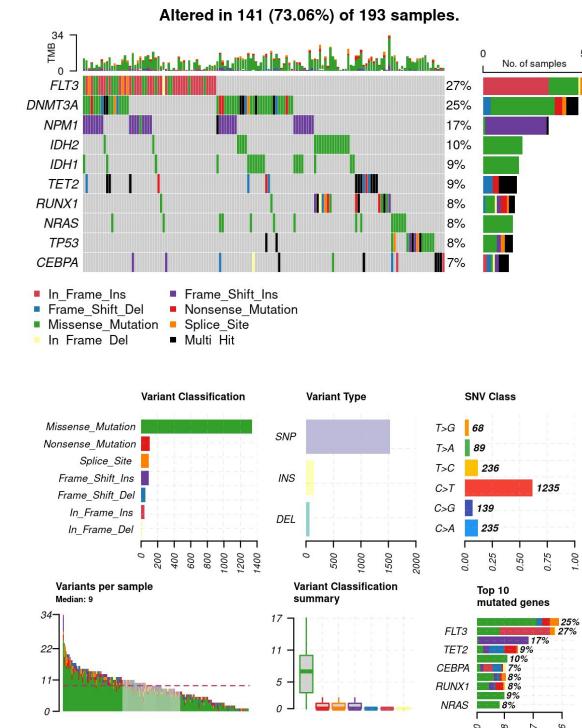
- |                      |                          |                                                 |
|----------------------|--------------------------|-------------------------------------------------|
| <a href="#">HTML</a> | <a href="#">R Script</a> | 1. Introduction                                 |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 10. Classifiers                                 |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 10. TCGAbiolinks_Extension                      |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 11. Stemness score                              |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 2. Searching GDC database                       |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 3. Downloading and preparing files for analysis |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 4. Clinical data                                |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 5. Mutation data                                |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 6. Compilation of TCGA molecular subtypes       |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 7. Analyzing and visualizing TCGA data          |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 8. Case Studies                                 |
| <a href="#">HTML</a> | <a href="#">R Script</a> | 9. Graphical User Interface (GUI)               |
| <a href="#">PDF</a>  |                          | Reference Manual                                |
| <a href="#">Text</a> |                          | NEWS                                            |

- Case study 1: Pan Cancer downstream analysis BRCA
- Case study 2: Pan Cancer downstream analysis LGG
- Case study 3: Integration of methylation and expression for ACC
- Case study 4: ELMER pipeline - KIRC



# TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages

- Studies: TCGA, ENCODE, Roadmap
- Data access
  - TCGA data
    - NCI Genomic Data Commons (GDC) (level 1 to 3 data)
      - Via TCGAbiolinks
    - GDC Legacy Archive
      - Via TCGAbiolinks
    - Broad Institute's GDAC Fire (level 3 and 4 data)
      - Via RTCGAToolbox
  - ROADMAP via AnnotationHub
  - ENCODE via ENCODExplorer
- Data analysis and visualization
  - maftools :
    - Summarize, Analyze and Visualize MAF Files
  - ELMER
    - Inferring Regulatory Element Landscapes and Transcription Factor Networks Using Cancer Methylomes



# cBioPortalData

- cBio Cancer Genomics Portal
  - <https://www.cbioportal.org/>
  - Platform for exploratory and interactive visualization, analysis and download of large-scale cancer genomic data sets.
  - Data sets
    - Public data (TCGA, ICGC, published sequencing studies)
    - Private instances
    - Visualize your own data
  - Open source
- cBioPortalData R package: Obtain data from the cBioPortal API using R
  - Identifying available studies
  - Download studies via cBioDataPack or cBioPortalData

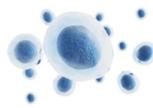
# Cloud Resources with Available Cancer Genomic Datasets



- Developed at the Broad Institute
- Uses Google Cloud Platform as compute and storage infrastructure
- Workflows need to be written in WDL (Workflow Description Language)
- Lots of common workflows available in Dockstore (<https://dockstore.org/>) and firecloud ([firecloud.org](https://firecloud.org))
- Easily shareable workspaces

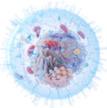
## Terra supports researchers in many biomedical disciplines

Cancer Genomics



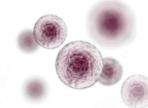
The Van Allen Lab is using Terra to advance clinical oncology through cancer genomics.

Single-Cell Transcriptomics



The Shalek Lab is using Terra to improve the scalability, accessibility, and reproducibility of single-cell analysis.

Medical and Population Genetics



The Natarajan Lab is using Terra to study genetic factors of heart diseases.

[Read More](#)

Infectious Diseases



The Broad's Viral Genomics Group is using Terra to advance genomic epidemiology and surveillance of viral pathogens.

## Datasets included in Terra:

- Human Reference Datasets
- [Human Cell Atlas](#)
- FireCloud resources  
(GTEX/TARGET/TCGA)

## How to access?



# Seven Bridges Genomics

- Backend is AWS
- One of the 3 available cloud resources for analysis if the data is in NCI Cancer Research Data Commons (CRDC)

## Datasets included:

- Cancer Genomics Cloud
- Gabriella Miller Kids First Data Center
- Trans-Omics for Precision Medicine (TOPMed), Genotype-Tissue Expression (GTEx)
- Model Organism Databases (MODs) datasets
- ICGC and PCAWG
- Blood Profiling Atlas in Cancer (BloodPAC)



Figure 3. Overview of the Seven Bridges Platform. A Cloud-based data and compute infrastructure underlies the discovery layer, which is built around features to streamline data management, search, and analysis. An application programming interface and collaboration features ensure flexibility for users. Data security and regulatory compliance controls operate at all levels.

[https://www.sevenbridges.com/wp-content/uploads/2016/11/WP\\_Scalable\\_Web.pdf](https://www.sevenbridges.com/wp-content/uploads/2016/11/WP_Scalable_Web.pdf)

# DNAnexus

- Backend is AWS and Azure
- Enables easy data sharing
- Provides easy to use tools, APIs and visualization
- Data accessible with DNAnexus:
  - St Jude Cloud Genomics Platform
  - UKBiobank
  - ICGC Pan-Cancer dataset
    - Hosted on AWS S3
    - DNAnexus provides ICGC Data Fetcher
  - TCGA
    - Hosted on AWS S3



[Log In to St. Jude Cloud](#)

St. Jude Cloud data is vended through the DNAnexus interface. If you have a DNAnexus employee, please log in with your St. Jude credentials.



# “Awesome” bioinformatics resources

# “Awesome” bioinformatics resources related to cancer genomic study

## Awesome genomics

Cancer Data Science's go to place for excellent genomics tools and packages.

## Awesome multi-omics

A community-maintained list of software packages for multi-omics data analysis.

## Awesome cancer variant databases

A community-maintained repository of cancer clinical knowledge bases and databases focused on cancer and normal variants

## Awesome cancer evolution

Papers for studying cancer evolution

## Awesome genome visualization

A list of interesting genome visualizers, genome browsers, or genome-browser-like implementations. [New website](#).

## Awesome Clonality

A curated list of awesome resources on clonality and tumor heterogeneity.

## Awesome expression browser

A curated list of software and resources for exploring and visualizing (browsing) expression data, but not only limited to that.

## Awesome microbes

List of resources, including software packages (and the people developing these methods) for microbiome (16S), metagenomics (WGS, Shot-gun sequencing), and pathogen identification/detection/characterization.

## Awesome bioinformatics benchmarks

A curated list of bioinformatics benchmarking papers and resources.

## Awesome single cell

List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc.

## Awesome bioinformatics

A curated list of awesome Bioinformatics software, resources, and libraries. Mostly command line based, and free or open-source.

## Awesome ChIP-Seq

A curated list of ChIP-Seq analysis

# Awesome mutational signature resource in mSigPortal

Original Research Papers Including Specific Mutational Signatures in mSigPortal

Cancer Type	Experimental Strategy	Year	Journal	Title
Human germline	WGS	2021	Science	Population sequencing data reveal a compendium of mutational processes in the human germ line
Glioma	WGS & WES	2021	Nature Genetics	Radiotherapy is associated with a deletion signature that contributes to poor outcomes in patients with cancer
iPSC	WGS	2021	Nature Cancer	A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage
Hematopoietic stem and progenitor cells (HSPCs)	WGS	2021	Cell Stem Cell	Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients
Melanoma	circle-damage-seq	2021	Science Advances	The major mechanism of melanoma mutations is based on deamination of cytosine in pyrimidine dimers as determined by circle damage sequencing
Colorectal Cancer	WES	2021	Cancer Discovery	Discovery and features of an alkylating signature in colorectal cancer
Skin Cancer	WGS	2021		Pre-mutagenic and mutagenic changes imprinted on the genomes of mammalian cells after irradiation with a nail polish dryer
PanCancer	WGS	2021		Mutational impact and signature of ionizing radiation
PanCancer	WGS & WES	2021	Brief Bioinform	Comprehensive analysis reveals distinct mutational signature and its mechanistic insights of alcohol consumption in human cancers
Liver Cancer	WGS & WES	2021	Hepatology	Mutational Signature Analysis Reveals Widespread Contribution of Pyrrolizidine Alkaloid Exposure to Human Liver Cancer

10 Showing 1 to 10 of 38 entries

Review Papers Focused on Mutational Signatures

Year	Journal	Title
2021	Nature Reviews Cancer	Mutational signatures: emerging concepts, caveats and clinical applications
2021	DNA Repair	Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures
2020	Nature Genetics	Are carcinogens direct mutagens?
2019	Briefings in Bioinformatics	Computational approaches for discovery of mutational signatures in cancer
2019	Nature Reviews Genetics	Switching APOBEC mutation signatures
2019	Cell	Local Determinants of the Mutational Landscape of the Human Genome
2018	Trends in Cancer	Mutation Signatures Depend on Epigenomic Contexts
2018	Journal of the National Cancer Institute	Biomarkers for Homologous Recombination Deficiency in Cancer
2018	Nature Communications	The therapeutic significance of mutational signatures from DNA repair deficiency in cancer
2018	Briefings in Bioinformatics	Mutational signatures and mutable motifs in cancer: Computational Methods, Tools, Databases or Websites for Mutational Signature Analyses

10 Showing 1 to 10 of 23 entries

Name	Method	Year	Journal	Title
RepairSig	Non-additive model	2021	Cell Systems	RepairSig: Deconvolution of DNA damage and repair contributions to the mutational landscape of cancer
TensorSignatures	Tensor factorisation framework	2021	Nature Communications	Learning mutational signatures and their multidimensional genomic properties with TensorSignatures
SparseSignatures	LASSO regularization	2021	PLOS Computational Biology	De novo mutational signature discovery in tumor genomes using SparseSignatures
Signal	NMF+KLD	2020	Nature Cancer	A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies
CANCERSIGN	NMF	2020	Scientific Reports	CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes
YAPSA	LCD	2020	Genes Chromosomes Cancer	Analysis of mutational signatures with yet another package for signature analysis
MutSignatures	NMF/fcnlms	2020	Scientific Reports	MutSignatures: an R package for extraction and analysis of cancer mutational signatures
iMutSig	Probabilistic Method	2020	F1000Research	iMutSig: a web application to identify the most similar mutational signature using shiny
Sigflow	Bayesian NMF	2020	Bioinformatics	Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis
pyCancerSig	NMF	2020	BMC Bioinformatics	pyCancerSig: subclassifying human cancer with comprehensive single nucleotide, structural and microsatellite mutational signature deconstruction from whole genome sequencing

10 Showing 1 to 10 of 32 entries

**THANKS FOR YOUR ATTENTION!**

**Questions?**

**Next: Practical session 2 (10:45am)**