



#AACRMolEpi



The Prostate Cancer Clinical Trials Consortium

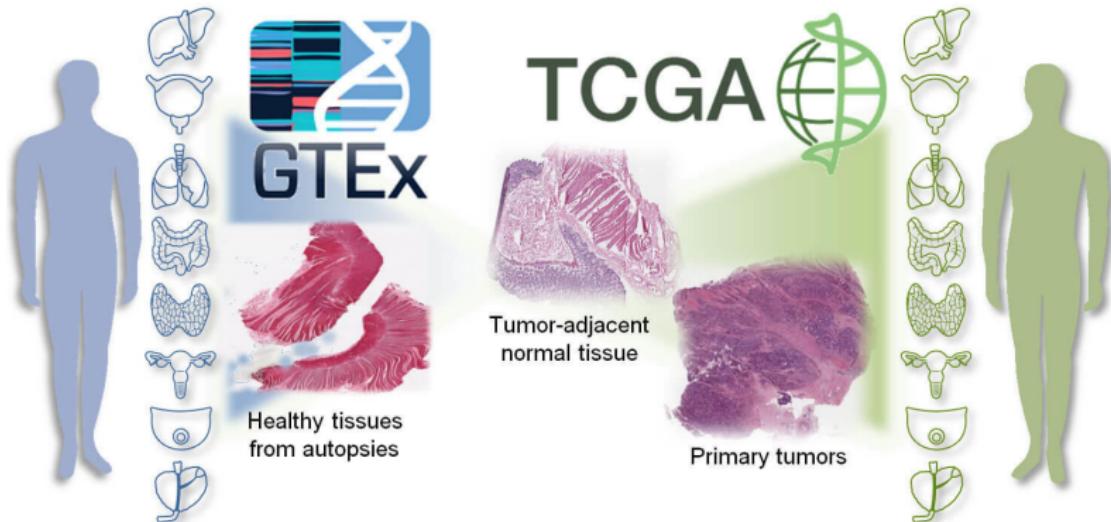
Publicly available genomics and ‘omics resources: Beyond DNA

AACR IME Workshop
July 27, 2021

Travis Gerke, ScD
Director of Data Science, PCCTC
gerket@mskcc.org
 @travisgerke

'Omics resources we're going to explore

1. The Cancer Genome Atlas
2. The Genotype-Tissue Expression project (GTEx)
3. Gene Expression Omnibus (GEO)
4. Molecular Signatures Database (MSigDB)



The Cancer Genome Atlas (TCGA)

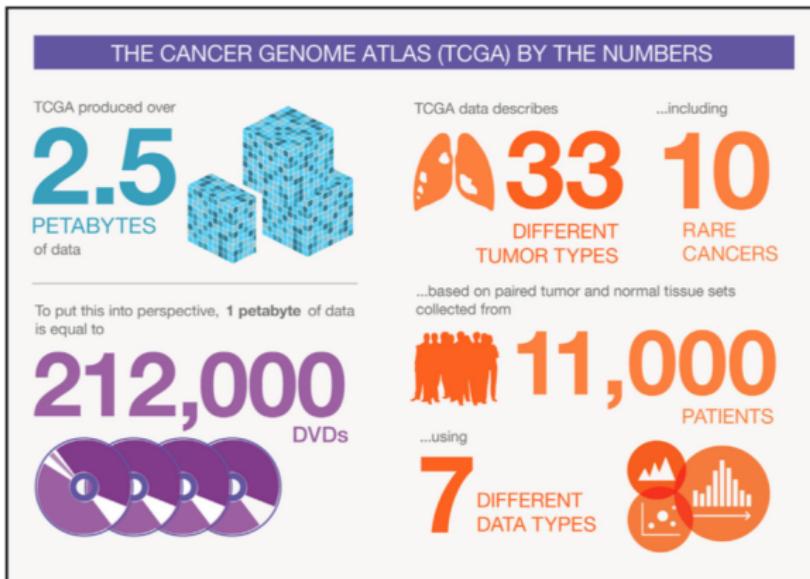
Purpose: characterize the molecular landscape of primary cancers

Joint project of NCI and NHGRI (> \$400 million investment)

Samples collected at 20 institutions across US/Canada 2006–2013

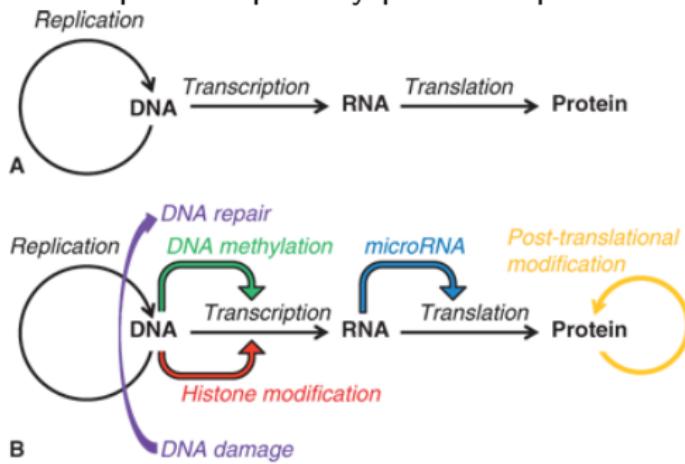
Data generation and analysis continued through 2017

Data made rapidly available to the public for research use



The 7 TCGA data types (in tumor + matched normal tissue)

1. Whole exome sequence: protein-coding segments of DNA
2. mRNA sequence: quantify transcript/gene expression
3. microRNA sequence: small RNAs that regulate gene expression
4. DNA copy number profile: how often sections of the genome repeat
5. DNA methylation profile: chemical additions to DNA that regulate gene expression
6. Whole genome sequence: coding and non-coding DNA sequence
7. RPPA expression profile: quantify protein expression



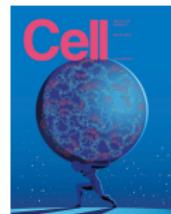
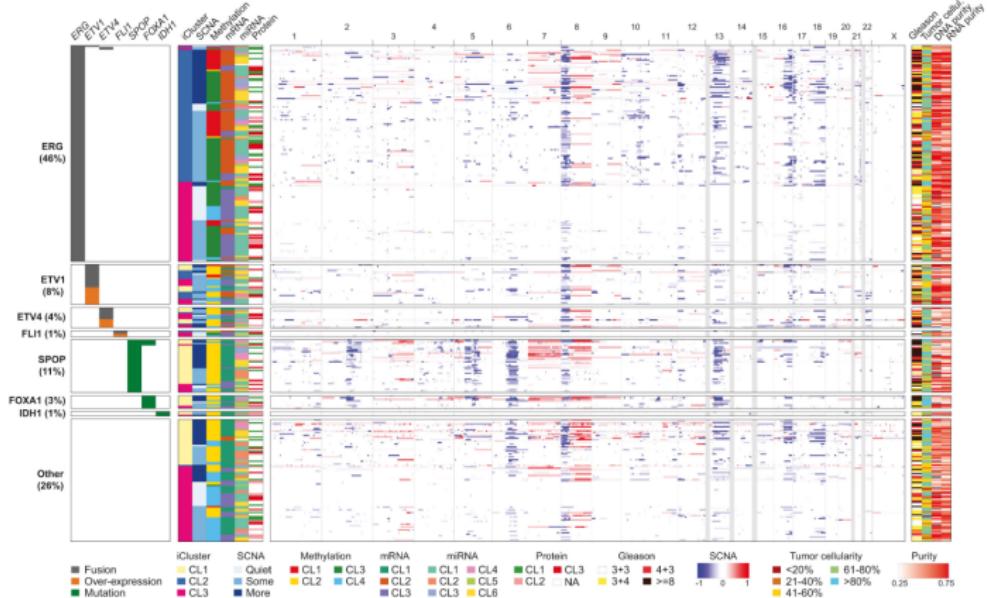
Bioinformatic insights gained through TCGA can't be overstated

Over the past 10 years, disease-specific working groups have molecularly characterized each cancer [[publication list](#)]

General theme of each paper: create taxonomy of molecular subtypes
Important details about collection/QC also appear in each

Advice: *read the publication(s) for your cancer site of interest!*

With conclusion of TCGA, Pan-Cancer series published in 2018



The comprehensiveness of TCGA can feel overwhelming

By comparison, scope/impact of your grant ideas may seem insignificant

But wait!

Notice that one of the 7 data types was not “clinical”

“

Obtaining comprehensive clinical annotation was neither a primary program objective nor a practical possibility, given the worldwide scope and severe time constraints for sample accrual goals determined at the time of TCGA program initiation and funding.

”

– Liu et al. Cell (2018)

Wide-open area: leverage molecular understanding from TCGA to guide clinical/epidemiologic investigations

Advice: in grants/manuscripts, point out how TCGA may provide preliminary evidence but is unable to fully answer your translational question (or use TCGA as validation data)

Getting to work: interrogating TCGA data

cBioPortal [link] is an extremely well-written (i.e. fast) application for querying/visualizing genomic data (“Google for cancer genomics”) It is most commonly used to analyze TCGA data in a web browser The browser hosts >250 cancer studies (not just TCGA) For hackers: you can put your own data in cBio [link]



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

Datasets

The portal currently contains data from cancer genomics studies. The table below lists the number of available samples per cancer study and data type.

Columns ▾

Name	Reference	All	Sequenced	CNA	RNA-Seq	Tumor mRNA (microarray)	Tumor miRNA	Methylation (HM27)	RPPA	Complete
Acinar Cell Carcinoma of the Pancreas (Johns Hopkins, J Pathol 2014)	Jia et al. J Pathol 2014	23	23	0	0	0	0	0	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)	Zhang et al. Nat Genet 2016	73	73	0	0	0	0	0	0	0
Acute Myeloid Leukemia (TCGA, NEJM 2013)	TCGA, NEJM 2013	200	200	191	173	0	0	194	0	162
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)		200	200	191	0	0	0	0	0	165
Acute Myeloid Leukemia (TCGA, Provisional)		200	197	191	173	0	0	194	0	163
Adenoid Cystic Carcinoma (FMI, Am J Surg Pathl. 2014)	Ross et al. Am J Surg Pathl 2014	28	15	28	0	0	0	0	0	0
Adenoid Cystic Carcinoma (MDA, Clin Cancer Res 2015)	Mitauri et al. Clin Cancer Res 2015	102	102	0	0	0	0	0	0	0
Adenoid Cystic Carcinoma (MSKCC, Nat Genet 2013)	Ho et al. Nat Genet 2013	60	60	60	0	0	0	0	0	0
Adenoid Cystic Carcinoma (Sanger/MDA, JCI 2013)	Stephens et al. J Clin Invest 2013	24	24	0	0	0	0	0	0	0
Adenoid Cystic Carcinoma of the Breast (MSKCC, J Pathol. 2015)	Marlefotto et al. J Pathol 2015	12	12	12	0	0	0	0	0	0
Adrenocortical Carcinoma (TCGA, PanCancer Atlas)		92	91	89	0	0	0	0	0	76
Adrenocortical Carcinoma (TCGA, Provisional)		92	90	90	79	0	0	0	46	75
Ampullary Carcinoma (Baylor College of Medicine, Cell Reports 2016)	Gingras et al. Cell Rep 2016	160	160	0	0	0	0	0	0	0
Bladder Cancer (MSKCC, Eur Urol 2014)	Kim et al. Eur Urol 2015	109	109	109	0	0	0	0	0	0
Bladder Cancer (MSKCC, JCO 2013)	Iyer et al. JCO 2013	97	97	97	0	58	0	24	0	58
Bladder Cancer (TCGA, Cell 2017)		413	412	408	408	0	0	0	344	404
Bladder Cancer, Plasmacytoid Variant (MSKCC, Nat Genet 2016)	Al-Ahmadie et al. Nat Genet 2016	34	34	33	0	0	0	0	0	0
Bladder Urothelial Carcinoma (BGI, Nat Genet 2013)	Guo et al. Nat Genet 2012	99	99	0	0	0	0	0	0	0
Bladder Urothelial Carcinoma (Dana Farber & MSKCC, Cancer Discov 2014)	Van Allen et al. Cancer Discov 2014	50	50	0	0	0	0	0	0	0
Bladder Urothelial Carcinoma (TCGA, Nature 2014)	TCGA, Nature 2014	131	130	128	129	0	0	0	120	125
Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)		411	410	408	0	0	0	0	0	402
Bladder Urothelial Carcinoma (TCGA, Provisional)		413	238	408	408	0	0	0	344	126
Brain Lower Grade Glioma (TCGA, PanCancer Atlas)		514	512	511	0	0	0	0	0	507
Brain Lower Grade Glioma (TCGA, Provisional)		532	286	513	530	27	0	0	435	283
Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)	Pereira et al. Nat Commun 2016	2509	2369	2173	0	0	0	0	0	0

TCGA through cBioPortal: landing page

Select studies, molecular profiles, and *genes* to query

Important: read the TCGA papers to understand difference between provisional/published/pan-cancer versions (often QC-driven)

QUERY DOWNLOAD DATA

Select Studies:

1 studies selected (563 samples) Deselect all View summary

Search...

PanCancer Studies 2 Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012)
MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017)
NCI-60 Cell Lines (NCI, Cancer Res. 2012)
SUMMIT (Nature, 2018)

Cell lines 2 1020 samples 1 2 3

Adrenal Gland 2 10945 samples 1 2 3

Ovary/Fallopian Tube 2 60 samples 1 2 3

Ampulla of Vater 1 141 samples 1 2 3

Ovarian Epithelial Tumor
— SEROUS OVARIAN CANCER 563 samples 1 2 3

Biliary Tract 6 Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)
Ovarian Serous Cystadenocarcinoma (TCGA, PanCancer Atlas)
Ovarian Serous Cystadenocarcinoma (TCGA, Provisional)

Bladder/Urinary Tract 10 585 samples 1 2 3

Bone 2 606 samples 1 2 3

— SMALL CELL CARCINOMA OF THE OVARY 12 samples 1 2 3

Small Cell Carcinoma of the Ovary (MSKCC, Nat Genet 2014)

Bowel 8

Pancreas 1 23 samples 1 2 3

Acinar Cell Carcinoma of the Pancreas

Breast 12 Acinar Cell Carcinomas of the Pancreas (Johns Hopkins, J Pathol 2014)

CNS/Brain 15

Select Genomic Profiles:

Mutations
 Putative copy-number alterations (GISTIC)
 mRNA Expression. Select one of the profiles below:
 mRNA expression Z-scores (all genes)
 microRNA expression Z-scores
 mRNA/miRNA expression Z-scores (all genes)

Select Patient/Case Set:

Tumors with sequencing and CNA data (316) X

To build your own case set, try our enhanced Study View.

Enter Genes:

User-defined List

Advanced: Onco Query Language (OQL)

BRCA1 BRCA2

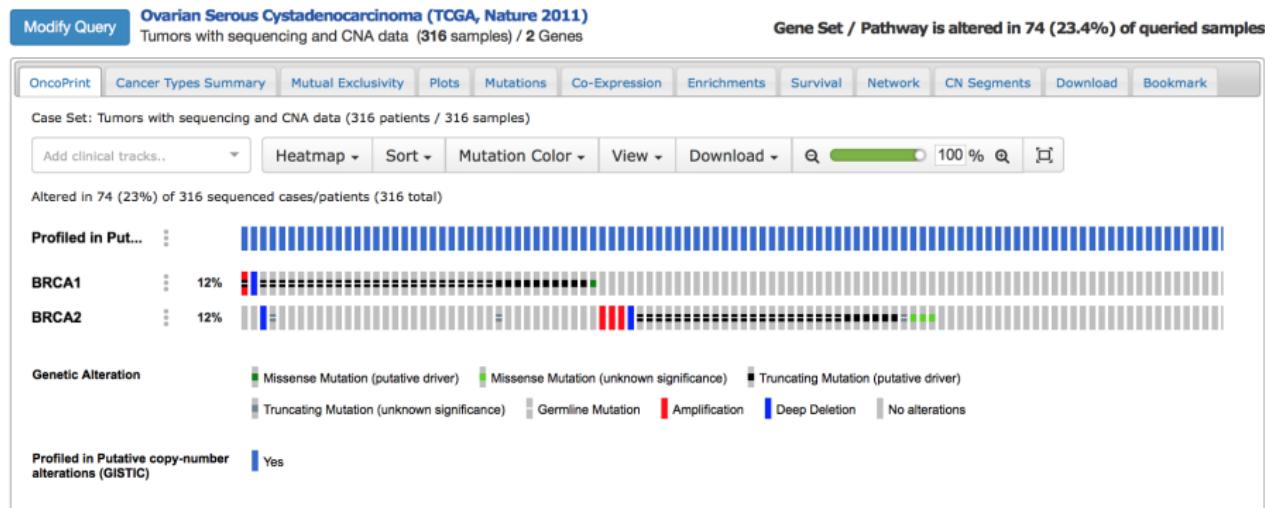
All gene symbols are valid.

TCGA through cBioPortal: basic query

Each tab is rich with info and visualizations

Note: the genomic profiles you select on the landing page determine how many patients have “altered” genes

E.g. I picked mutations + CNVs, but could also have selected mRNA



TCGA through cBioPortal: mutual exclusivity

Among all pairwise combinations of your genes, tells you how likely alterations are to occur together/exclusively

Modify Query Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)
Tumors with sequencing and CNA data (316 samples) / 2 Genes Gene Set / Pathway is altered in 74 (23.4%) of queried samples

OncoPrint Cancer Types Summary Mutual Exclusivity Plots Mutations Co-Expression Enrichments Survival Network CN Segments Download Bookmark

The query contains **1** gene pair with mutually exclusive alterations (none significant), and **no** gene pair with co-occurrence alterations.

Mutual exclusivity Co-occurrence Significant only

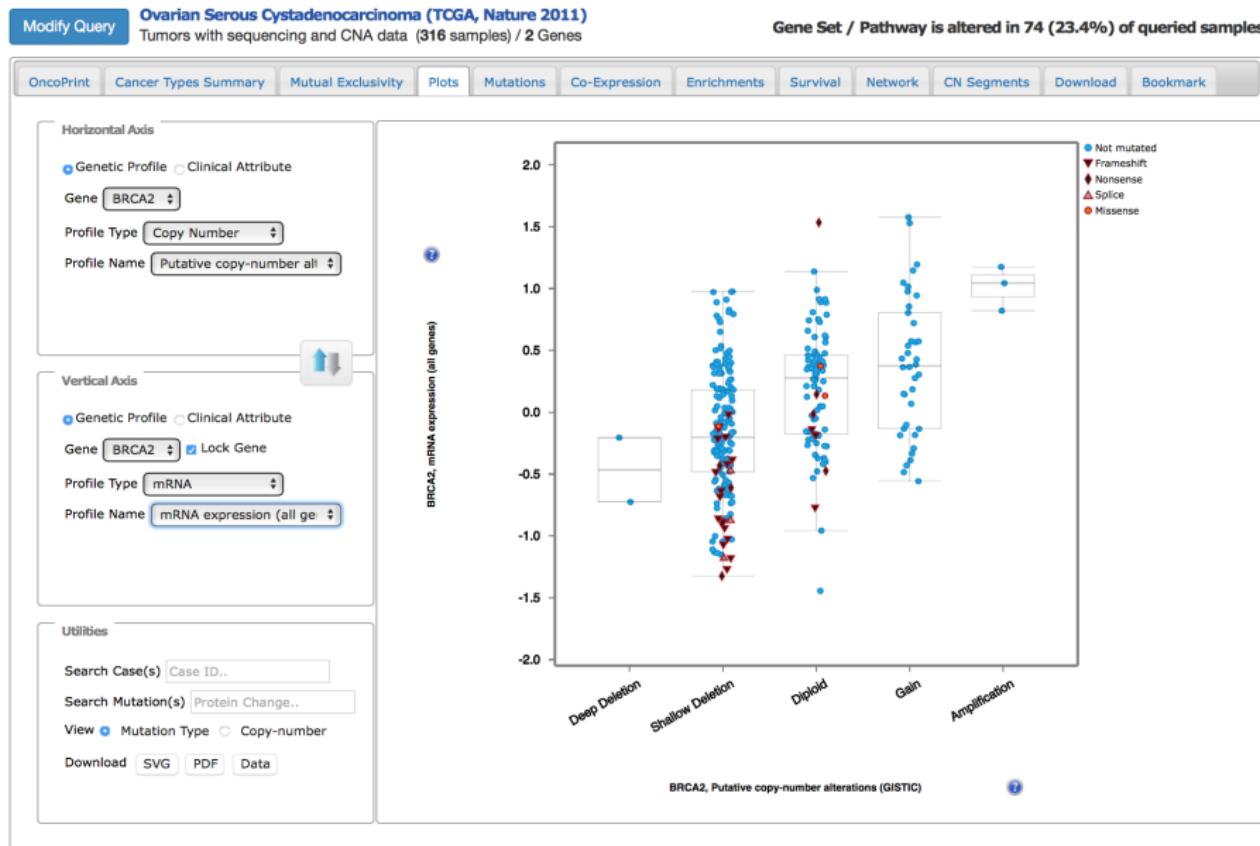
Columns ▾

Gene A	Gene B	Neither	A Not B	B Not A	Both	Log Odds Ratio	p-Value	Adjusted p-Value ▲	Tendency
BRCA1	BRCA2	242	35	36	3	-0.551	0.277	0.277	Mutual exclusivity

Showing 1-1 of 1

TCGA through cBioPortal: plots

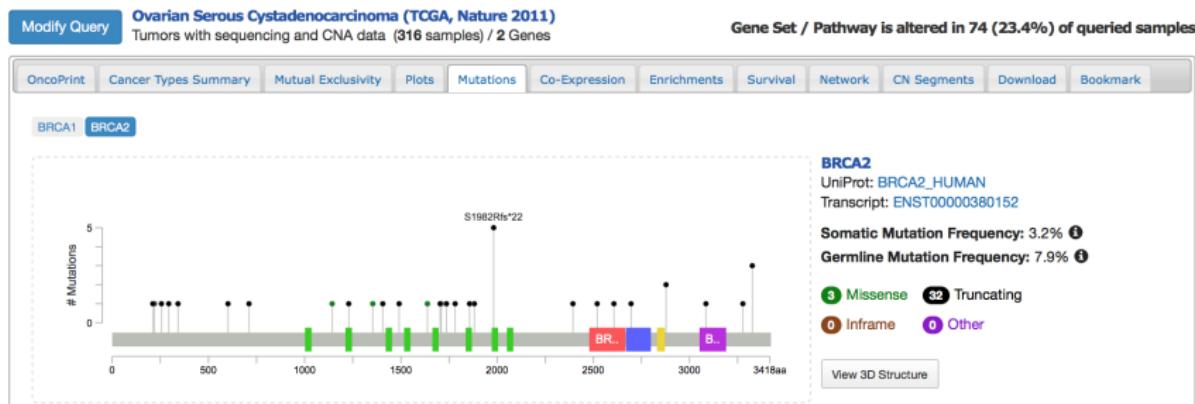
Very useful: works with genomic and clinical features



TCGA through cBioPortal: mutations

Tells you the frequency and mutation details (somatic vs germline, nonsense, etc)

Note: synonymous mutations are not reported



35 Mutations (page 1 of 2)

Columns ▾

Q

Sample ID	Protein Change	Annotation ▾	Mutation Type	Copy #	COSMIC	# Mut in Sample
TCGA-24-2288-01	V220Ifs*4 Germline	●	FS del	ShallowDel		85
TCGA-13-0886-01	S1982Rfs*22 Germline	●	FS del	ShallowDel	1	49
TCGA-13-1498-01	S1982Rfs*22 Germline	●	FS del	Diploid	1	87
TCGA-24-2280-01	S1982Rfs*22 Germline	●	FS del	ShallowDel	1	113
TCGA-13-1499-01	S1982Rfs*22 Germline	●	FS del	ShallowDel	1	51
TCGA-59-2351-01	S1982Rfs*22 Germline	●	FS del	ShallowDel	1	70
TCGA-57-1584-01	N1784Hfs*2 Germline	●	FS del	ShallowDel	2	26
TCGA-04-1367-01	E294* Germline	●	Nonsense	ShallowDel		75
TCGA-04-1231-01	G741A	●	Nonsense	ShallowDel	1	75

TCGA through cBioPortal: co-expression

Shows co-expression with your selected genes

Modify Query Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)
Tumors with sequencing and CNA data (316 samples) / 2 Genes Gene Set / Pathway is altered in 74 (23.4%) of queried samples

OncoPrint Cancer Types Summary Mutual Exclusivity Plots Mutations Co-Expression Enrichments Survival Network CN Segments Download Bookmark

Data Set: mRNA expression (all genes) (316 samples)

BRCA1 BRCA2

Show Any Correlation Enter gene or cytoband.

Correlated Gene	Cytoband	Pearson's Correlation	Spearman's Correlation
TOP2A	17q21.2	0.48	0.53
SPAG5	17q11.2	0.46	0.52
CIT	12q24.23	0.45	0.50
NCAPH	2q11.2	0.44	0.48
TPX2	20q11.21	0.41	0.47
CENPF	1q41	0.43	0.47
C17orf53	17q21.31	0.44	0.46
NBR2	17q21.31	0.57	0.45
MYBL2	20q13.12	0.39	0.45
BUB1B	15q15.1	0.43	0.45
RACGAP1	12q13.12	0.41	0.44
TROAP	12q13.12	0.43	0.44
HJURP	2q37.1	0.38	0.44
KIF14	1q32.1	0.40	0.44
UBE2C	20q13.12	0.40	0.44
PSMC3IP	17q21.2	0.41	0.43
RNF72	12q24.22	0.41	0.43
CENPA	2p23.3	0.38	0.43
CCNB1	5q13.2	0.39	0.43
TUBG1	17q21.2	0.37	0.42
EXO1	1q43	0.38	0.42
KIFC1	6p21.32	0.38	0.42
ESPL1	12q13.13	0.41	0.42
KNTC1	12q24.31	0.37	0.42
ATAD5	17q11.2	0.41	0.42

BRCA1 Show Mutations SVG PDF

mRNA expression (all genes): BRCA1 vs. TOP2A

TOP2A (Cytoband: 17q21.2)

BRCA1 (Cytoband: 17q21.31)

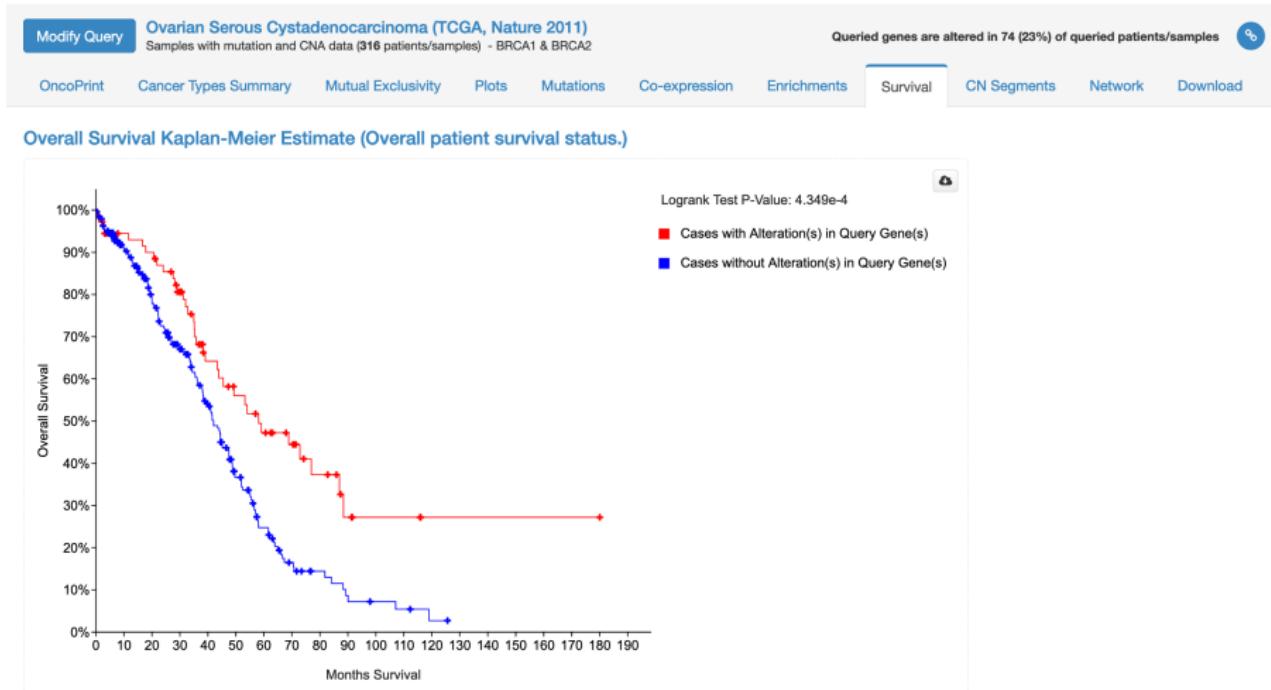
BRCA1 mutated
TOP2A mutated
Both mutated
Neither mutated

Pearson: 0.48
Spearman: 0.53

Showing 1-25 of 18505 Show more

TCGA through cBioPortal: survival

Compares survival among gene altered and non-altered patients



TCGA through cBioPortal: study summaries

Access by clicking the study name wherever you see it

Very useful high-level study summaries

Click Clinical Data tab → DATA to download all clinical data*

Merge with Download tab from gene queries for custom analyses

Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011) ▲

Whole exome sequencing (316 samples with matched normals), mRNA expression, miRNA expression, promoter methylation, and DNA copy number in 489 high-grade serous ovarian adenocarcinomas (HGS-OvCa). The Cancer Genome Atlas (TCGA) Serous Ovarian Cancer project. 489 cases.

Raw data via the TCGA Data Portal, PubMed

Summary Clinical Data Heatmaps CN Segments Selected: 489 patients | 489 samples Custom Selection + Add Chart Groups ▾

Quick Filters: 316 samples with mutation data 489 samples with CNA data

Mutation Count vs Fraction of Genome Altered

Mutated Genes (316 profiled samples)

Gene	# Mut	#	Freq ▾
TP53	306	303	95.9%
BRCA1	37	37	11.7%
BRCA2	35	34	10.8%
LRP1B	14	13	4.1%
NF1	13	12	3.8%
FAT1	11	11	3.5%
LRRK2	10	10	3.2%
CDK12	9	9	2.8%
RNF213	9	9	2.8%
CHD4	8	8	2.5%
PRKDC	8	8	2.5%

CNA Genes (489 profiled samples)

Gene	Cytoband	CNA	#	Freq ▾
MYC	8q24.21	AMP	154	31.5%
NDRG1	8q24.22	AMP	133	27.2%
RECL4	8q24.3	AMP	129	26.4%
MECOM	3q26.2	AMP	121	24.7%
TERC	3q26.2	AMP	120	24.5%
AGO2	8q24.3	AMP	115	23.5%
PRKCI	3q26.2	AMP	108	22.1%
CNE1	19q12	AMP	106	21.7%
EXT1	8q24.11	AMP	98	20.0%
TBL1XR1	3q26.32	AMP	93	19.0%
PIK3CA	3q26.32	AMP	88	18.0%

Mutation Count

Fraction Genome Altered

ACGH Data

Complete Data

Disease Free (Months)

Disease Free Status

MRNA Data

Neoplasm Histologic Grade

Overall Survival Status

Search... Search...

TCGA through cBioPortal: patient views

Access by clicking a patient ID wherever you see one

Includes detailed summary of mutations + clinical data

Patient: TCGA-31-1956, Ovarian Cancer (High-Grade Serous Ovarian Cancer), **LIVING** (2 months), **Recurred/Progressed**
Samples: TCGA-31-1956-01, N

Ovarian Serous Cystadenocarcinoma (TCGA)

Summary Clinical Data Pathology Report Heatmap

Specimen Type:
A: RIGHT OVARY AND UTERUS
B: LEFT OVARY
C: Omentum
D: ANTERIOR ABDOMINAL WALL NODE

TCGA-31-1956

Clinical Details:
Ovarian Debulking.

Macroscopic Description:

A) **RIGHT OVARY AND UTERUS:** The ovary weighs 53.3 grams and measures 6.5 x 5.5 x 4.5 cm. The ovarian cyst wall is ruptured. The cut surface shows a multicellular cyst with solid areas. The solid area shows tan coloured and brownish areas. The uterus corpus measures 6 cm longitudinally, 3 cm transversally and 3.2 cm anteroposteriorly. The fallopian tubes are not included.
BLOCKS: A1 to A8 – 1 pc in each, rep sections from right ovarian tumour
A9 and A10 – 1 pc in each, lower uterine segment
A11 to A14 – 1 pc in each, endomyometrium with rep sections [REDACTED]

B) **LEFT OVARY:** The ovary measures 2.5 x 1.5 x 0.8 cm. The fallopian tube measures 5.2 cm and the maximum diameter measures 0.6 cm.
BLOCKS: B1 and B2 – 1 pc in each, left ovary bisected
B3 – 3 pcs, fallopian tube with rep sections [REDACTED]

C) **OMENTUM:** A piece of omentum measuring 28 x 8 x 1.2 cm. There are two nodular areas each measuring 6.5 x 4.5 x 1.5 cm and 7.5 x 4 x 1.3 cm. The cut surface of the nodules shows tan coloured homogenous.
BLOCKS: C1 to C7 – 1 pc in each, rep sections from nodular areas [REDACTED]

Page 1 / 2 | - | +

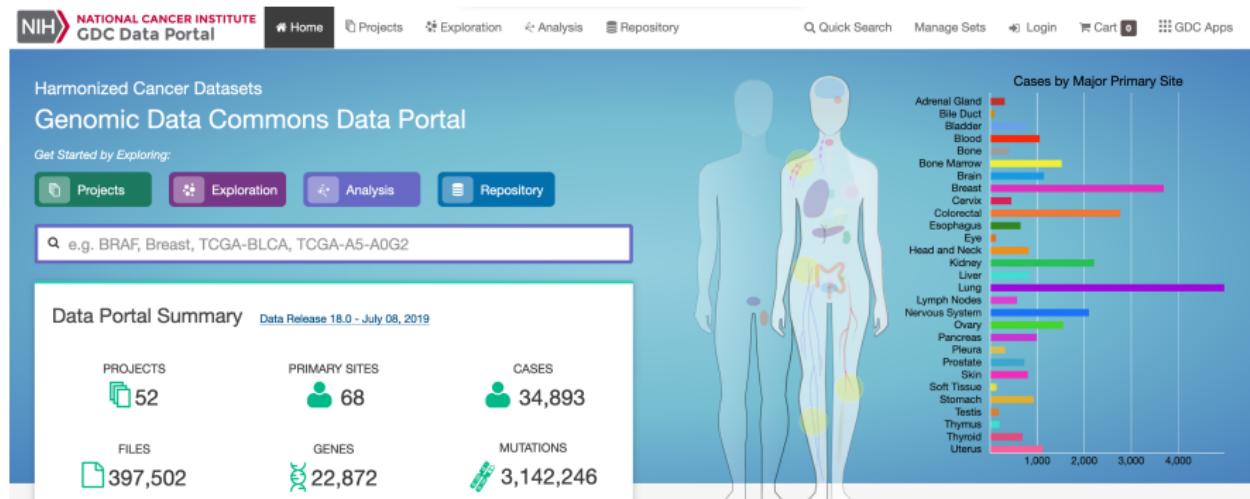
TCGA through GDC

Sometimes, you just need all of the data (e.g. transcriptome-wide studies)

I don't know what the gene # max is for cBio, but it can get slow

Also, analyses/UI within cBio are designed for small # of genes

Genomic Data Commons (GDC) is the official data portal [[link](#)]



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



Data Transfer Tool



API



Data Submission Portal



Documentation



Legacy Archive

TCGA through GDC: querying

In simple cases (< 10,000 files), GDC works like a shopping cart

The screenshot shows the GDC search interface with the following details:

- Project ID:** TCGA-PRAD
- Number of Files:** 14,287
- Number of Cases:** 900
- Total Size:** 17.56 TB

The interface includes several filters and summary charts:

- Data Category:** Simple Nucleotide Variation (4,038), Transcriptome Profiling (2,755), Biopsies (2,168), New Sequencing Data (2,133), Copy Number Variation (2,076).
- Data Type:** Aligned Reads (2,152), Annotated Somatic Mutation (2,013), Raw Simple Somatic Mutation (2,012), Gene Expression Quantification (1,493), Slide Image (1,172).
- Experimental Strategy:** WXS (5,081), RNA-Seq (2,905), Genotyping Array (2,076), miRNA-Seq (1,653).

The main table displays 900 cases across various projects and primary sites, with columns for Case ID, Project, Primary Site, Gender, and file counts (Seq, Snp, Cnv, Meth, Clinical, Bio) per category.

Case ID	Project	Primary Site	Gender	Files	Seq	Snp	Cnv	Meth	Clinical	Bio	Annotations	Slides	
TCGA-EU-A4EB	TCGA-PRAD	Prostate	Male	52	5	5	16	5	1	8	14	2	3 (2)
TCGA-EU-A4EE	TCGA-PRAD	Prostate	Male	51	5	5	18	5	1	7	14	2	3 (2)
TCGA-EU-A4EH	TCGA-PRAD	Prostate	Male	51	5	5	16	5	1	7	14	2	3 (2)
TCGA-HC-7741	TCGA-PRAD	Prostate	Male	22	0	0	2	0	0	2	13	1	3 (1)
TCGA-HG-7850	TCGA-PRAD	Prostate	Male	52	5	5	18	5	1	8	14	1	3 (2)
TCGA-O9-6386	TCGA-PRAD	Prostate	Male	57	5	5	18	5	2	8	16	1	3 (2)
TCGA-HC-A4B8	TCGA-PRAD	Prostate	Male	59	5	5	16	5	1	7	13	1	3 (1)
TCGA-EU-7123	TCGA-PRAD	Prostate	Male	84	7	10	18	8	2	8	18	1	3 (2)
TCGA-EU-A6033	TCGA-PRAD	Prostate	Male	51	5	5	16	5	1	7	14	1	3 (2)
TCGA-CH-5790	TCGA-PRAD	Prostate	Male	52	5	5	18	5	1	8	14	1	3 (2)
TCGA-XK-AAK1	TCGA-PRAD	Prostate	Male	51	5	5	16	5	1	8	13	1	3 (1)
TCGA-EU-7312	TCGA-PRAD	Prostate	Male	51	5	5	18	5	1	8	13	1	3 (1)
TCGA-CH-5784	TCGA-PRAD	Prostate	Male	54	5	5	16	5	2	8	16	1	3 (2)
TCGA-XK-AAK3	TCGA-PRAD	Prostate	Male	51	5	5	16	5	1	8	13	1	3 (1)

For larger downloads, use the manifest file and Data Transfer Tool [link]

Additional steps needed for Controlled-Access Data (raw sequence data, SNPs, exon array, VCFs, [link])

Submit an application (next slide) and get Authentication Token using eRA Commons credentials

Can do entirely in R with GenomicDataCommons package [link]

Controlled TCGA data access

Don't let this extra step deter you!

Go to dbGaP Authorized Access page [[link](#)]

Follow the prompts, select the data set as below, write a brief RUS

In my experience, approval \leq 2 weeks

db GaP project and archive

Logged in as [Travis Gerke](#) | [Log out](#)

Browse/Search Authorized Access Help

Beacon Data Browser My Projects My Requests Downloads My Profile

Project Request

#13369: In silico functional analyses of low-grade glioma and glioblastoma
SO: Margaret Fonner

+ OMB control number: 0925-0670 Expiration date: 03/31/2019

[Project Details](#) [Research Project](#) [Collaborators](#) [IT Director](#) [Choose Datasets](#) [Confirm Datasets](#) [Review DUC](#) [Review DUL](#) [Review Applications](#) [Feedback](#)

Please use this form to submit a Project Request to the NIH for the first time or if you are asked to make changes to your original submission. If you are submitting a final report at the end of your approved access period and are not seeking renewal, please go to [close out project](#).

Principal Investigator's (PI) Name: Travis Gerke

Institutional Signing Official (ISO): Margaret Fonner

Institutional Affiliation: H. LEE MOFFITT CANCER CTR & RES INST

Project ID: 13369

Project Name: In silico functional analyses of low-grade glioma and glioblastoma

Initial Request Date:

Date of Last Renewal:

Research Use Statement for "In silico functional analyses of low-grade glioma and glioblastoma" [hide..](#)

GWAS studies have identified several genetic loci that increase risk of low-grade glioma or glioblastoma, but the mechanistic role of these genetic variations is largely unknown. In parallel, cell line and animal models have elucidated the functional role of certain variants, but their relevance in human tissue remains to be validated.

In the proposed study, we aim to perform QTL analyses that link genomic changes (genotype array and WXS data) to other molecular profiles (RNA-seq, miRNA-Seq, methylation, as available) and tumor/clinical phenotypes (tumor subtype, cancer aggressiveness, as available). To conduct the analyses, we will leverage the low-grade glioma and glioblastoma samples available from The Cancer Genome Atlas (TCGA). We aim to use results towards peer-reviewed publications and as supportive data for grant applications.

Datasets for research project "In silico functional analyses of low-grade glioma and glioblastoma" [hide..](#)

DAR #	Study, Consent	Status	Expiration	Application
S1960-1	TCGA - The Cancer Genome Atlas (phs000178.v9.p8) General Research Use (phs000178.v9.p8.c1), TCGA	✓Approved GRANTED	2017-02-10	View

[Return to My Projects](#) [Continue](#)

A final note about TCGA clinical data

Be aware of which version/source you use!

Pan-Cancer group released a version with recommendations

Liu Cell 2018

Often Pan-Cancer \neq Provisional \neq Published \neq cBio \neq ...

E.g. we previously explored a list of such differences [link]

The screenshot shows a GitHub repository page for 'TCGA Pancancer Clinical Data'. At the top, there's a navigation bar with links for GitHub, Inc. (US), the repository URL, and search functions. Below the header, a commit history is displayed:

Author	Commit Message	Time
jhcreck	add cbio	Latest commit 8a95a11 3 days ago
[data]	add cbio	3 days ago
[README.md]	add cbio	3 days ago

Below the commit history, the repository content is shown. A file named 'README.md' is expanded, displaying the following text:

TCGA Pancancer Clinical Data

This repo contains the combined clinical data with follow up and outcome data for the TCGA PanCancer Atlas in a single text file and RData file. All data in its original format can be found at <https://gdc.cancer.gov/about-data/publications/pancanatlas>. All original files had been previously downloaded in June of 2018.

The PanCancer Atlas

The original [flagship](#) paper (Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer by Hoadley et al) represented efforts to provide "comprehensive integrative molecular analyses of the complete set of tumors in TCGA". This paper was accompanied by a paper from [Liu et al](#) (An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics) who attempted to create a standardized dataset for the clinical data across the PanCancer Atlas called the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR). This was necessary as the original TCGA project consisted of two parts 1) the pilot study which focused on GBM, OV, and LUSC and 2) the full project which encompassed 33 cancer types. Due to the relatively short time frame of establishing TCGA (2006-2015) clinical data is often limited and follow-up times are short. Furthermore, the data fields available differ by cancer type as each type had a Disease Working Group who decided what data to collect.

Thus, the PanCancer Atlas provides two data resources for clinical annotations:

`clinical_PANCAN_patient_with_followup.tsv` which is available under Additional Resources/Supplemental Data and is not associated with a specific publication and details on it's creation are scarce and `TCGA-CDR-SupplementalTables1.xlsx`

The Genotype-Tissue Expression project (GTEx)

Data resource to study link between genetic variation and gene expression across multiple human tissues [[link](#)]

Rapid autopsy program > 700 donors \Rightarrow > 11,000 samples

Not cancer patients

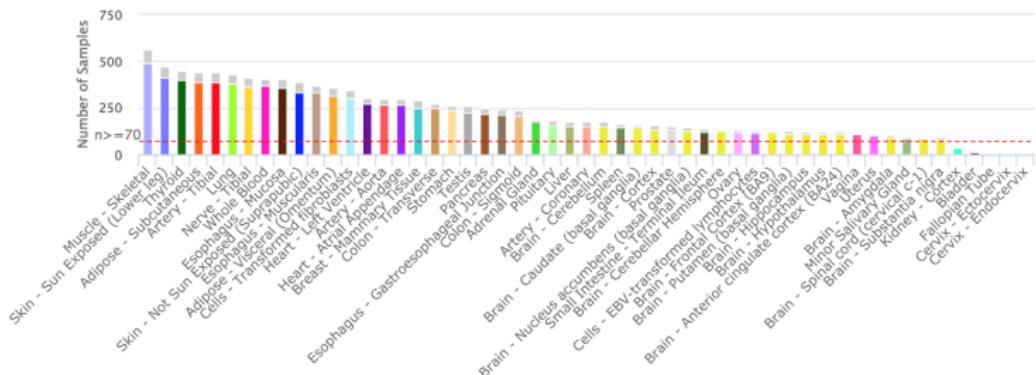
Can be useful to inform what genes are expressed in which tissues in a “normal” state

V7 Sample Counts by Tissues

Sort tissues by: Count

 Download

Vi



GTEx access

Like TCGA, can apply through dbGaP [[link](#)]

But don't use a hammer to swat a fly!

The GTEx web browser [[link](#)] handles most queries easily

Here's an example for BRCA1: [[link](#)]

Where it may fall short: trans-eQTLs

- GTEX Analysis V7
- Additional Datasets
- Biobank Inventory
- GTEX Analysis V6p
- GTEX Analysis V6**
- GTEX Analysis V4
- GTEX Analysis Pilot V3

Please use the Chrome or Firefox browsers to download GTEx files. There is a known issue that prevents Internet Explorer from properly downloading GTEx files.

-GTEX Analysis V7 (dbGaP Accession phs000424.v7.p2)

The GTEX Analysis V7 release is the most current release of the GTEX Portal.

Protected Data and Raw Data

Due to the nature of our donor consent agreement, raw data and attributes which might be used to identify the donors are not publicly available on the GTEX Portal.
You may apply for access to the data through [dbGaP](#).

Data available include:

- BAM files for RNA-Seq, Whole Exome Seq, and Whole Genome Seq
- Genotype Calls (.vcf) for OMNI SNP Arrays, WES, and WGS
- OMNI SNP Array Intensity files (.ida and .gtc)
- Affymetrix Expression Array Intensity files (.cel)
- Allele Specific Expression (ASE) tables
- All expression matrices from the Portal, including samples that did not pass the Analysis Freeze QC
- Sample Attributes
- Subject Phenotypes

On dbGaP, the VCF used for V6p eQTL analyses is in the archive (under 'Genotype Files') phg000520.v2.GTEx_MidPoint_imputation.genotype-calls-vcf.c1.GRU.tar

Gene Expression Omnibus (GEO) [[link](#)]

Published studies of high-throughput data are very often publicly available

Typically, funding mechanism/journal required for reproducibility

When GEO is most useful: studies with long-term follow-up or epidemiologic data (that resources like TCGA/GTEX lack)

Using this data requires programming (R/similar), so not covered here

If interested, check out GEOquery for R [[link](#)]

Alternatively, I often just download the data and import from scratch

A typical query: [[link](#)]

The screenshot shows the NCBI GEO homepage. At the top, there's a navigation bar with links for 'Resources', 'How To', 'GEO Home', 'Documentation', 'Query & Browse', and 'Email GEO'. On the right, there's a 'Sign in to NCBI' link. The main title 'Gene Expression Omnibus' is prominently displayed with its logo. Below the title, a brief description states: 'GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.' To the right of this text is a search bar with 'Keyword or GEO Accession' and a 'Search' button. The page is divided into three main sections: 'Getting Started' (with links to Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, and How to Download Data), 'Tools' (with links to Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, Studies with Genome Data Viewer Tracks, Programmatic Access, and FTP Site), and 'Browse Content' (with links to Repository Browser, DataSets: 4348, Series: 115443, Platforms: 19877, and Samples: 3133399).

Molecular Signatures Database (MSigDB) [link]

Rich compendium of annotated gene set collections

Once registered, can use to look at overlaps with your list of genes,
or generate lists to query in tools like cBioPortal

The screenshot shows the MSigDB homepage. At the top, there's a navigation bar with links for "GSEA Home", "Downloads", "Molecular Signatures Database" (which is highlighted in blue), "Documentation", and "Contact". On the left, a sidebar titled "MSigDB Home" contains links for "About Collections", "Browse Gene Sets", "Search Gene Sets", "Investigate Gene Sets", "View Gene Families", and "Help". The main content area features a large logo for "MSigDB Molecular Signatures Database". Below the logo, there's a section titled "Overview" which describes the database as a collection of annotated gene sets for use with GSEA software. It includes a list of features: "Search for gene sets by keyword.", "Browse gene sets by name or collection.", "Examine a gene set and its annotations. See, for example, the GO_NOTCH_SIGNALING_PATHWAY gene set page.", "Download gene sets.", "Investigate gene sets: Compute overlaps between your gene set and gene sets in MSigDB.", "Categorize members of a gene set by gene families.", and "View the expression profile of a gene set in a provided public expression compendia.".

On the right, there's a "Collections" section with a table showing 8 major collections:

Collection Type	Description
H	hallmark gene sets: coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	positional gene sets: for each human chromosome and cytogenetic band.
C2	curated gene sets: from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	motif gene sets: based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	computational gene sets: defined by mining large collections of cancer-oriented microarray data.
C5	GO gene sets: consist of genes annotated by the same GO terms.
C6	oncogenic gene sets: defined directly from microarray gene expression data from cancer gene perturbations.
C7	immunologic gene sets: defined directly from microarray gene expression data from immunologic studies.

At the bottom, there's a "Current Version" section with the note: "MSigDB database v6.1 updated October 2017. Release notes: GSEA (MSigDB) version 6.1 released January 2018". The footer contains the URL "Software.broadinstitute.org/gsea/msigdb/collections.jsp".

An MSigDB overlap query

These are TCGA's top 50 associated genes with BRCA1 in ovarian

The screenshot shows the GSEA interface. The top navigation bar includes links for 'GSEA Home', 'Downloads', 'Molecular Signatures Database', 'Documentation', and 'Contact'. A sidebar on the left lists 'MSigDB Home', 'About Collections', 'Browse Gene Sets', 'Search Gene Sets', 'Investigate Gene Sets' (which is selected and highlighted in blue), 'View Gene Families', and 'Help'. The main content area is titled 'Investigate Gene Sets' and contains a section for 'Gain further insight into the biology behind a gene set by using the following tools:' with three bullet points: 'compute overlaps with other gene sets in MSigDB (more...)', 'display the gene set expression profile based on a selected compendium of expression data (more...)', and 'categorize members of the gene set by gene families (more...)'. Below this is a 'Gene Identifiers' section listing genes: TOP2A, SPAG5, CIT, NCAP4, TPX2, CENPF, C17orf53, NBR2, MYBL2, BUB1B, RACGAP1, TRIO, HJURP, KIF14, UBE2C, PSMC3IP, RNF12, CENPF, CCNB1, TUBG1, EXO1, KIFC1, ESR1, KNTC1, ATAD5, TIMELESS, E2F7. To the right are sections for 'Compute Overlaps' (with checkboxes for various gene set types like Hallmark, Curated, Chemical & Genetic Perturbations, etc.) and 'Compendia expression profiles' (with checkboxes for Human tissue compendium (Novartis) and NCI-60 cell lines (National Cancer Institute)). At the bottom, there are buttons for 'show gene families' and 'show [top 10] geneses'.

Compute Overlaps for Selected Genes

Converted 50 submitted identifiers into 46 entrez genes. [click here for details.](#)

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
H	3	50	46	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates

Exercise

You've got some genes from the earlier exercise in 15q25 (I hope!)

How can TCGA (via cBioPortal) extend your story around the locus?

What are TCGA's limitations for this story?

Are you able to identify any datasets in GEO that may help you?

You don't need to download/use it right now, just identify one

Bonus: Identify a relevant gene pathway in MSigDB

Bonus: What can you learn from GTEx about this locus/related genes?

What would your next study be?

Give two specific aims