The Cancer Genomics Atlas (TCGA)

Mikhail Dozmorov

Spring 2018

The Cancer Genome Atlas (TCGA)

- Started December 13, 2005, phase II in 2009, ended in 2014
- Mission - to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.
- Data generation
  - Clinical information about participants
  - Metadata about the samples (e.g. the weight of a sample portion, etc.)
  - Histopathology slide images from sample portions
  - Molecular information derived from the samples (e.g. mRNA/miRNA expression, protein expression, copy number, etc.)

https://cancergenome.nih.gov/

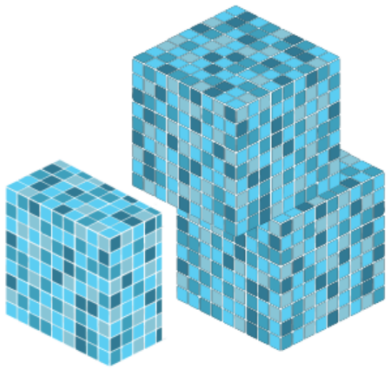
TCGA by the numbers

TCGA produced over

2.5

PETABYTES

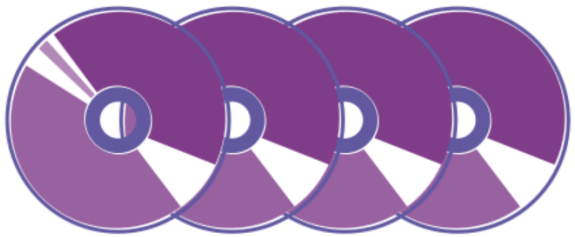
of data



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

https://cancergenome.nih.gov/abouttcga

Major TCGA Research Components

- **Biospecimen Core Resource (BCR)** - Collect and process tissue samples
- **Genome Sequencing Centers (GSCs)** - Use high-throughput Genome Sequencing to identify the changes in DNA sequences in cancer
- **Genome Characterization Centers (GCCs)** - Analyze genomic and epigenomic changes involved in cancer
- **Data Coordinating Center (DCC)** - The TCGA data are centrally managed at the DCC
- **Genome Data Analysis Centers (GDACs)** - These centers provide informatics tools to facilitate broader use of TCGA data.

TCGA Data Access Policy

- An access control policy is in place for TCGA data to ensure that personally identifiable information is kept from unauthorized users.
- **Open access** - Houses data that cannot be aggregated to generate a data set unique to an individual. This tier does not require user certification for data access.
- **Controlled access** - Houses individually-unique information that could potentially be used to identify an individual. This tier requires user certification for data access.

TCGA Controlled Access Data

Access to controlled data is available to researchers who:

- Agree to restrict their use of the information to biomedical research purposes only
- Agree with the statements within TCGA Data Use Certification (DUC)
- Have their institutions certifiably agree to the statements within TCGA DUC
- Complete the Data Access Request (DAR) form and submit it to the Data Access Committee to be a TCGA Approved User. This form is available electronically through dbGaP.

https://wiki.nci.nih.gov/display/TCGA/TCGA+Home

TCGA data types

Omics-Profile	data.type	type	Description
Gene Expression	mRNA_Array	**G450**	Microarray, Agilent 244K Custom Gene Expression G4502A-07
	mRNA_Array	U133	Microarray, Affymetrix Human Genome U133A 2.0 Array
	mRNA_Array	Huex	Microarray, Affymetrix Human Exon 1.0 ST Array
	RNASeq	**count**	RNA-Seq, raw counts
	RNASeq	RPKM	RNA-Seq, normalized counts
	RNASeq2		RNA-Seq second analysis pipeline, RSEM
miRNA expression	miRNA_Array		miRNA array, Agilent 8 x 15K Human miRNA-specific microarray (H-miRNA_8x15K)
	miRNA_Array		miRNA array, Agilent Human miRNA Microarray Rel12.0 (H-miRNA_8x15Kv2- for OV only).
	miRNASeq	**count**	miRNA-Seq, raw counts
	miRNASeq	rpmmm	miRNA-Seq, reads per million miRNA mapped
Mutation	Mutation	**somatic**	Somatic non-silent mutations
	Mutation	all	All mutations called
Methylation	Methylation	**27K**	Illumina Infinium HumanMethylation27 BeadChip
		450K	Illumina Infinium HumanMethylation450 BeadChip
Copy number changes	CNA_SNP		CNA, Affymetrix Genome-Wide Human SNP Array 6.0
	CNV_SNP		CNV, Affymetrix Genome-Wide Human SNP Array 6.0
	CNA_CGH	**415K**	CNA, Agilent Human Genome CGH Custom Microarray 2x415K
	CNA_CGH	244A	CNA, Agilent Human Genome CGH Microarray 244A

http://www.liuzlab.org/TCGA2STAT/DataPlatforms.pdf

TCGA cancer types

Cancer name	Acronym	RNASeq V2	RNASeq	miRNASeq	CNA_SNP	CNV_SNP	CNA_CGH	Methylation (27K)	Methylation (450K)	Mutation	mRNA_Array	miRNA_Array
Adrenocortical carcinoma	**ACC**	Y		Y	Y	Y			Y	Y		
Bladder urothelial carcinoma	**BLCA**	Y	Y	Y	Y	Y			Y	Y		
Breast invasive carcinoma	**BRCA**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Cervical and endocervical cancers	**CESC**	Y		Y	Y	Y			Y	Y		
Cholangiocarcinoma	**CHOL**	Y		Y	Y	Y			Y			
Colon adenocarcinoma	**COAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Colorectal adenocarcinoma	**COADREAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	**DLBC**	Y		Y	Y	Y			Y			
Esophageal carcinoma	**ESCA**		Y	Y	Y	Y			Y			
FFPE Pilot Phase II	**FPPP**			Y								
Glioblastoma multiforme	**GBM**	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Glioma	**GBMLGG**	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Head and Neck squamous cell carcinoma	**HNSC**	Y	Y	Y	Y	Y			Y	Y		
Kidney Chromophobe	**KICH**	Y		Y	Y	Y			Y	Y		
Pan-kidney cohort (KICH+KIRC+KIRP)	**KIPAN**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal clear cell carcinoma	**KIRC**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal papillary cell carcinoma	**KIRP**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Acute Myeloid Leukemia	**LAML**	Y	Y	Y	Y	Y		Y	Y	Y		
Brain Lower Grade Glioma	**LGG**	Y		Y	Y	Y			Y	Y	Y	
Liver hepatocellular carcinoma	**LIHC**	Y	Y	Y	Y	Y			Y	Y		
Lung adenocarcinoma	**LUAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Lung squamous cell carcinoma	**LUSC**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Mesothelioma	**MESO**	Y		Y	Y	Y			Y			
Ovarian serous cystadenocarcinoma	**OV**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Pancreatic adenocarcinoma	**PAAD**	Y		Y	Y	Y			Y	Y		
Pheochromocytoma and Paraganglioma	**PCPG**	Y		Y	Y	Y			Y	Y		
Prostate adenocarcinoma	**PRAD**	Y		Y	Y	Y			Y	Y		
Rectum adenocarcinoma	**READ**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Sarcoma	**SARC**	Y		Y	Y	Y			Y			
Skin Cutaneous Melanoma	**SKCM**	Y		Y	Y	Y			Y	Y		
Stomach adenocarcinoma	**STAD**		Y	Y	Y	Y		Y	Y	Y		
Testicular Germ Cell Tumors	**TGCT**	Y		Y	Y	Y			Y	Y		
Thyroid carcinoma	**THCA**	Y	Y	Y	Y	Y			Y	Y		
Thymoma	**THYM**	Y		Y	Y	Y			Y			
Uterine Corpus Endometrial Carcinoma	**UCEC**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Uterine Carcinosarcoma	**UCS**	Y		Y	Y	Y			Y	Y		
Uveal Melanoma	**UVM**	Y		Y	Y	Y			Y	Y		

http://www.liuzlab.org/TCGA2STAT/CancerDataChecklist.pdf

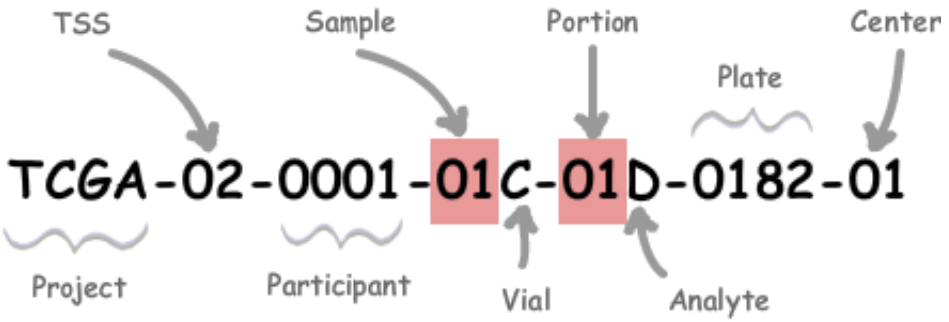
TCGA Clinical data

cvars	Description	Values
yearstobirth	Age (at first diagnosis)	Integer
vitalstatus	Vital status	Binary; 1 - dead, 0 - alive/censored
daystodeath	Number of days to death (overall survival time)	Integer or NA if the vitalstatus is 0
daystolastfollowup	Number of days to the last follow-up (last known survival time)	Integer or NA (sometimes) if the vitalstatus is 1
gender	Gender	Categorical; "female" or "male"
race	Race	Categorical; "asian", "white", "black or african american", "american indian or alaska native", or "native hawaiian or other pacific islander"
ethnicity	Ethnicity	Categorical; "hispanic or latino" or "not hispanic or latino"
pathologicstage	Pathologic stage	Categorical; vary slightly based on cancer types but in general could range from Stage I to IV, and sub-stages such as Stage Ia, Ib, etc.
pathologyTstage	Tumor stage (in TNM staging system) describing the size and location of the tumor	Categorical; vary based on cancer types but in general could be "TX", "T0", "Tis", "T1", "T2", "T3", "T4", and substages like "T2a","T2b" etc.
pathologyNstage	Lymph nodes status (in TNM staging system) describing if the cancer has spread into nearby lymph nodes	Categorical; vary based on cancer types but in general could be "NX", "N0", "N1", N2", "N3", and substages such as "N1a", "N2a", etc
pathologyMstage	Metastasis status (in TNM staging system) describing if the cancer has spread to other parts of the body	Categorical; vary based on cancer types but in general could be "MX", "M0", "M1", and substages "M1a", "M1b", etc.

http://www.liuzlab.org/TCGA2STAT/ClinicalVariables.pdf

TCGA sample identifiers

- Each sample has a unique ID (barcode), like `TCGA-AO-A128` or `TCGA-A1-AOSK-01A`
- Each barcode can and should be parsed



- Can be used to distinguish normal and tumor samples (Sample: Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29)
- Not to be confused with case UUIDs, like `7eea2b6e-771f-44c0-9350-38f45c8dbe87`, which are bound to filenames

PAM50

- Breast cancer can be classified into 4 major intrinsic subtypes: Luminal A, Luminal B, Her2-enriched, Basal
- Subtypes are clinically relevant for drug sensitivity and long-term survival
- Determine tumor subtype by looking at the gene expression of 50 genes

Parker, Joel S., Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, et al. "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes." Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 27, no. 8 (March 10, 2009): 1160–67. https://doi.org/10.1200/JCO.2008.18.1370.

https://xenabrowser.net/datapages/?dataset=TCGA.BRCA.sampleMap/BRCA_clinicalMatrix&host=https://tcga.xenahubs.net

`genefu` R package for PAM50 classification and survival analysis.
https://www.bioconductor.org/packages/release/bioc/html/genefu.html

The Broad Institute Genome Data Analysis Center (GDAC) Firehose

- Standardized, analysis-ready TCGA datasets
  - Aggregated, version-stamped
  - Analysis-ready format / semantics
- Standardized analyses upon them
  - For vetted algorithms: GISTIC, MutSig, CNMF, ...
  - Companioned with biologist-friendly reports

http://gdac.broadinstitute.org/

Firehose data access

- `fbget` - Python application programming interface (API) with >27 functions for Sample-level data, Firehose analyses, Standard data archives, Metadata access
- Unix command-line access, `firehose_get`
- `FirebrowseR` - An R client for broads firehose pipeline, providing TCGA data sets
- `web-TCGA` - a shiny app to access TCGA data from Firebrowse

http://firebrowse.org/

https://confluence.broadinstitute.org/display/GDAC/fbget

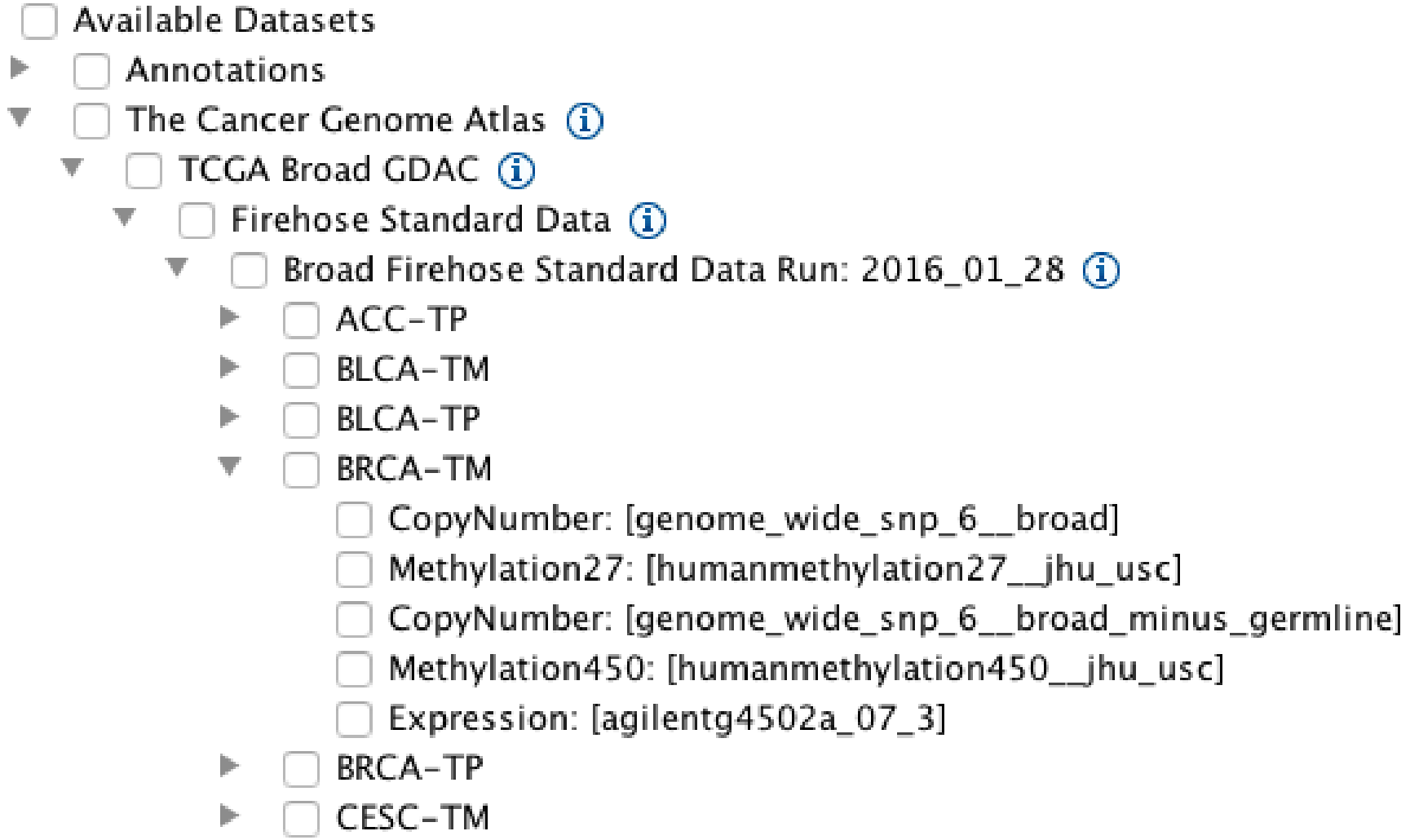
https://confluence.broadinstitute.org/display/GDAC/Download

https://github.com/mariodeng/FirebrowseR

https://github.com/mariodeng/web-TCGA

Firehose data visualization

Firehose data comes pre-loaded in IGV (File/Load from server)



NCI's Genomic Data Commons (GDC)

- Launched on June 6, 2016
- Provides standardized genomic and clinical data from
  - **The Cancer Genome Atlas (TCGA)**
  - **Therapeutically Applicable Research To Generate Effective Treatments (TARGET)** - A comprehensive genomic approach to determine molecular changes that drive childhood cancers. (AML and Neuroblastoma)
  - **Cancer Cell Line Encyclopedia (CCLE)** - Genome-wide information of ~1000 cell lines under baseline condition. Pharmacologic response profiles (IC50) and mutation status analysis.
  - **Stand Up To Cancer (SU2C)** - 50 Breast cancer cell lines. GI50 to 77 therapeutic compounds.
  - **Connectivity Map** - 4 cell lines and 1309 perturbagens at several concentrations. Gene expression change after treatment.

https://ocg.cancer.gov/programs/target.

https://portals.broadinstitute.org/ccle

http://www.standuptocancer.org/

Accessing GDC

- The GDC Application Programming Interface (API)
- `GenomicDataCommons` - GDC access in R

https://docs.gdc.cancer.gov/API/Users_Guide/Getting_Started/#api-endpoints

https://bioconductor.org/packages/release/bioc/html/GenomicDataCommons.html

cBioPortal

- Rich set of tools for visualization, analysis and download of large-scale cancer genomics data sets.
  - Mutations (OncoPrint display)
  - Mutual exclusivity of genetic events (log-odds ratio)
  - Correlations among genetic events (boxplots)
  - Survival (Kaplan-Meier plots)
- The Onco Query Language (OQL) to fine-tune queries

http://www.cbioportal.org/index.do

http://www.cbioportal.org/tutorial.jsp - short tutorials

Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, et al. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the CBioPortal." Science Signaling 6, no. 269 (April 2, 2013): pl1. https://doi.org/10.1126/scisignal.2004088.

cBioPortal data

- REST-based web API
- `CGDS-R` package provides a basic set of functions for querying the Cancer Genomic Data Server (CGDS)
- `MATLAB CGDS Cancer Genomics Toolbox` - data access functionality in the MATLAB environment

http://www.cbioportal.org/web_api.jsp

http://www.cbioportal.org/cgds_r.jsp

https://cran.r-project.org/web/packages/cgdsr/vignettes/cgdsr.pdf

R resources to access TCGA data

- `curatedTCGAData` - Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects
  - MultiAssayExperiment objects integrate multiple assays (e.g. RNA-seq, copy number, mutation, microRNA, protein, and others) with clinical / pathological data.
  - Patient IDs are matched (same number and order) across multiple assays, enabling harmonized subsetting of rows (features) and columns (patients / samples) across the entire experiment.
- `HarmonizedTCGAData` - Processed Harmonized TCGA Data of Five Selected Cancer Types

https://bioconductor.org/packages/release/data/experiment/html/curatedTCGAData.html

MultiAssayExperiment TCGA data, http://tinyurl.com/MAEOurls

https://bioconductor.org/packages/release/data/experiment/html/HarmonizedTCGAData.html

R resources to access TCGA data

- `curatedOvarianData`
  - 30 datasets, $>$ 3K unique samples
  - survival, surgical debulking, histology...
- `curatedCRCData` (colorectal)
  - 34 datasets, ~4K unique samples
  - many annotated for MSS, gender, stage, age, N, M
- `curatedBladderData`
  - 12 datasets, ~1,200 unique samples
  - many annotated for stage, grade, OS

TCGA packages

- `TCGAbiolinks` - an R/Bioconductor package for integrative analysis of TCGA data

Features	Sub-features	TCGAbiolinks	TCGA Assembler	can Envolve	TCGA2stat	Firehose-FirebrowseR	RTCGA Toobox	cBio Portal CGDS-R
		B	R	W	C	CW	B	CW
Availability	Platform							
Query TCGA Cases	Different Versions	x					x	
	Individual TCGA samples (e.g. TCGA-01-0001)	x	x			x		
Download	All TGCA platforms	x						
Data Type Analysis	mRNA	x		x	x	x	x	x
	miRNA	x		x	x	x	x	x
	Copy number	x		x	x	x	x	x
	DNA Methylation	x			x	x	x	x
	Clinical	x		x	x	x	x	x
	Protein			x		x		x
	Mutation	x		x	x	x	x	x
Integrative Analysis	DNA Meth. and Gene Exp.	x				x		
	Clinical and Exp. (dnet)	x				x	x	x
Other	Extensible to other BioC packages	x						

TCGA2STAT

- Well-structured TCGA data access in R

http://www.liuzlab.org/TCGA2STAT/

GDCRNATools

- Downloading, organizing, and integrative analyzing RNA data in the GDC
- Differential gene expression analysis, ceRNAs regulatory network analysis, univariate survival analysis, and functional enrichment analysis.
- Considers ceRNAs - Competing endogenous RNAs, RNA molecules that indirectly regulate other RNA transcripts by competing for the shared miRNAs.

https://github.com/Jialab-UCR/GDCRNATools

Li, Ruidong, Han Qu, Shibo Wang, Julong Wei, Le Zhang, Renyuan Ma, Jianming Lu, Jianguo Zhu, Wei-De Zhong, and Zhenyu Jia. "GDCRNATools: An R/Bioconductor Package for Integrative Analysis of LncRNA, MiRNA, and MRNA Data in GDC," December 11, 2017. https://doi.org/10.1101/229799.

https://github.com/Jialab-UCR/GDCRNATools

Xena Functional Genomics Explorer

- Former UCSC Cancer Genomics Browser. Now UCSC Xena
- Includes TCGA, Cancer Cell Line Encyclopedia, the Stand Up To Cancer (SU2C) Breast Cancer data, custom datasets
- A tool to visually explore and analyze cancer genomics data and its associated clinical information.
- Gene- and genome-centric view
- Survival analysis on user-defined subgroups

https://xenabrowser.net/, https://xenabrowser.net/datapages/, http://xena.ucsc.edu/getting-started/

Cline, Melissa S., Brian Craft, Teresa Swatloski, Mary Goldman, Singer Ma, David Haussler, and Jingchun Zhu. "Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser." Scientific Reports 3 (October 2, 2013): 2652. https://doi.org/10.1038/srep02652.

Gitools

- A framework for analysis and visualization of multidimensional genomic data using interactive heatmaps
- User-provided and precompiled datasets: TCGA, IntOGen
- Analyses: Enrichment, Group Comparison, Mutual exclusion and co-occurrence test, Correlations, Overlaps, Combination of p-values

http://www.gitools.org/

TCGA analysis on the cloud

- Goal - simplify centralized access to TCGA data and provide easy analysis
- Three centers were awarded to develop cloud access
  - Institute for Systems Biology Cancer Genomics Cloud (ISB-CGC)
  - Broad Institute FireCloud
  - Seven Bridges Cancer Genomics Cloud

http://cgc.systemsbiology.net/

https://software.broadinstitute.org/firecloud/

http://www.cancergenomicscloud.org/

Other resources for cancer genomics

- `IntOgen` - catalog of cancer driver mutations,
- `Regulome Explorer` - exploratory analysis of integrated TCGA data
- `Oncomine research edition` - coexpression, differential analysis of cancer datasets, including TCGA
- `CPTAC` - Clinical Proteomics Tumor Analysis Consortium

https://www.intogen.org/search

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. "IntOGen-Mutations Identifies Cancer Drivers across Tumor Types." Nature Methods 10, no. 11 (September 15, 2013): 1081–82. https://doi.org/10.1038/nmeth.2642.

http://explorer.cancerregulome.org

https://www.oncomine.org/resource/login.html

International Cancer Genome Consortium

- International effort
- A comprehensive catalog of somatic changes in the major cancers
- *10,000 cancer genomes*
- Similar to other large-scale genome projects, the ICGC has a Data Coordination Center (DCC)

http://icgc.org/

ICGC data portal http://dcc.icgc.org/