Emerging Approaches for Tumor Analyses In Epidemiological Studies

Session 2: Public Databases

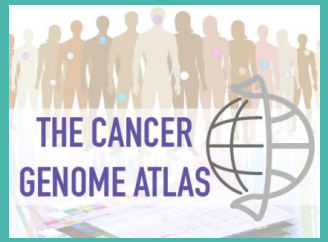
November 9, 2022

Session Overview

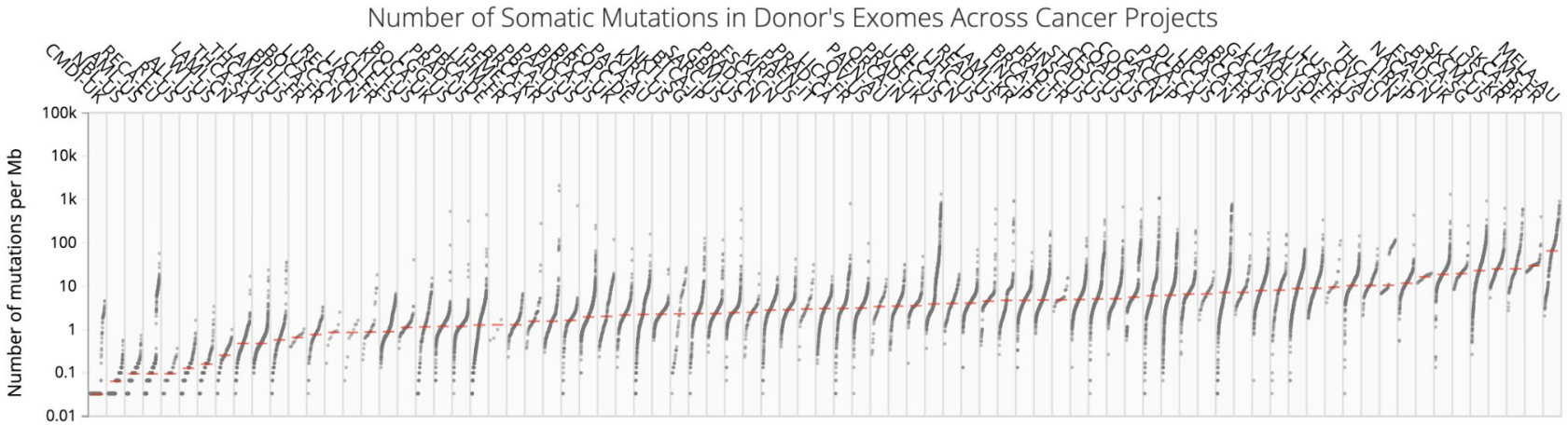
- Major Cancer Genomic Studies with Data Portal Available
- Common Genetics Data Resources
- Specialized Databases for Genomic Analyses
- Analytical Programming Packages for Cancer Genomic Datasets
- Cloud Resources with Available Cancer Genomic Datasets
- “Awesome” Bioinformatics Resources

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

DCC: ICGC Data Coordination Center
ARGO: Accelerating Research in Genomic Oncology

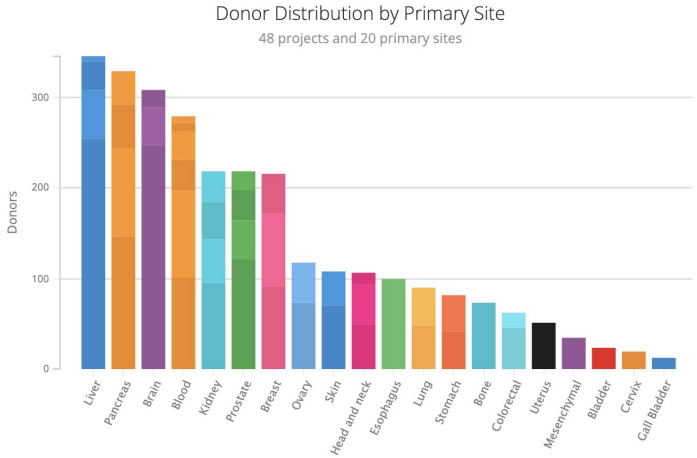
234,022 Files | 16,307 Donors | 1.73 PB

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

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

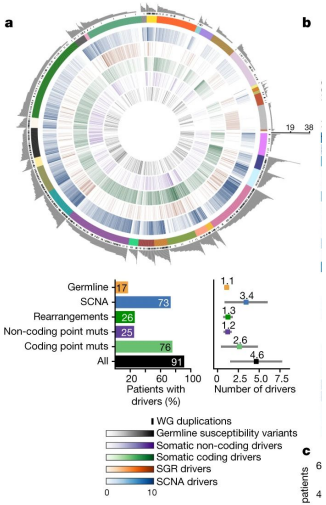


2,793 Donors | **71,709** Files | **795.14 TB**

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

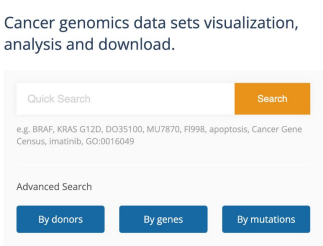
2658 cancer whole genomes with matched normal tissues across 38 tumor types

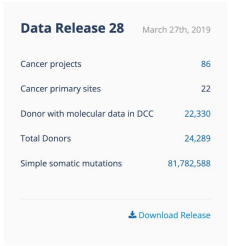


Major Publications

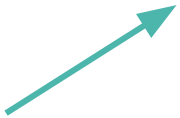
1. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. **Pan-cancer analysis of whole genomes**. Nature (2020).
2. Rheinbay, E. et al. **Analyses of non-coding somatic drivers in 2,693 cancer whole genomes**. Nature (2020).
3. PCAWG Transcriptome Core Group et al. **Genomic basis of RNA alterations 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).

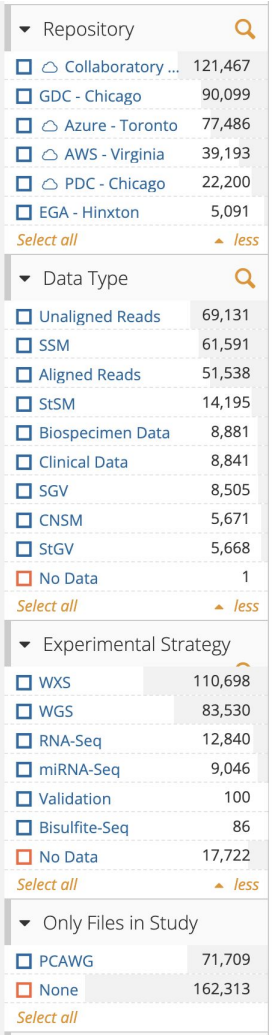
ICGC Data Portal



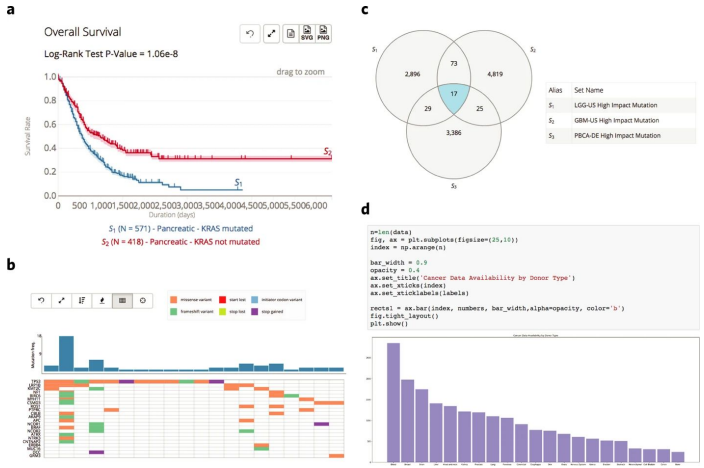


The ICGC Data Portal Facet Search interface





Using the ICGC Data Portal



WGS:100,000 Genomes Project | Genomics England

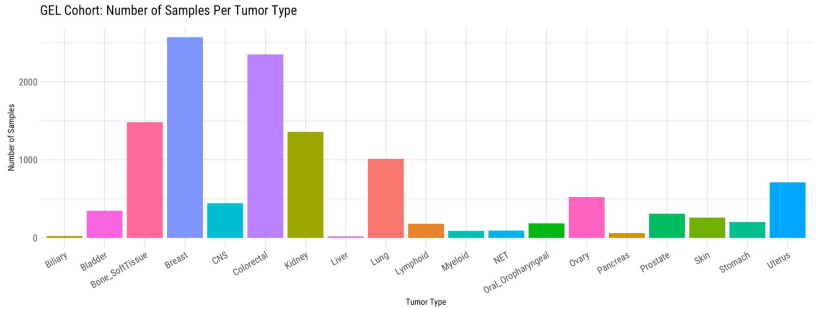
Genomics England's very first initiative – sequencing 100,000 genomes from around 85,000 NHS patients affected by rare disease or cancer – is leading to groundbreaking insights and continued findings into the role genomics can play in healthcare.

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.



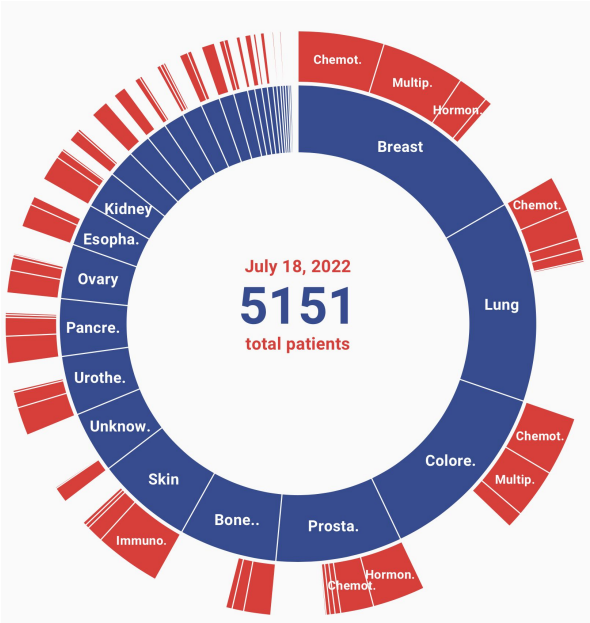
The GEL cohort:

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).

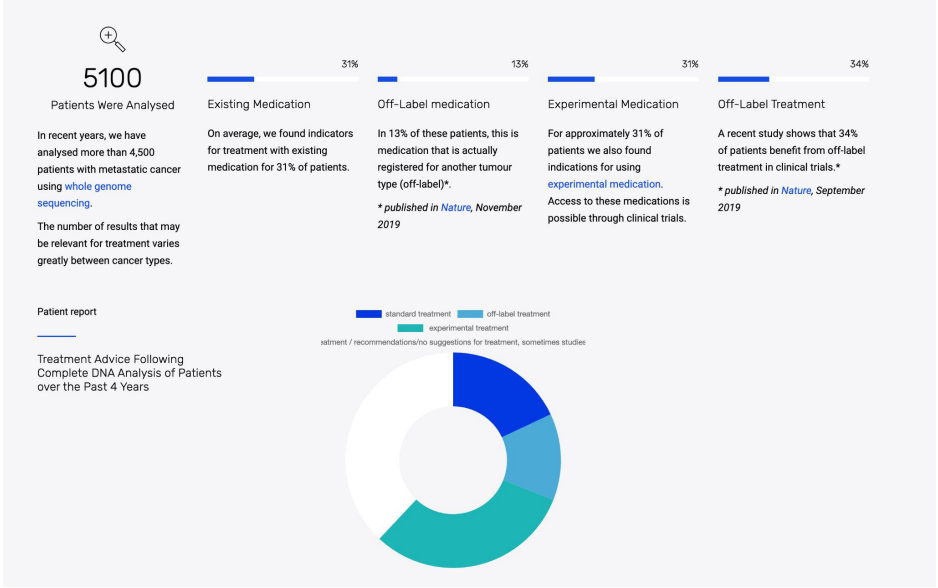
19 tumor types: skin, lung, stomach, colorectal, bladder, liver, uterus, ovary, biliary, kidney, pancreas, breast, prostate, bone and soft tissue, central nervous system (CNS), lymphoid, oropharyngeal, neuroendocrine tumors (NETs), and myeloid

WGS: Hartwig Medical Foundation



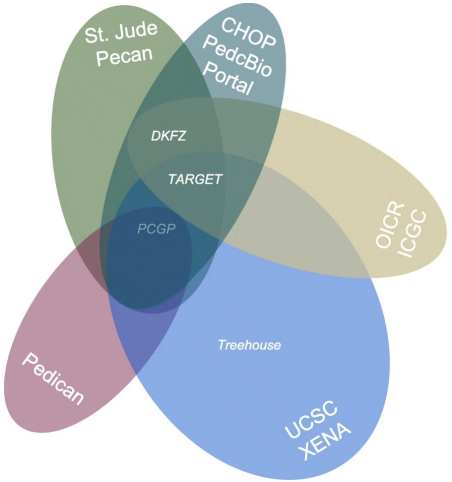
https://database.hartwigmedicalfoundation.nl

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.



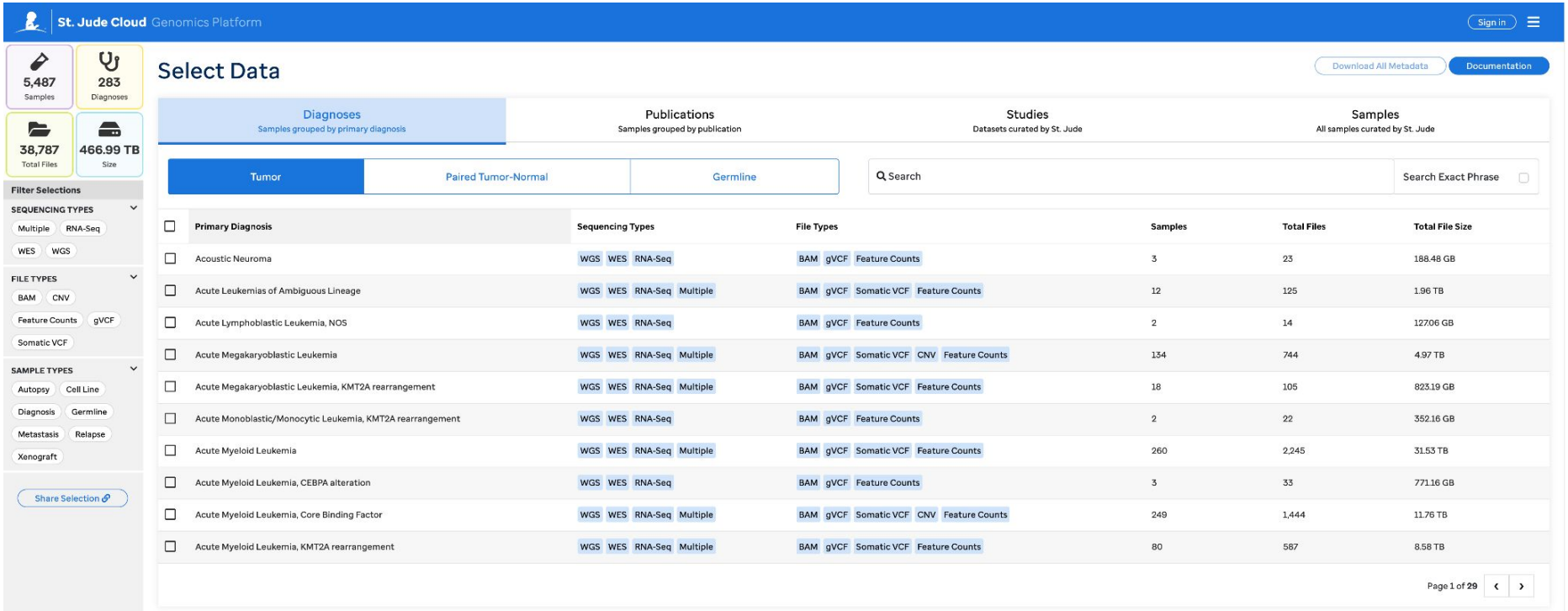
WGS: Pediatric Cancer Genome Project (PCGP) and Pediatric Cancer Databases (PeCan)

- 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.

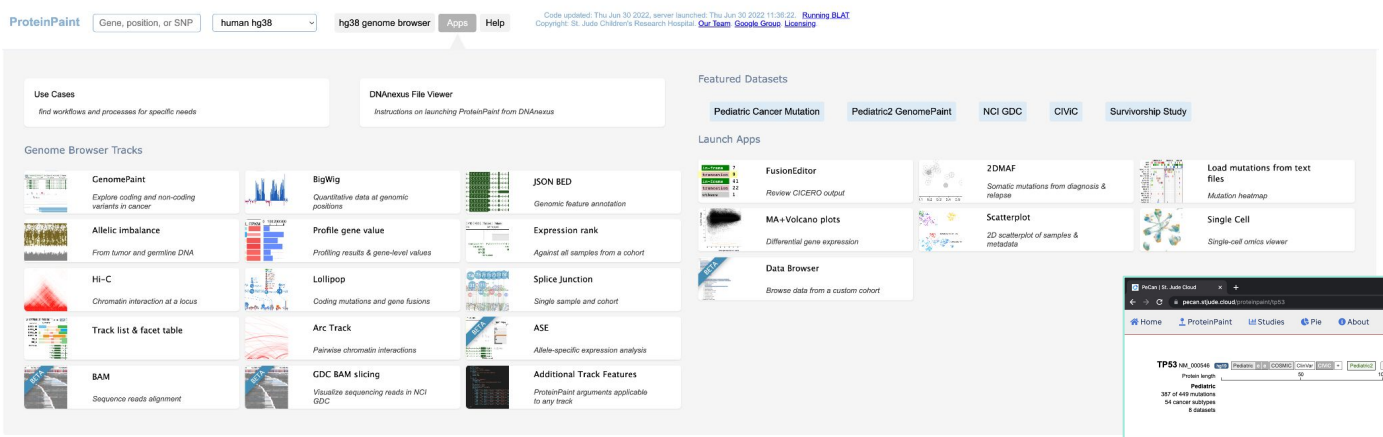


5,510 Samples
5,213 Patients
32 Diagnoses
20,537 Genes
126,414 Mutations

St. Jude Cloud Genomics Platform



St. Jude Cloud Visualization Community

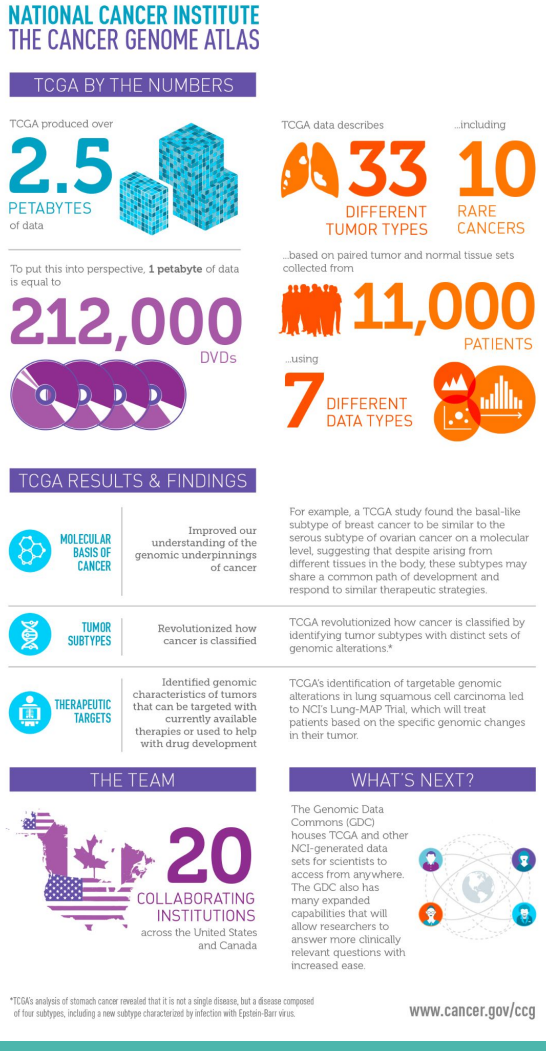


ProteinPaint

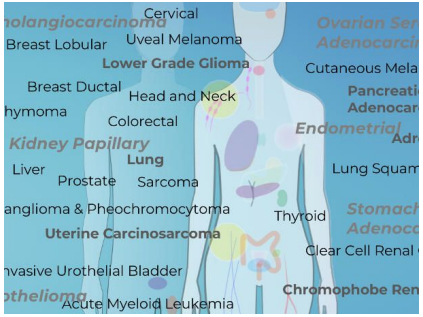
WES: The Cancer Genome Atlas (TCGA)

Analysis platforms supporting TCGA dataset:

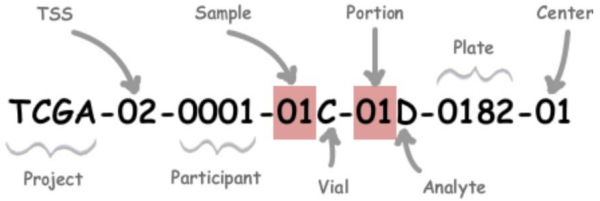
- Query based platforms: cBioPortal, FireBrowse, GEPIA, Genomic Data Commons, UCSC, TCGA Batch Effects Viewer, Tumor Map, TANRIC, SurvNet, Regulome Explorer, Xena
- Cloud-based platforms: Seven Bridges, Terra, DNANexus etc.
- Programming-based packages: TCGAbiolinks, RTCGA etc.



TCGA Cancers



TCGA Barcode



Cancer Type Studied	# Cases Characterized (# Cases in Marker Paper)	Publication	Study Abbreviation
Acute Myeloid Leukemia	200 (200)	NEJM 2013	LAML
Adrenocortical Carcinoma	92 (91)	Cancer Cell 2016	ACC
Bladder Urothelial Carcinoma	412 (131)	Nature 2014, Cell 2017	BLCA
Breast Ductal Carcinoma	778 (430)	Nature 2012	BRCA/DCIS
Breast Lobular Carcinoma	201 (127)	Cell 2015	BRCA/LCIS
Cervical Carcinoma	307 (228)	Nature 2017	CESC
Cholangiocarcinoma	51 (38)	Cell Reports 2017	CHOL
Colorectal Adenocarcinoma	633 (276)	Nature 2012	COAD
Esophageal Carcinoma	185 (164)	Nature 2017	ESCA
Gastric Adenocarcinoma	443 (295)	Nature 2014	STAD
Glioblastoma Multiforme	617 (206)	Nature 2008, Cell 2013	GBM
Head and Neck Squamous Cell Carcinoma	528 (279)	Nature 2015	HNSC
Hepatocellular Carcinoma	377 (363)	Cell 2017	LIHC
Kidney Chromophobe Carcinoma	113 (66)	Cancer Cell 2014	KICH
Kidney Clear Cell Carcinoma	537 (446)	Nature 2013	KIRC
Kidney Papillary Cell Carcinoma	291 (161)	NEJM 2016	KIRP
Lower Grade Glioma	516 (293)	NEJM 2015	LGG
Lung Adenocarcinoma	585 (230)	Nature 2014, Nature Genetics 2016	LUAD
Lung Squamous Cell Carcinoma	504 (178)	Nature 2012, Nature Genetics 2016	LUSC
Mesothelioma	74 (87)	Cancer Discovery 2018	MESO
Ovarian Serous Adenocarcinoma	608 (489)	Nature 2011	OV
Pancreatic Ductal Adenocarcinoma	185 (150)	Cancer Cell 2017	PAAD
Paraganglioma & Pheochromocytoma	179 (173)	Cancer Cell 2017	PCPG
Prostate Adenocarcinoma	500 (333)	Cell 2015	PRAD
Sarcoma	261 (206)	Cell 2017	SARC
Skin Cutaneous Melanoma	470 (331)	Cell 2015	SKCM
Testicular Germ Cell Cancer	150 (137)	Cell Reports 2018	TGCT
Thymoma	124 (117)	Cancer Cell 2018	THYM
Thyroid Papillary Carcinoma	507 (496)	Cell 2014	THCA
Uterine Carcinosarcoma	57 (57)	Cancer Cell 2017	UCS
Uterine Corpus Endometrioid Carcinoma	560 (373)	Nature 2013	UCEC
Uveal Melanoma	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	3730/3730xl DNA Analyzers	Primers	N/A
McDonnell Genome Institute at Washington University	ABI	Applied Biosystems Sequence Data	3730/3730xl DNA Analyzers	Primers	N/A
Human Genome Sequencing Center at Baylor College of Medicine	ABI	Applied Biosystems Sequence Data	3730/3730xl DNA Analyzers	Primers	N/A
University of North Carolina	AgilentG4502A_07_1	Agilent 244K Custom Gene Expression G4502A-07-1	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
University of North Carolina	AgilentG4502A_07_2	Agilent 244K Custom Gene Expression G4502A-07-2	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
University of North Carolina	AgilentG4502A_07_3	Agilent 244K Custom Gene Expression G4502A-07-3	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
Memorial Sloan Kettering Cancer Center	CGH-1x1M_G4447A	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
Broad Institute of MIT and Harvard	Genome_Wide_SNP_6	Affymetrix Genome-Wide Human SNP Array 6.0	Genome-Wide Human SNP Array 6.0	FASTA	TCGA ADF
McDonnell Genome Institute at Washington University	Genome_Wide_SNP_6	Affymetrix Genome-Wide Human SNP Array 6.0	Genome-Wide Human SNP Array 6.0	FASTA	TCGA ADF
University of North Carolina	H-miRNA_8x15K	Agilent 8 x 15K Human miRNA-specific microarray	Human mouse and rat miRNA Microarray	FASTA	TCGA ADF
University of North Carolina	H-miRNA_8x15Kv2	Agilent Human miRNA Microarray Rel12.0	Human mouse and rat miRNA Microarray	FASTA	TCGA ADF
Memorial Sloan Kettering Cancer Center	HG-CGH-244A	Agilent Human Genome CGH Microarray 244A	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
Harvard Medical School	HG-CGH-244A	Agilent Human Genome CGH Microarray 244A	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
Harvard Medical School	HG-CGH-415K_G4124A	Agilent Human Genome CGH Custom Microarray 2x415K	SurePrint G3 CGH+SNP Microarray	FASTA	TCGA ADF
McDonnell Genome Institute at Washington University	HG-U133_Plus_2	Affymetrix Human Genome U133 Plus 2.0 Array	Human Genome U133 Array	FASTA	TCGA ADF
Broad Institute of MIT and Harvard	HT_HG-U133A	Affymetrix HT Human Genome U133 Array Plate Set	Human Genome U133 Array	TSV	TCGA ADF
Berkeley Lab	HuEx-1_0-st-v2	Affymetrix Human Exon 1.0 ST Array	Exon 1.0 ST Array	FASTA	TCGA ADF
HudsonAlpha Institute for Biotechnology	Human1MDuo	Illumina Human1M-Duo BeadChip	Infinium HD BeadChip DNA	FASTA	TCGA ADF
HudsonAlpha Institute for Biotechnology	HumanHap550	Illumina 550K Infinium HumanHap550 SNP Chip	Infinium BeadChip DNA	FASTA	TCGA ADF
Johns Hopkins-USC Epigenome Center	HumanMethylation27	Illumina Infinium Human DNA Methylation 27	Infinium Human DNA Methylation 27	See ADF File	TCGA ADF
Johns Hopkins-USC Epigenome Center	HumanMethylation450	Illumina Infinium Human DNA Methylation 450	Infinium Human DNA Methylation 450	See ADF File	TCGA ADF
Johns Hopkins-USC Epigenome Center	IlluminaDNAMethylation_OMA002_CPI	Illumina DNA Methylation OMA002 Cancer Panel I	GoldenGate DNA Methylation	FASTA	TCGA ADF
Johns Hopkins-USC Epigenome Center	IlluminaDNAMethylation_OMA003_CPI	Illumina DNA Methylation OMA003 Cancer Panel I	GoldenGate DNA Methylation	See ADF File	TCGA ADF
McDonnell Genome Institute at Washington University	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	Genome Analyzer IIx	N/A	N/A
Broad Institute of MIT and Harvard	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	Genome Analyzer IIx	N/A	N/A
Human Genome Sequencing Center at Baylor College of Medicine	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	Genome Analyzer IIx	N/A	N/A
University of California Santa Cruz	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	Genome Analyzer IIx	N/A	N/A
University of North Carolina	IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	Genome Analyzer IIx	N/A	N/A
BC Cancer Agency	IlluminaGA_miRNASeq	Illumina Genome Analyzer miRNA Sequencing	Genome Analyzer IIx	N/A	N/A
Harvard Medical School	IlluminaGA_mRNA_DGE	Illumina Genome Analyzer mRNA Digital Gene Expression	Genome Analyzer IIx	N/A	N/A
BC Cancer Agency	IlluminaGA_RNASeq	Illumina Genome Analyzer RNA Sequencing	Genome Analyzer IIx	N/A	N/A
University of North Carolina	IlluminaGA_RNASeqV2	Illumina Genome Analyzer RNA Sequencing Version 2 analysis	Genome Analyzer IIx	N/A	N/A
Harvard Medical School	IlluminaHiSeq_DNASeqC	Illumina HiSeq for Copy Number Variation	HiSeq 2000	N/A	N/A
BC Cancer Agency	IlluminaHiSeq_miRNASeq	Illumina HiSeq 2000 miRNA Sequencing	HiSeq 2001	N/A	N/A
University of North Carolina	IlluminaHiSeq_RNASeq	Illumina HiSeq 2000 RNA Sequencing	HiSeq 2002	N/A	N/A
University of North Carolina	IlluminaHiSeq_RNASeqV2	Illumina HiSeq 2000 RNA Sequencing Version 2 analysis	HiSeq 2003	N/A	N/A
University of North Carolina	IlluminaHiSeq_TotalRNASeqV2	Illumina HiSeq 2000 Total RNA Sequencing Version 2 analysis	HiSeq 2004	N/A	N/A
Johns Hopkins-USC Epigenome Center	IlluminaHiSeq_WGBS	Illumina HiSeq 2000 Bisulfite-converted DNA Sequencing	HiSeq 2005	N/A	N/A
MD Anderson Cancer Center	MDA_RPPA_Core	M.D. Anderson Reverse Phase Protein Array Core	MD Anderson RPPA Core Facility	N/A	N/A
Nationwide Children's Hospital	microsat_i	Microsatellite Instability Analysis	SOLiD 3 Plus	N/A	N/A
University of California Santa Cruz	SOLiD_DNASeq	ABI SOLiD DNA System Sequencing	SOLiD 3 Plus	N/A	N/A
Broad Institute of MIT and Harvard	SOLiD_DNASeq	ABI SOLiD DNA System Sequencing	SOLiD 3 Plus	N/A	N/A
Human Genome Sequencing Center at Baylor College of Medicine	SOLiD_DNASeq	ABI SOLiD DNA System Sequencing	SOLiD 3 Plus	N/A	N/A

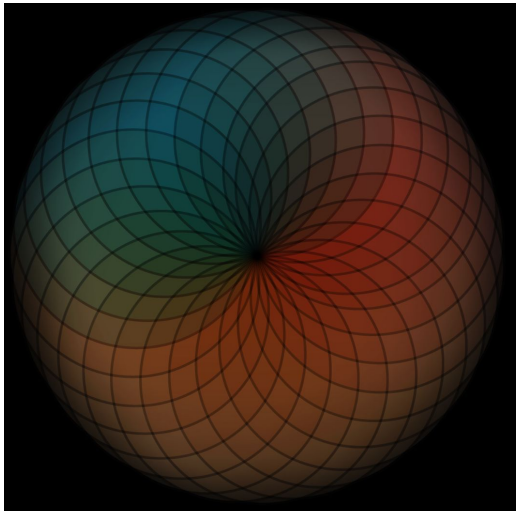
https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/technology

TCGA Findings

TCGA outcomes & impact

TCGA's Pan-Cancer Atlas

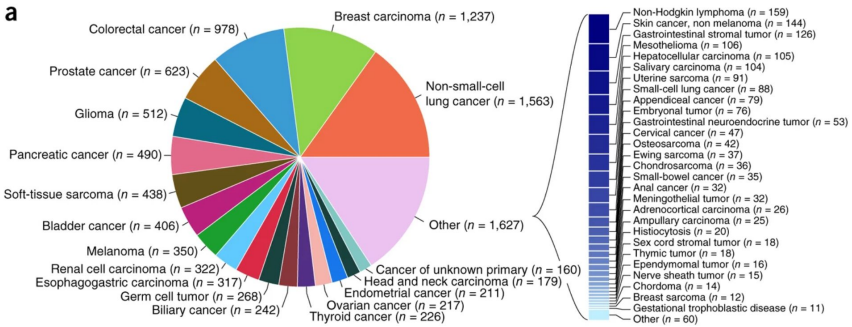
TCGA Research Network Publications

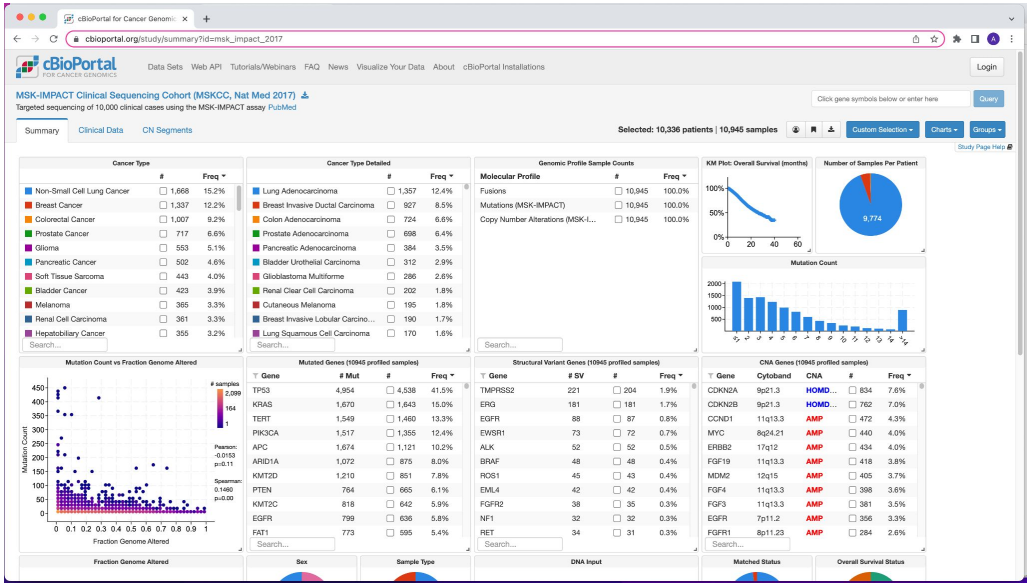


TS:MSK-IMPACT

- 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**

10,945 Tumor Samples
10,336 Individuals
62 Principal Tumor Types
>300 Detailed Tumor Types

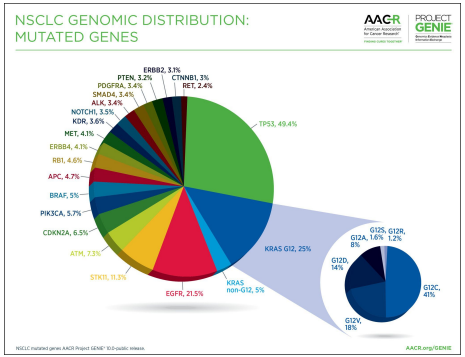


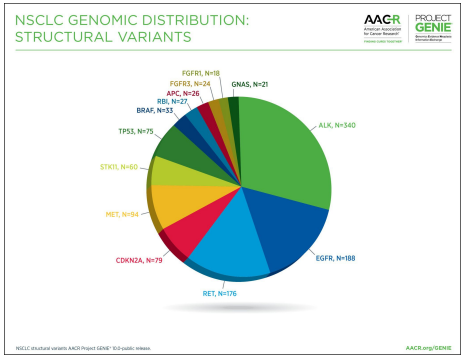


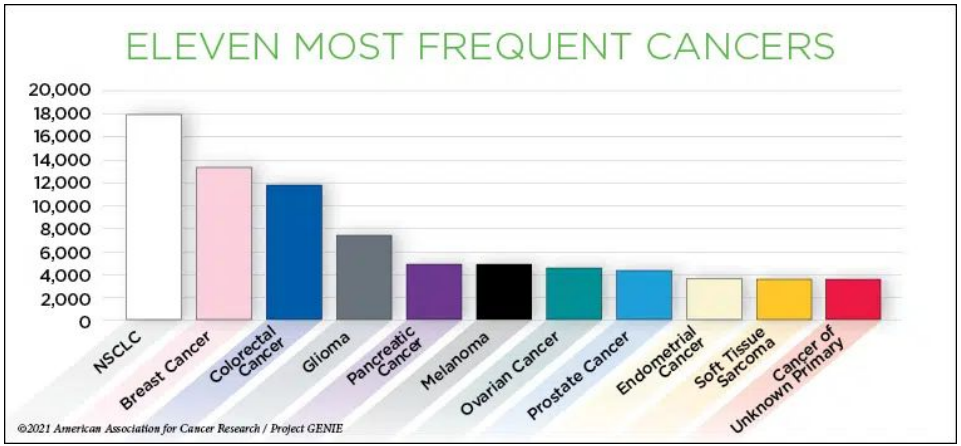
TS: Genomics Evidence Neoplasia Information Exchange (GENIE)

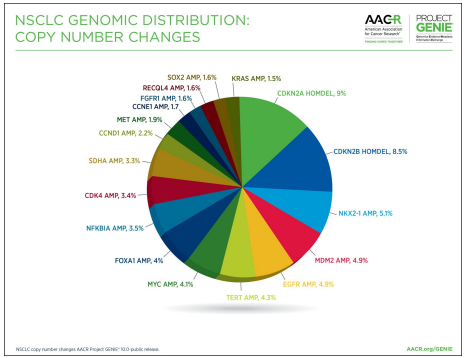
The AACR Project GENIE is an international data-sharing consortium focused on generating an evidence base for precision cancer medicine by integrating clinical-grade cancer genomic data with clinical outcome data for tens of thousands of cancer patients treated at multiple institutions worldwide.

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 (HCMI)	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 transciptomic 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.

Genomic Data Commons (GDC)

Proteomic Data Commons (PDC)

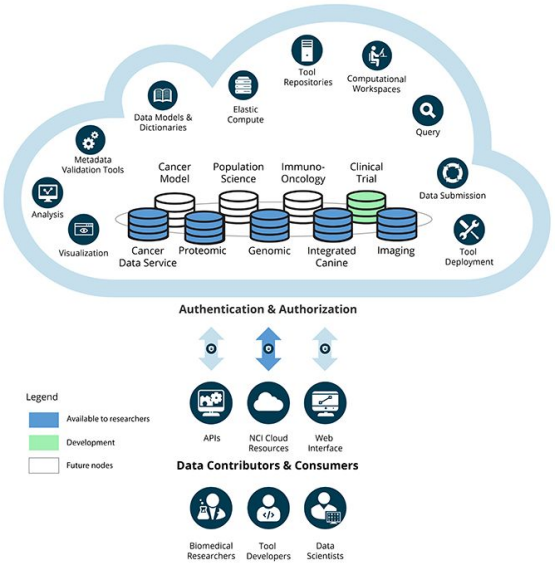
Integrated Canine Data Commons (ICD

Imaging Data Commons (IDC)

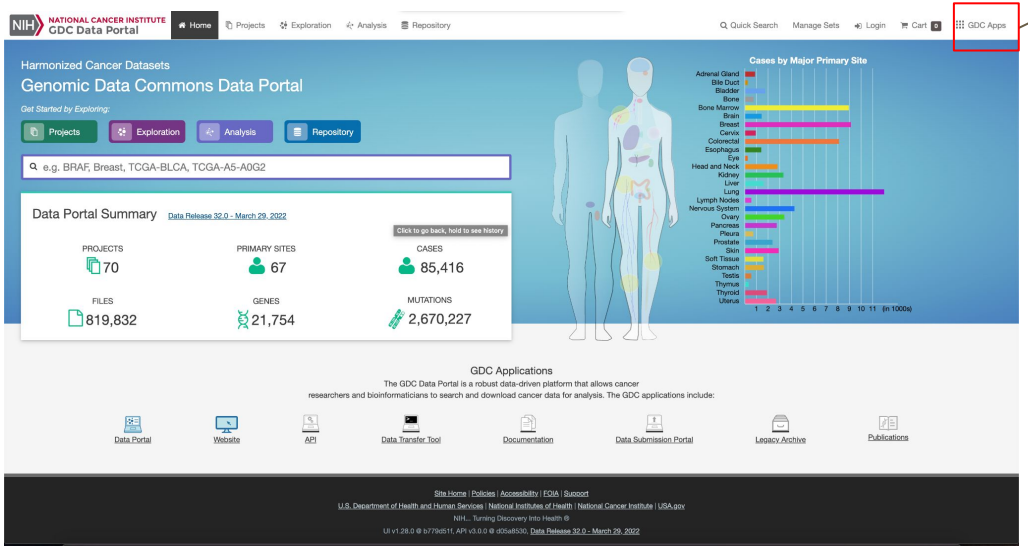
Cancer Data Commons (CDS)

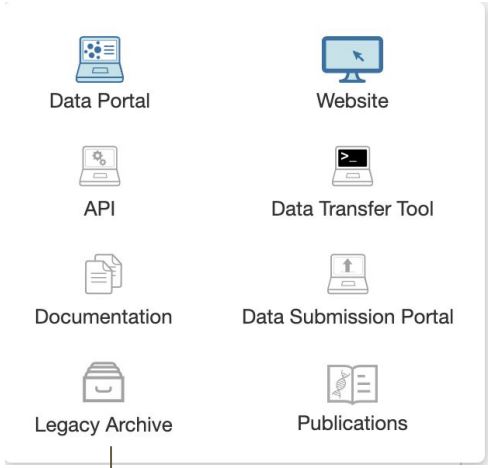
NCI Cloud Resources

NCI Cancer Research Data Commons (CRDC)



GDC Data Portal





TCGA hg19 version

[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

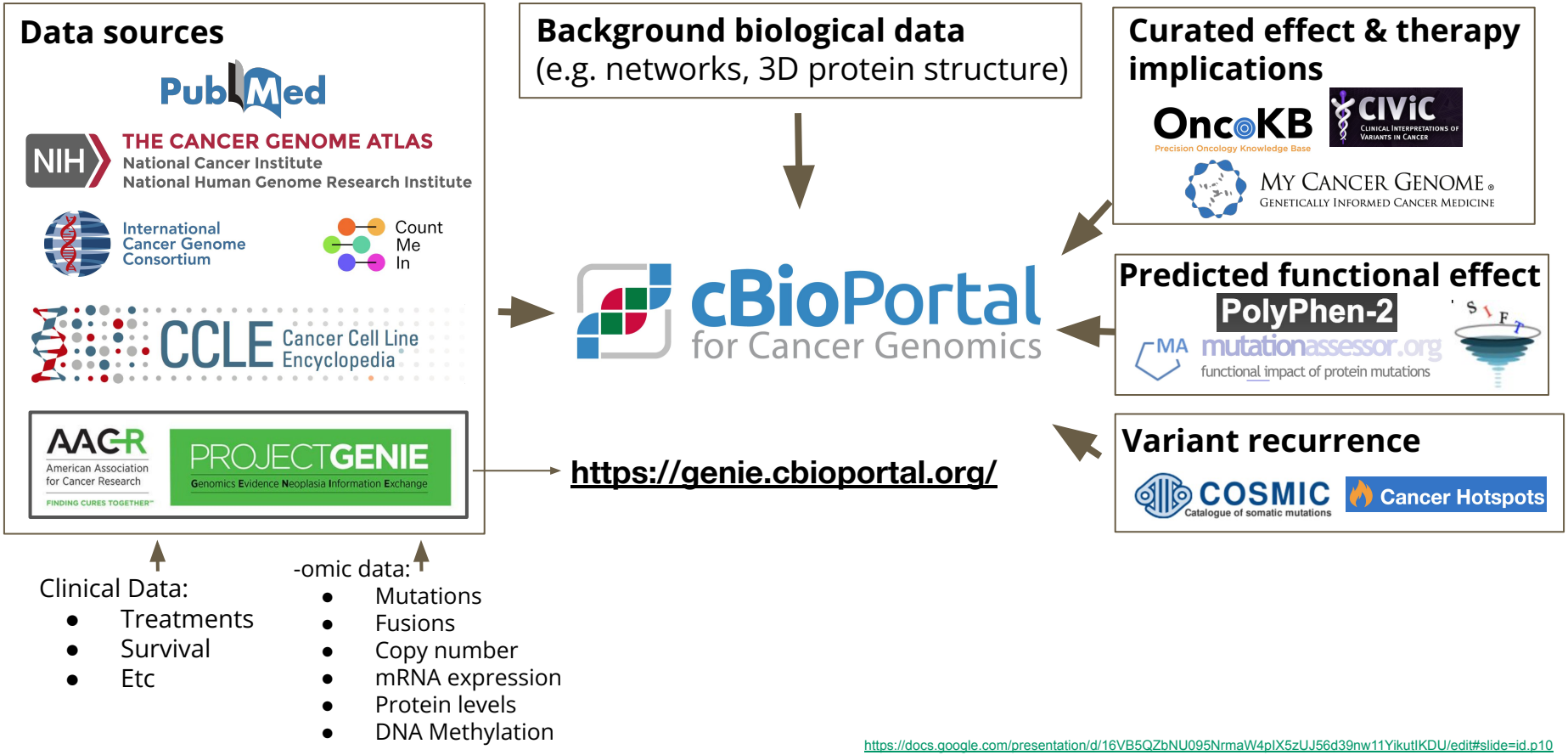
Major features:

- 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 !!

cBioPortal for Cancer Genomics

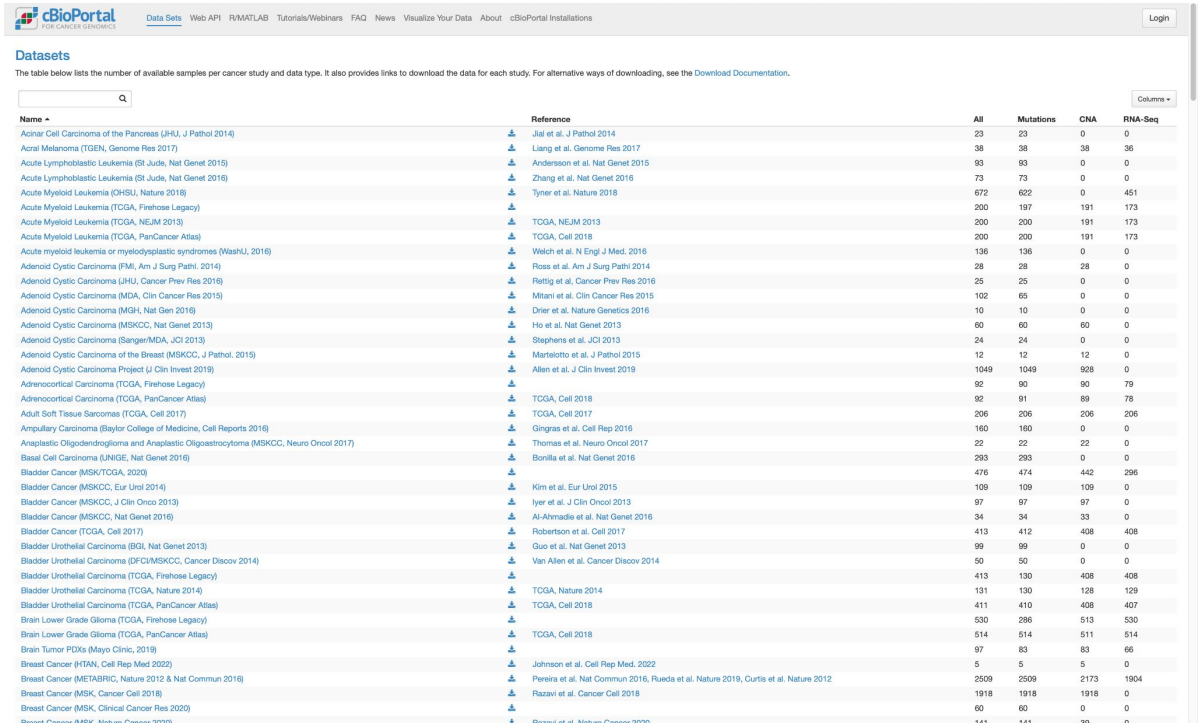
What data is in cBioPortal?

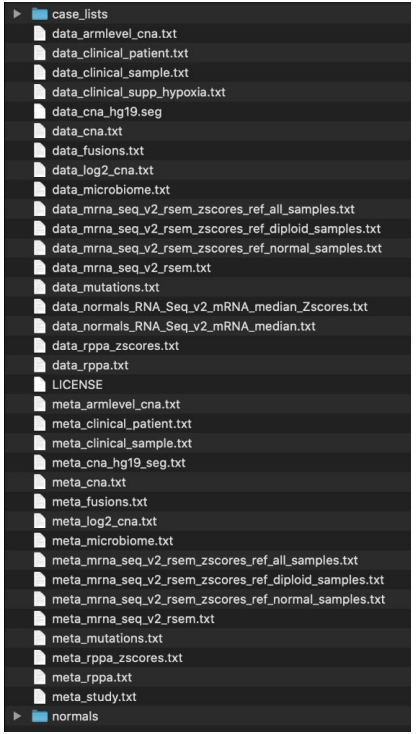


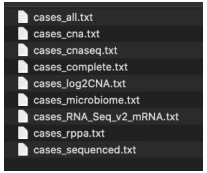
https://docs.google.com/presentation/d/16VB5QZbNU095NrmaW4pIX5zUJ56d39nw11YikutIKDU/edit#slide=id.p10

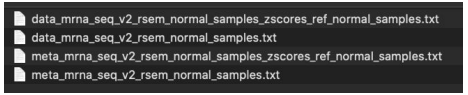
Download Datasets

luad_tcga_pan_can_atlas_2018.tar.gz

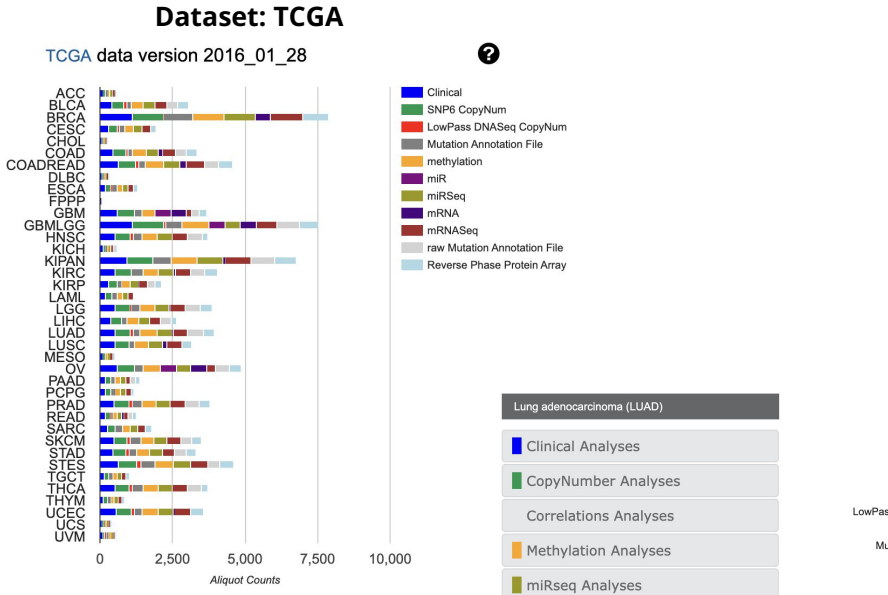


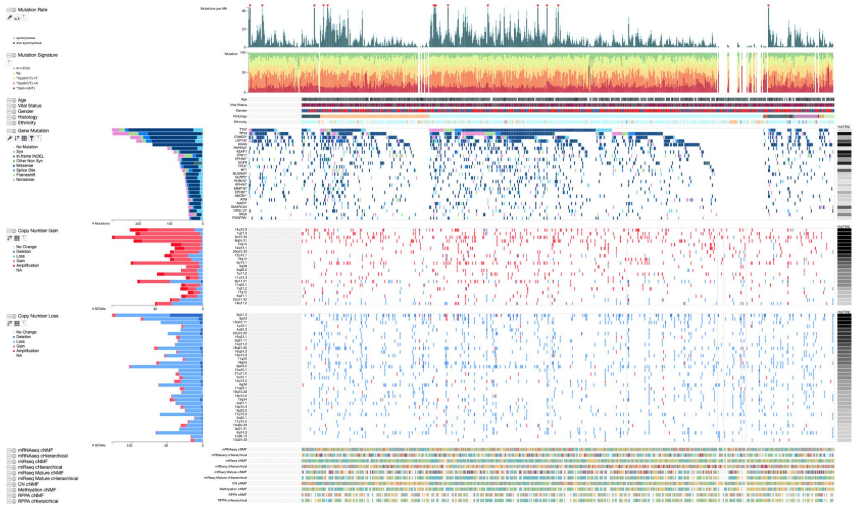


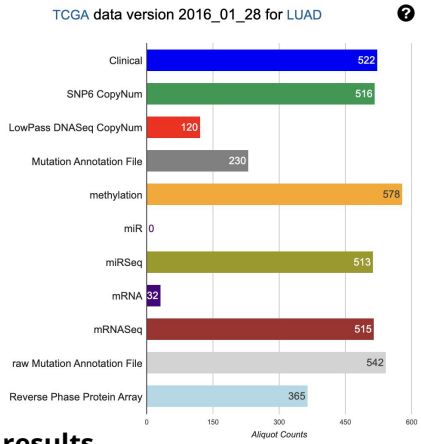




FIREBROWSE (Broad GDAC)







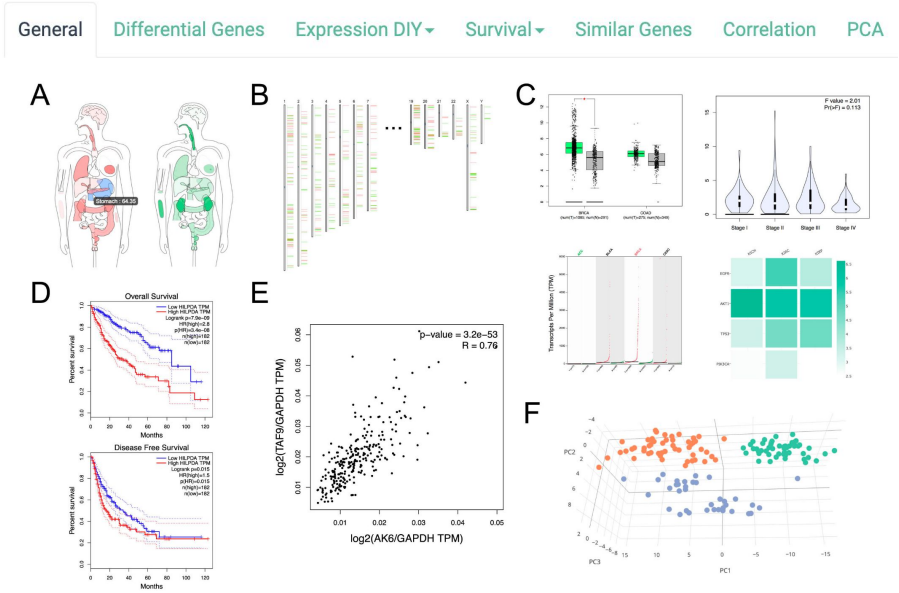
[iCoMut for FireBrowse]

Major genomic analysis results

Gene Expression Profiling Interactive Analysis (GEPIA)

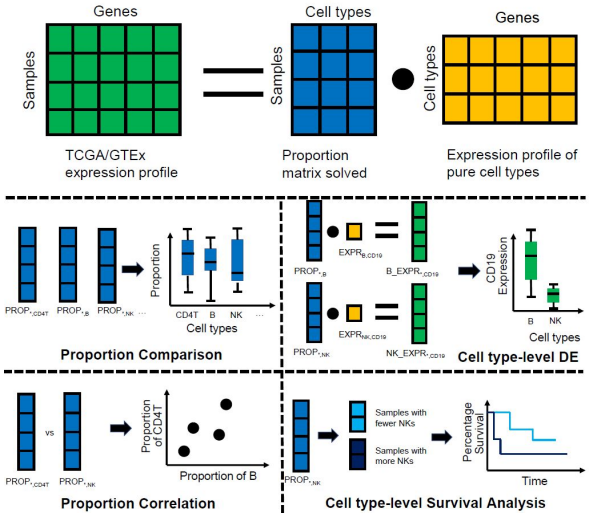
Datasets: **TCGA** and **GTEx** data

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



GEPIA2

GEPIA2021

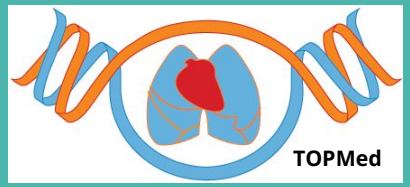


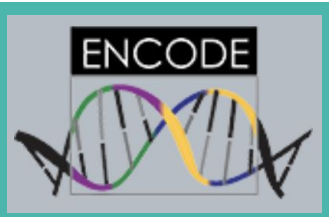
a standalone extension with multiple deconvolution based analysis for GEPIA. We deconvolute each sample tool in TCGA/GTEx with the bioinformatics tools CIBERSORT, EPIC and quanTIseq. Based on the inferred cell proportions in each bulk-RNA sample, we can then perform multiple downstream analysis:

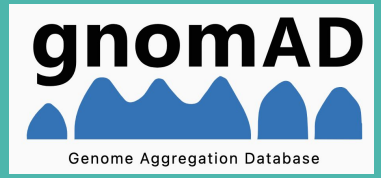
Summary of Data Portals

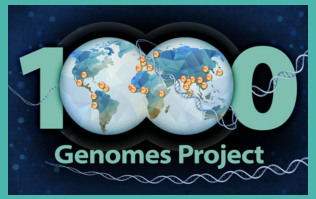
Data Portals	Studies	Data type	Interactive analysis	Integrative analysis	Visualization	Access Control	Recommendation
GEPIA	TCGA; GTEx	Expression data	Yes	No	Limited	No	Tumor-Normal expression analysis
GTEx Portal	GTEx	Gene expression and genotyping data	Limited	Limited	Yes	Raw data and most analyzed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

Other Genetic Data Resources









Genotype-Tissue Expression (GTEx)

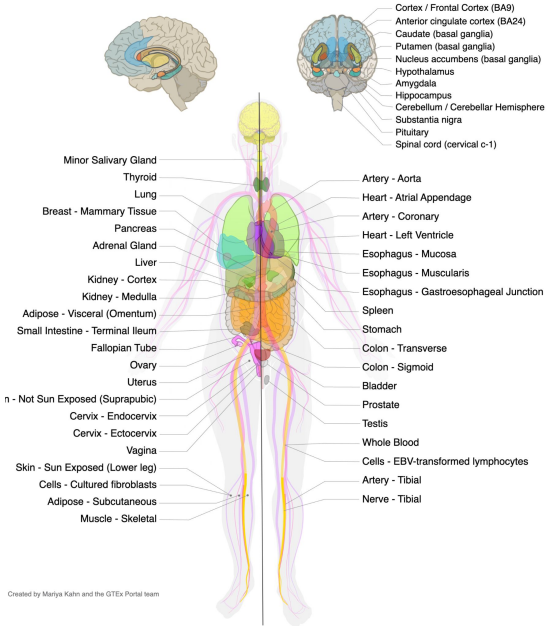
Public resource of tissue-specific gene expression

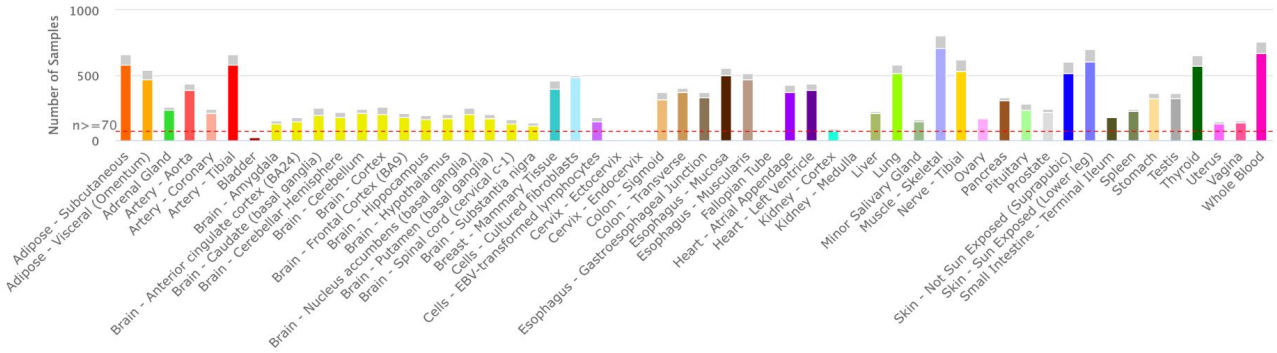
Samples collected from **54** non-diseased tissue sites across over 900 individuals

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





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

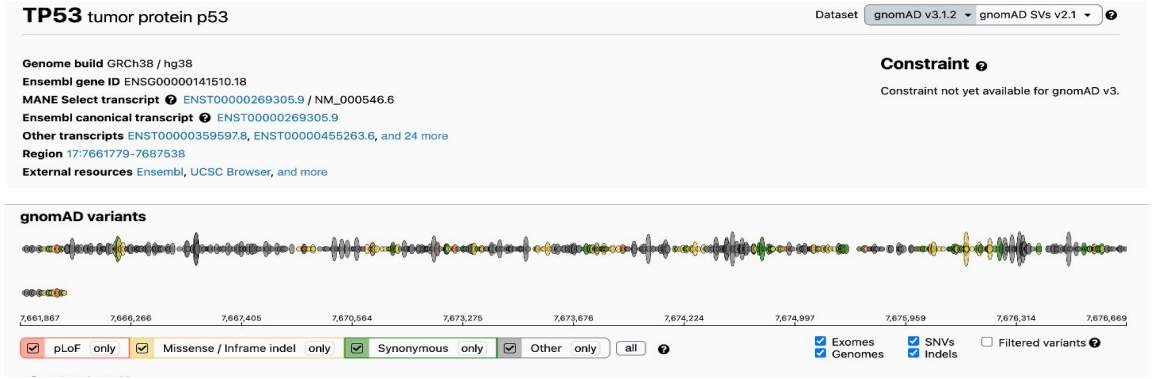
* Number of samples with genotype >= 70

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:

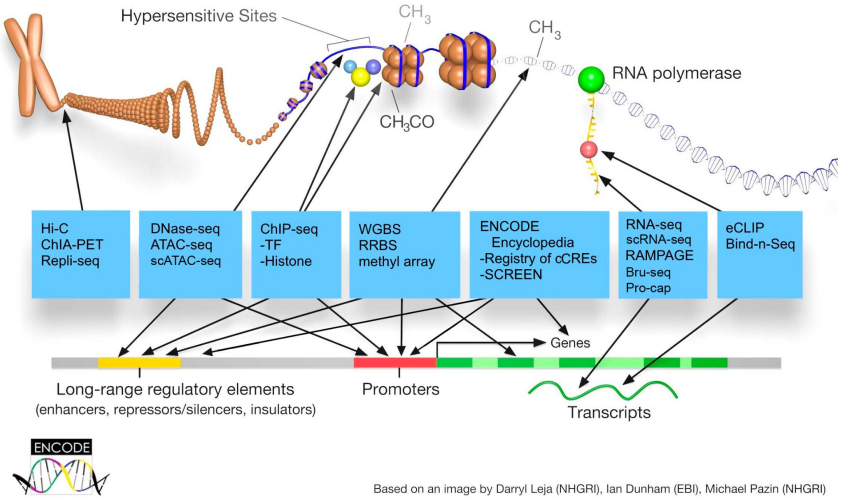
- gnomAD v2.1 data set contains data from **125,748 exomes** and **15,708 whole genomes,** all mapped to the **GRCh37/hg19** reference sequence.
- gnomAD v3.1 data set contains **76,156 whole genomes**, all mapped to the **GRCh38** reference sequence.
- gnomAD v3.1 contains a substantially larger number 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)

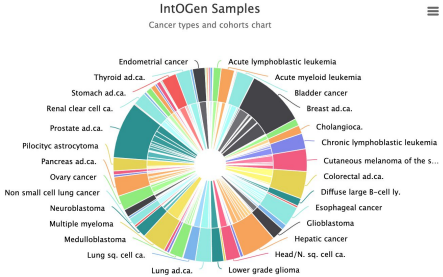
Specialized Databases for Genomic Analyses

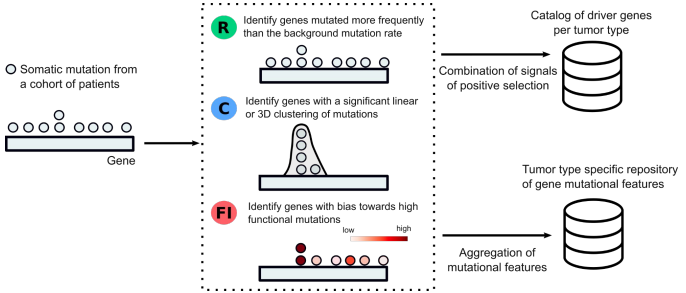
Driver Gene Analysis

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

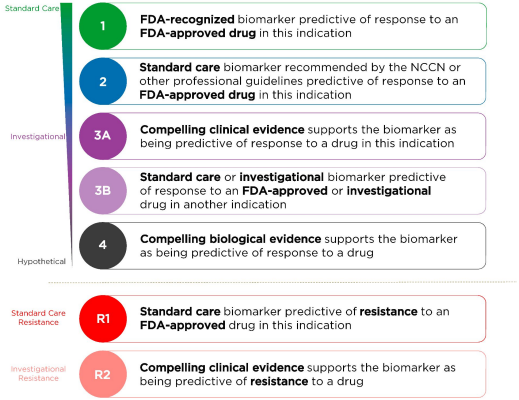
MSK's Precision Oncology Knowledge Base

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

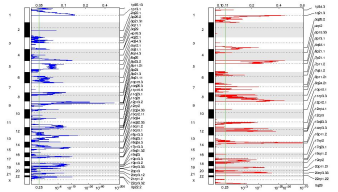
Single residue and in-frame indel mutation hotspots in 24,592 tumor samples

Can download hotspot results and mutation data (MAF)

FIREBROWSE (Broad GDAC) for Somatic Analysis

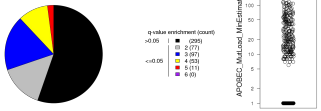
CopyNumber Analyses

- 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



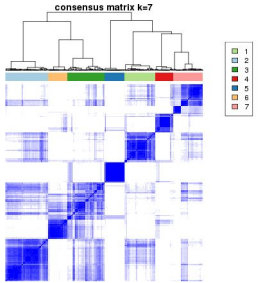
Mutation Analyses

- 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

[TSVdb](#)

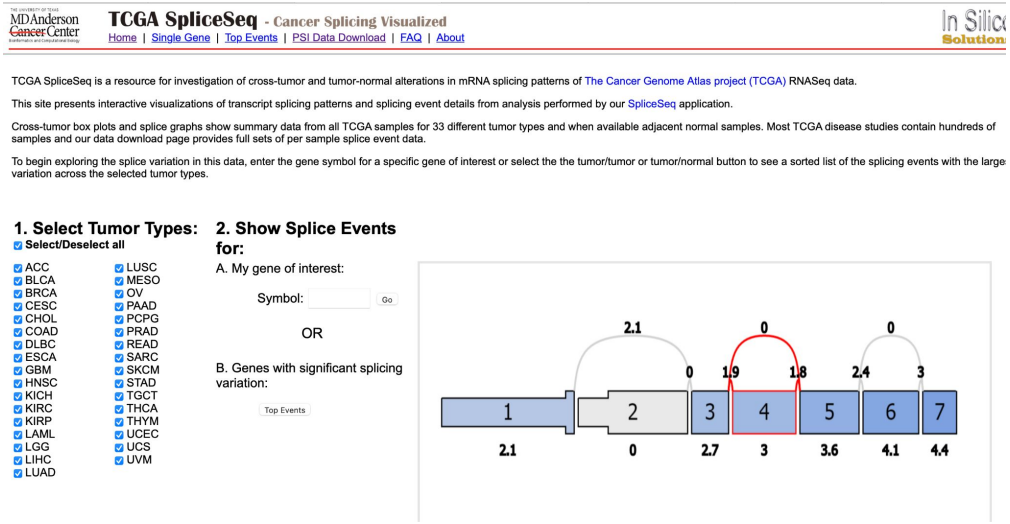
TCGA Splicing Variants DB

OncoSplicing

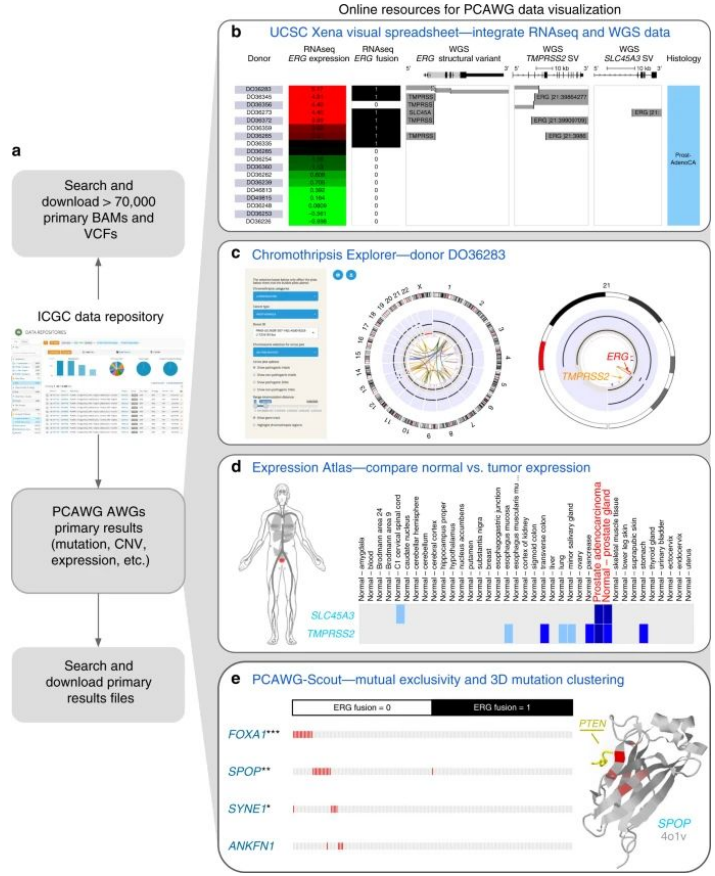
OncoSplicing is a database to systematically study clinically relevant alternative splicing in 33 TCGA cancers and 31 GTEx tissues.

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



PCAWG Data Portal

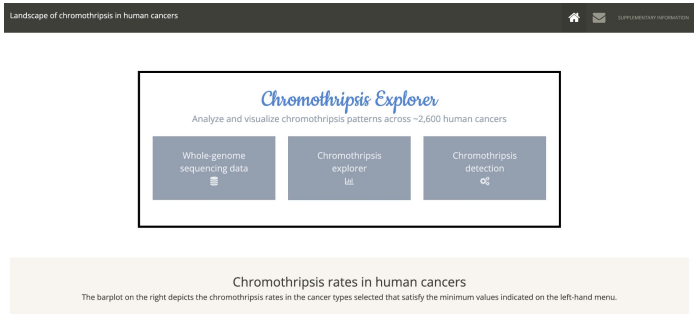
DCC / PCAWG / Filter by file name...

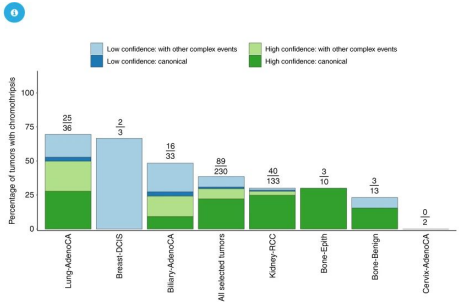
Name

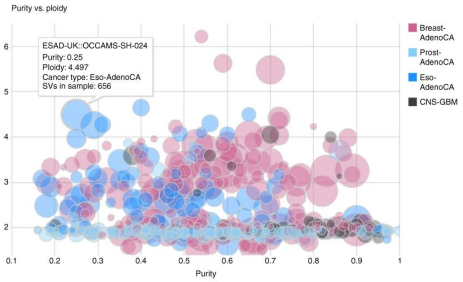
- README.md
- APOBEC_mutagenesis
- benchmarking_data
- cell_lines
- clinical_and_histology
- consensus_cnv
- consensus_snv_indel
- consensus_sv
- data_releases
- donors_and_biospecimens
- driver_mutations
- drivers
- evolution_and_heterogeneity
- germline_variations
- Hartwig
- hla_and_neoantigen
- msi
- mutational_signatures
- networks
- pathogen_analysis
- pcawg_dkfz_caller
- quality_control_info
- reference_data
- retrotransposition
- sequencing_metadata
- smufin_indel_calls
- subclonal_reconstruction
- terminology_and_standard_colours
- thesaurus_snv
- transcriptome
- unaligned_bams

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

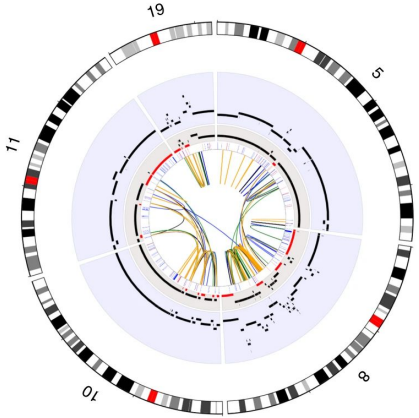
Chromothripsis Explorer

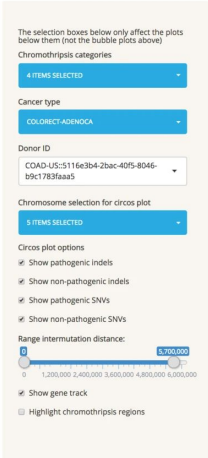


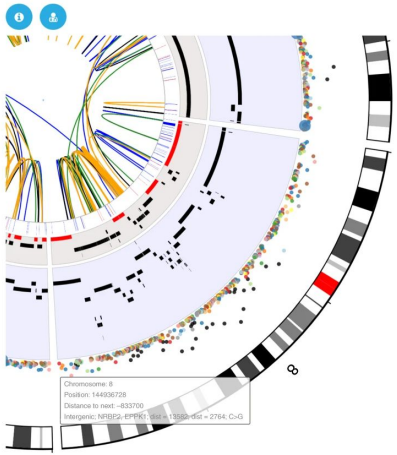




Analyze and visualize chromothripsis patterns across ~2,600 human cancers

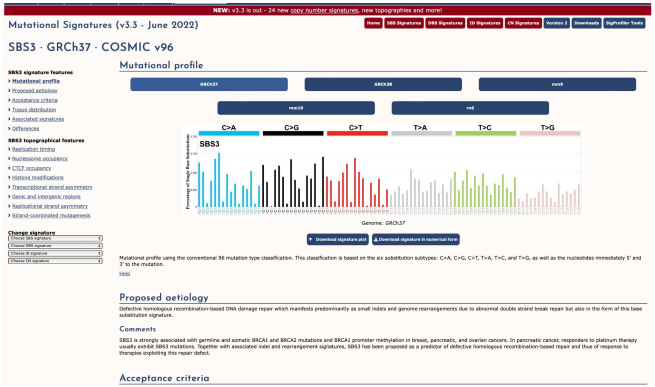






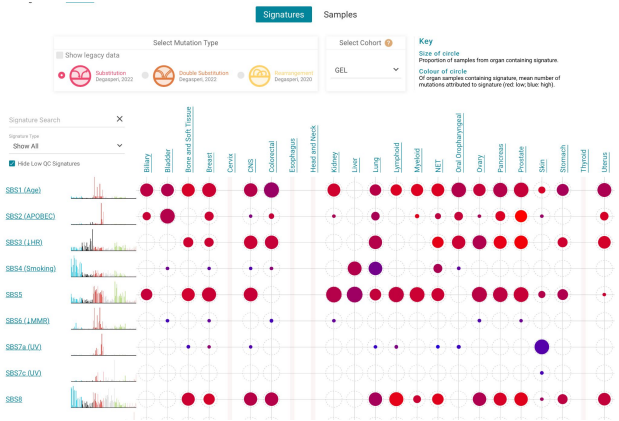
Mutational Signature Analysis

COSMIC Mutational Signatures



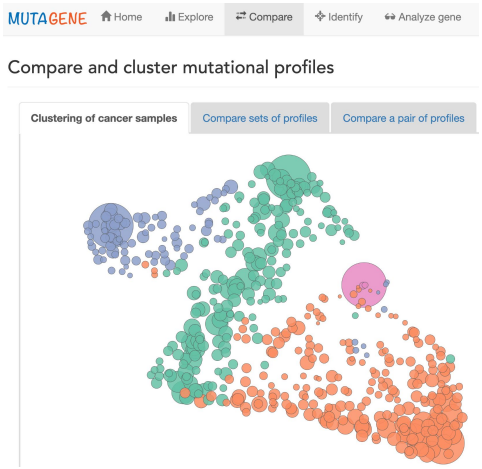
- Support different mutational profiles (SBS/ID/DBS/CNV)
- Detail annotations (proposed aetiology and acceptance criteria, tissue distribution, signature associations, version difference)
- Topographies associations (Replication timing, Nucleosome occupancy, CTCF occupancy, Histone modifications, Transcriptional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

Signal



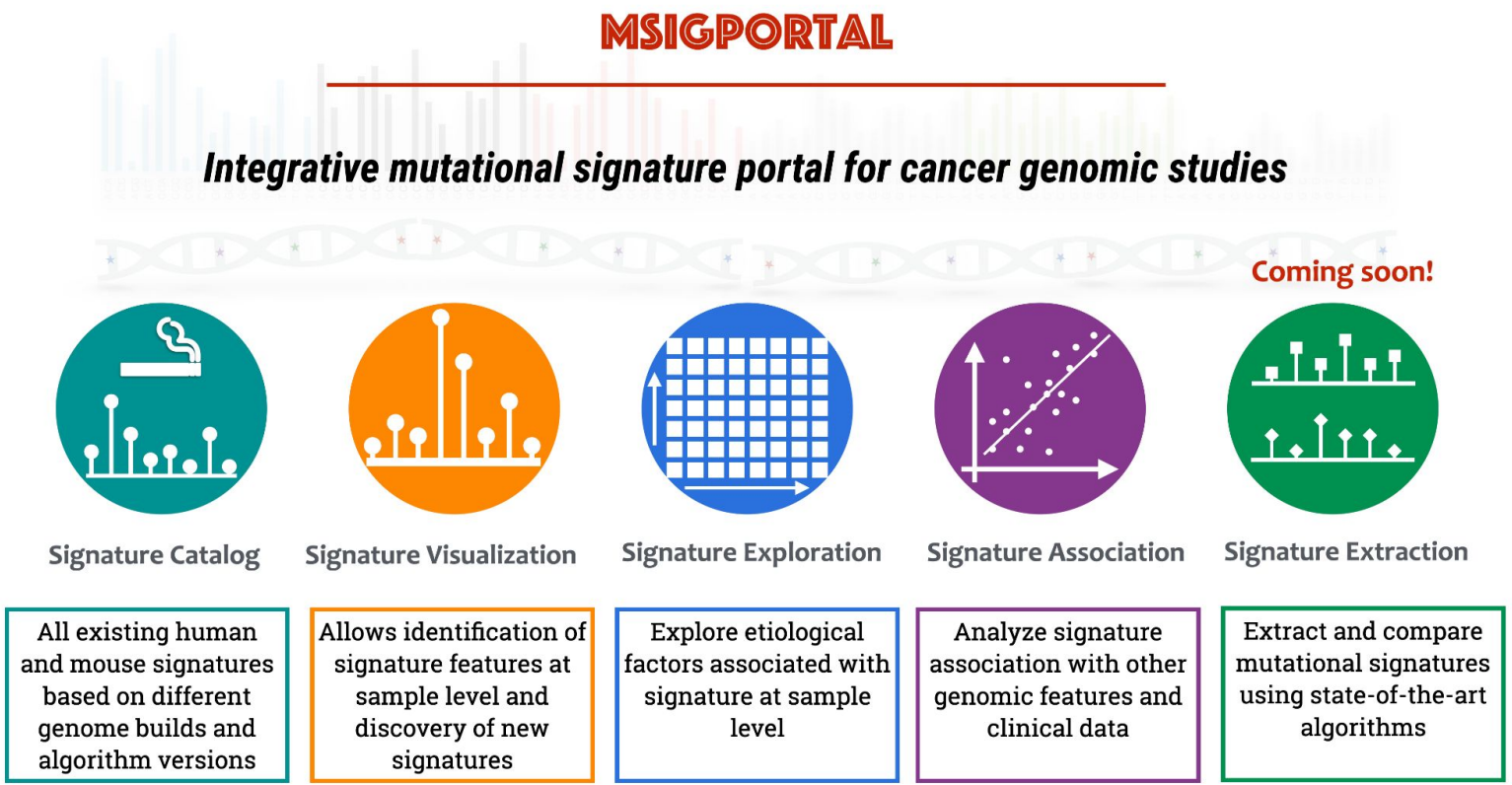
- Explore different source of signatures (cancer, gene edits, environmental mutagenesis)
- Support different mutational profiles (SBS/ID/DBS/SV)
- Support signature data analyses
- Include largest cancer genomic studies, allow to discovery the rare mutational signatures in cancer
- Enhanced signature and study visualizations with API available

MUTAGENE



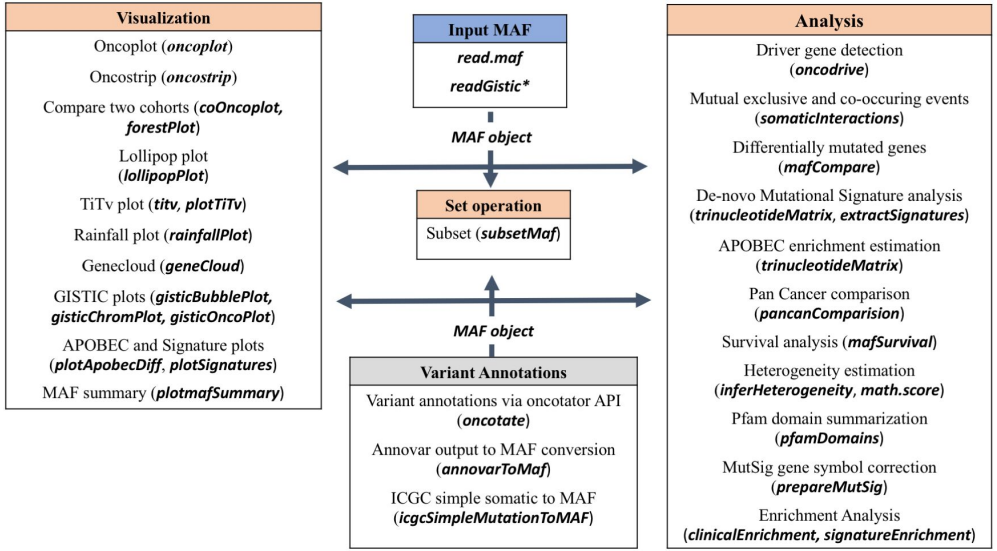
- Explore context-dependent mutational profiles and signatures
- Compare and cluster mutational profiles
- Identify mutational processes based on NMF algorithm
- Compares observed mutational frequencies to expected background mutability to identify potential drivers

mSigPortal: *Integrative mutational signature portal for cancer genomic studies*



Analytical Programming Packages for Cancer Genomic Datasets

MAFtools: Summarize, Analyze and Visualize MAF Files



TCGAmutations - An R data package for TCGA somatic mutations

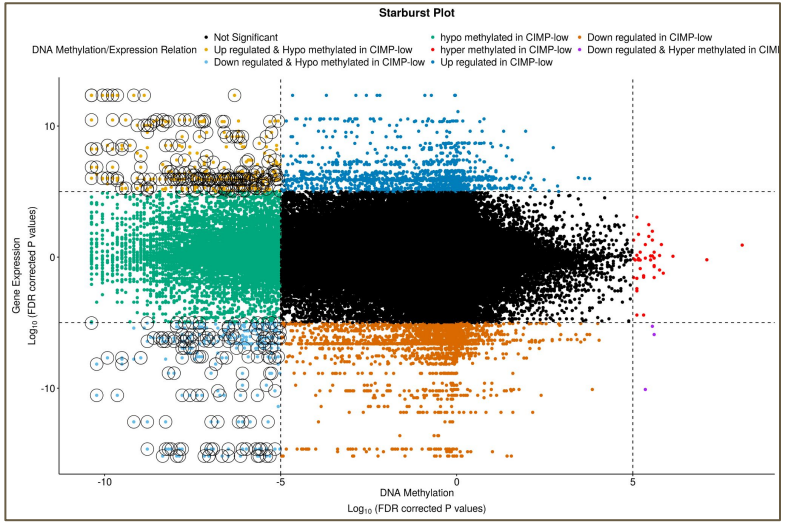
Data source:
Broad firehose
TCGA MC3
(Multi-Center Mutation Calling in Multiple Cancers)

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

TCGAbiolinks

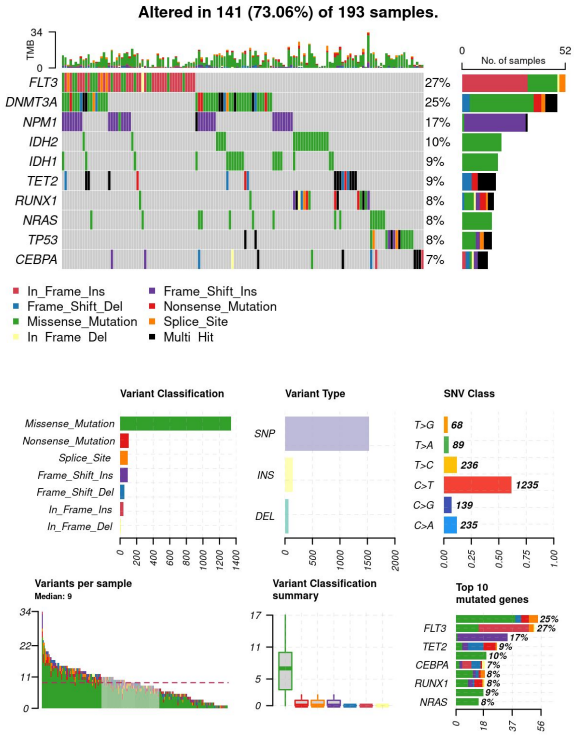
HTML	R Script	1. Introduction
HTML	R Script	10. Classifiers
HTML	R Script	10. TCGAbiolinks_Extension
HTML	R Script	11. Stemness score
HTML	R Script	2. Searching GDC database
HTML	R Script	3. Downloading and preparing files for analysis
HTML	R Script	4. Clinical data
HTML	R Script	5. Mutation data
HTML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



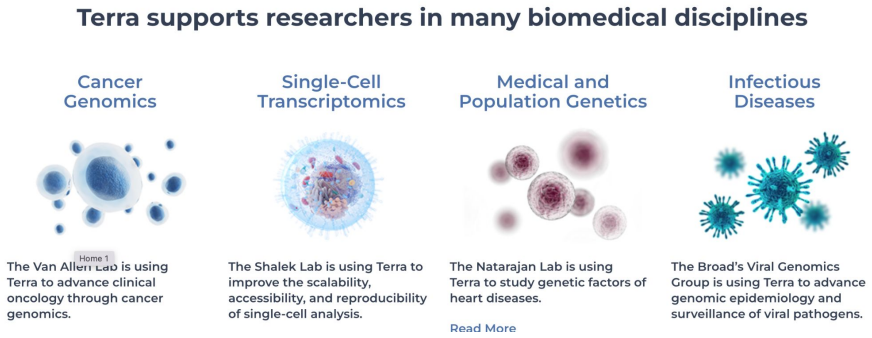
cBioPortalData

- 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

Terra

- 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)
- Easily shareable workspaces



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)

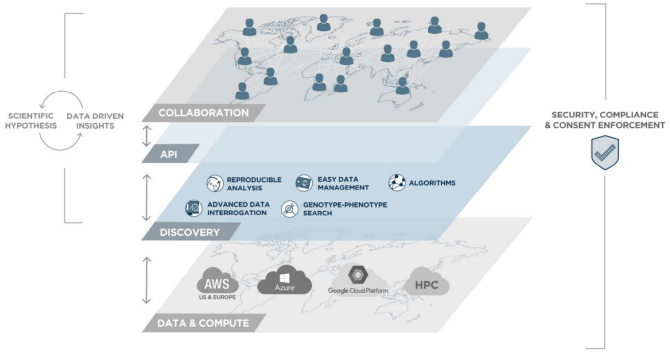


Figure 3. Overview of the Seven Bridges Platform. A Cloud-based data and compute infrastructure underlies the discovery layer, which is built around features to streamline data management, search, and analysis. An application programming interface and collaboration features ensure flexibility for users.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

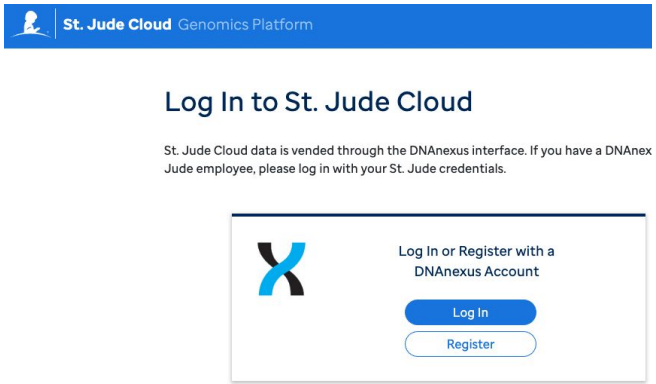
Datasets included:

- 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)

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



“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 ChiP-Seq
A curated list of CHiP-Seq analysis

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 mutational signature resource in mSigPortal

Original Research Papers Including Specific Mutational Signatures in mSigPortal

Search | Columns

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

Search | Columns

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 canc

10 | Showing 1 to 10 of 23 entries

Computational Methods, Tools, Databases or Websites for Mutational Signature Analyses

Search | Columns

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/fcnnls	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 | << < > >> | Page: 1

THANKS FOR YOUR ATTENTION!
Questions?

Next: Practical session 2 (10:45am)