#AACRMolEpi

The Prostate Cancer Clinical Trials Consortium

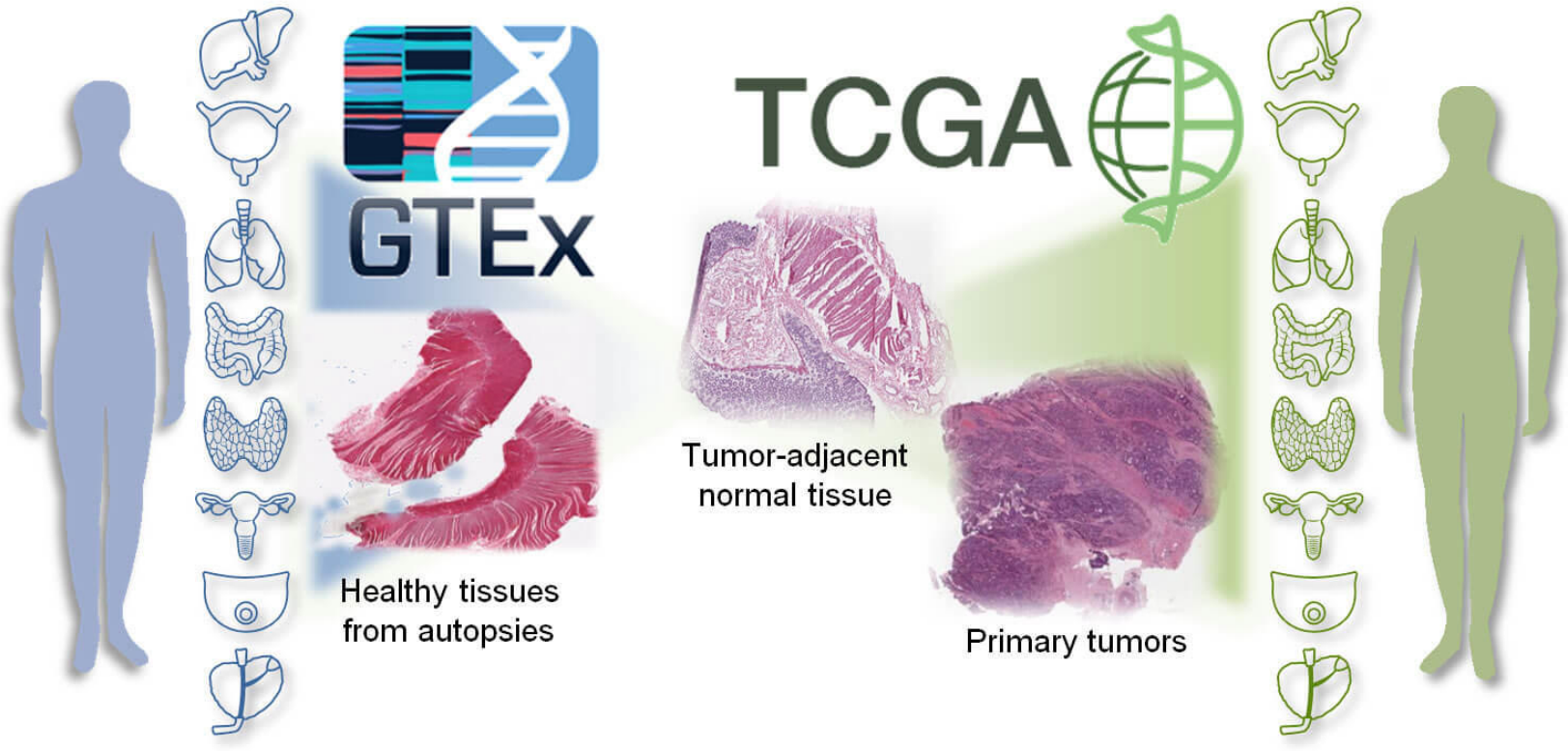
Publicly available genomics and 'omics resources: Beyond DNA

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

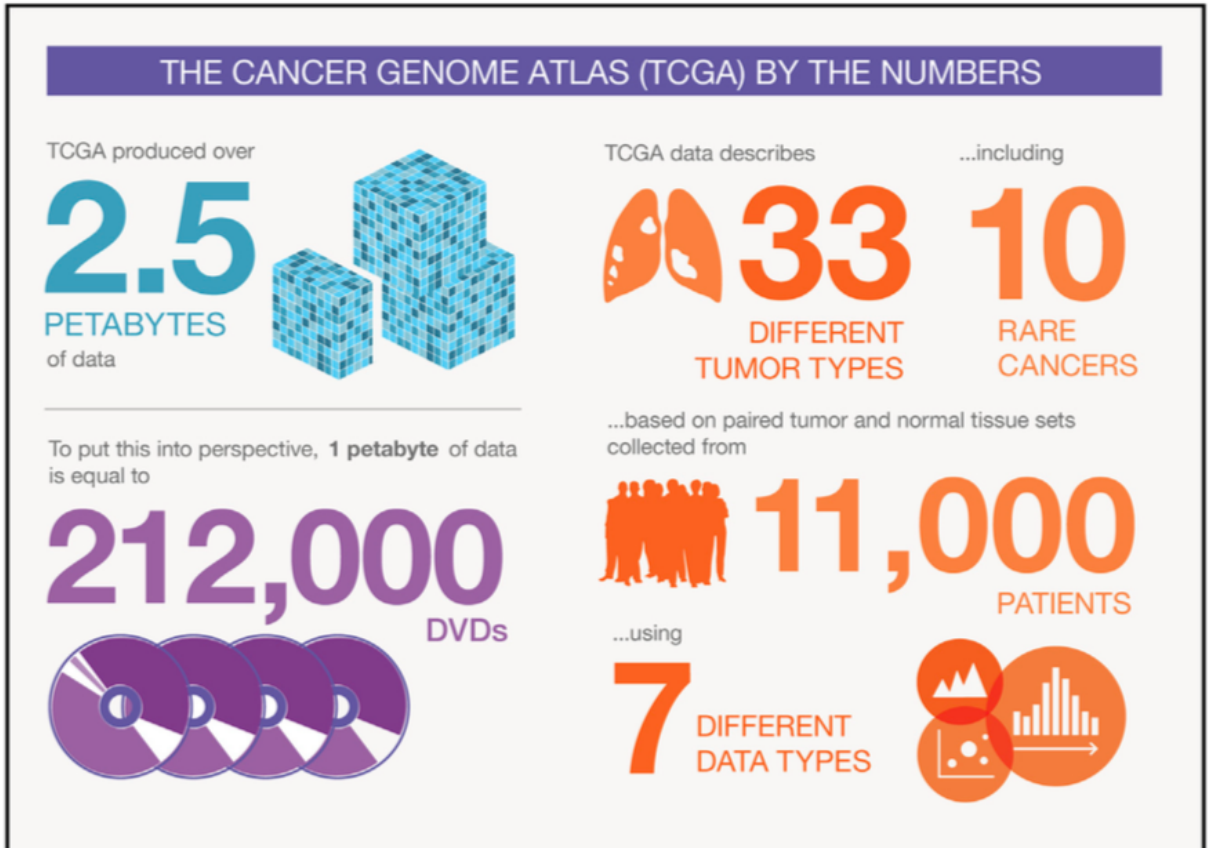
1. The Cancer Genome Atlas
2. The Genotype-Tissue Expression project (GTEx)
3. Gene Expression Omnibus (GEO)
4. Molecular Signatures Database (MSigDB)



The Cancer Genome Atlas (TCGA)

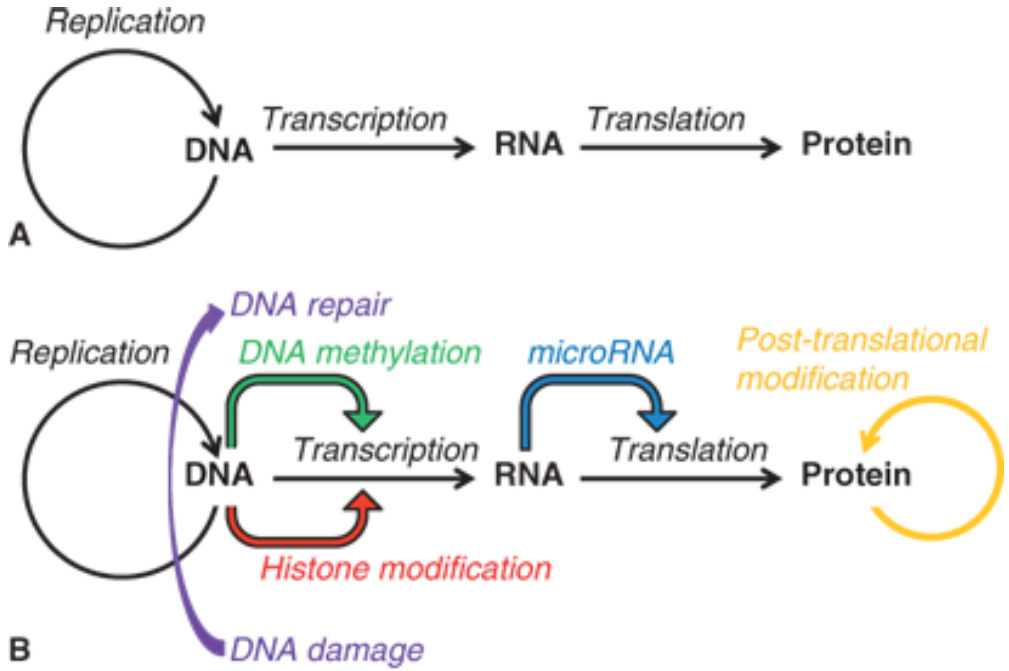
Purpose: characterize the molecular landscape of primary cancers

- 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



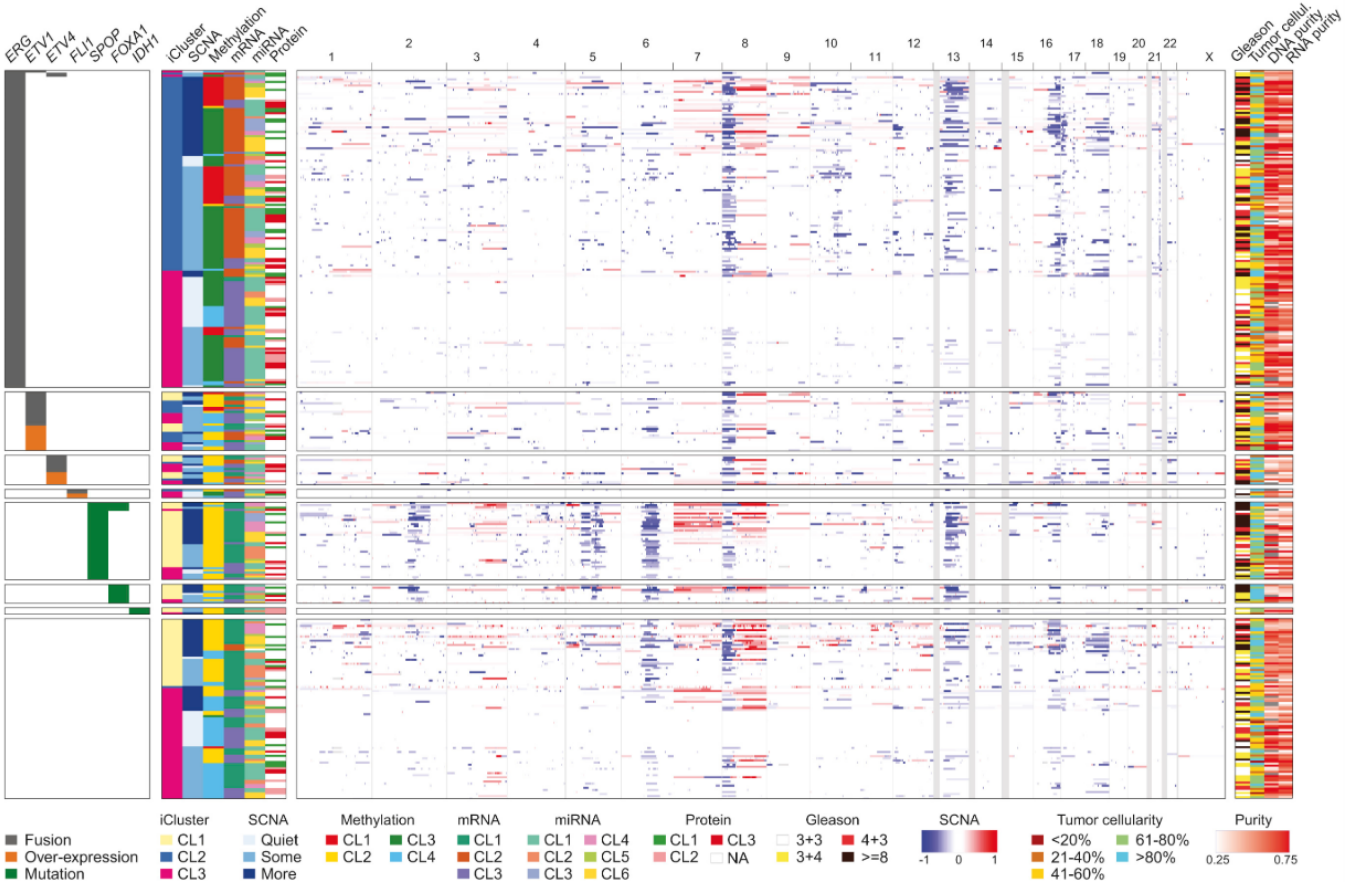
Source: Howard M. Reisner: Pathology: A Modern Case Study
www.accessmedicine.com

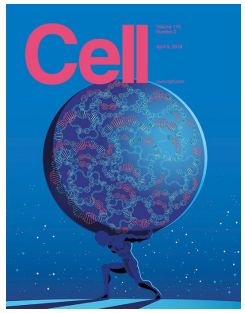
Bioinformatic insights gained through TCGA can't be overstated

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)

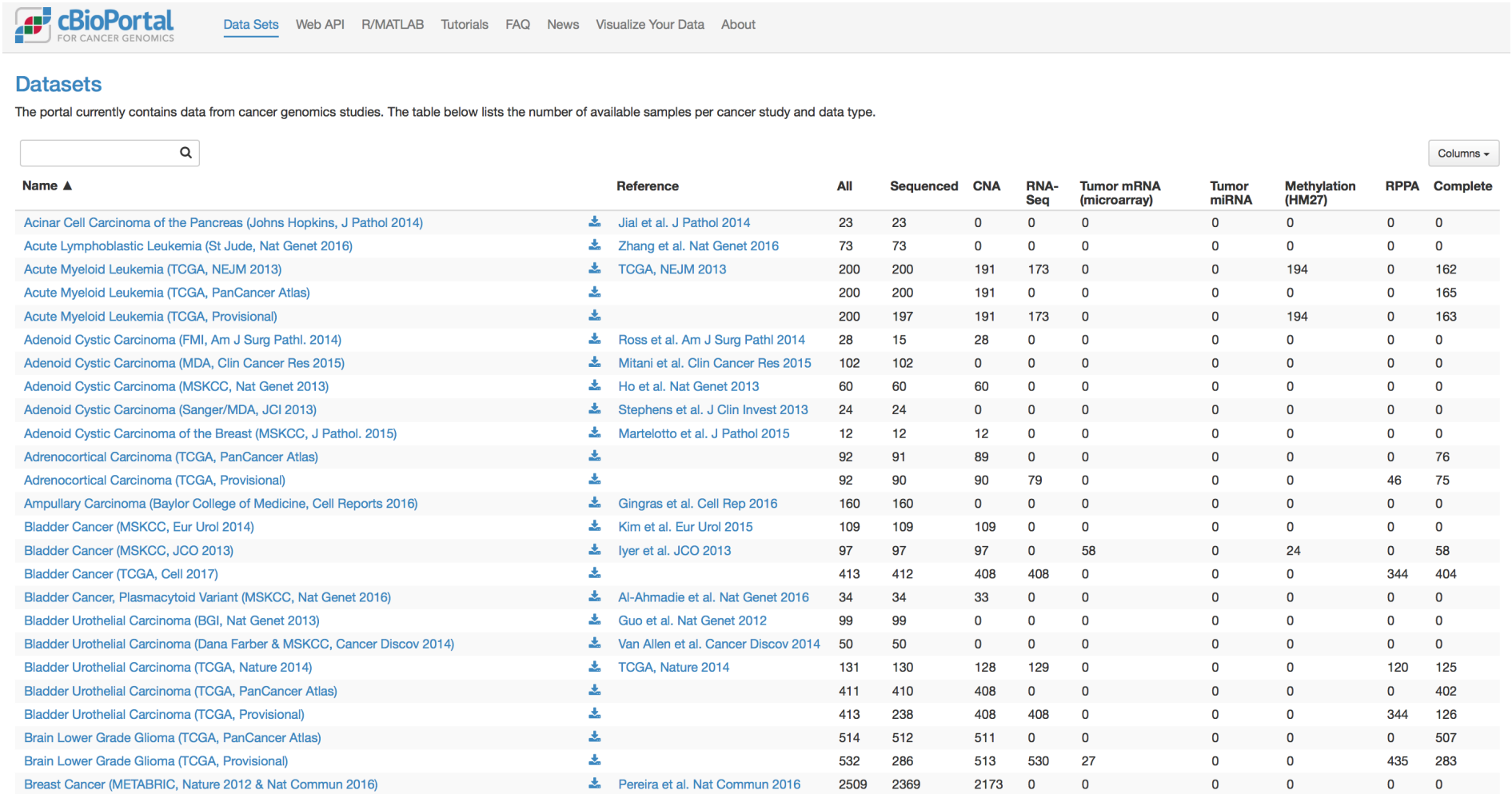
Wide-open area: leverage molecular understanding from TCGA to guide clinical/epidemiologic investigations

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]

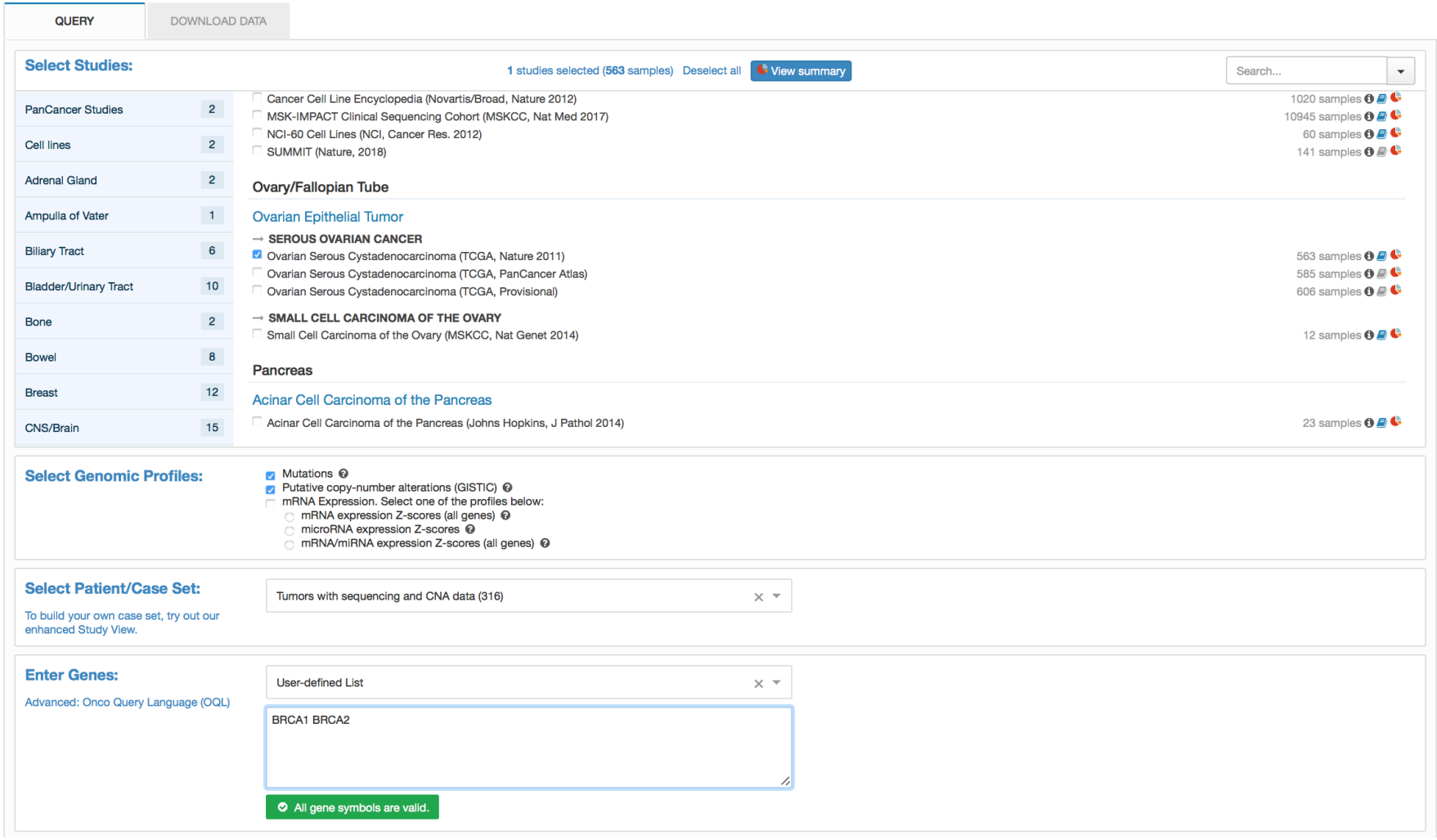


Name ▲	Reference	All	Sequenced	CNA	RNA-Seq	Tumor mRNA (microarray)	Tumor miRNA	Methylation (HM27)	RPPA	Complete
Acinar Cell Carcinoma of the Pancreas (Johns Hopkins, J Pathol 2014)	Jial et al. J Pathol 2014	23	23	0	0	0	0	0	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)	Zhang et al. Nat Genet 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)	Mitani 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)	Martelotto 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)

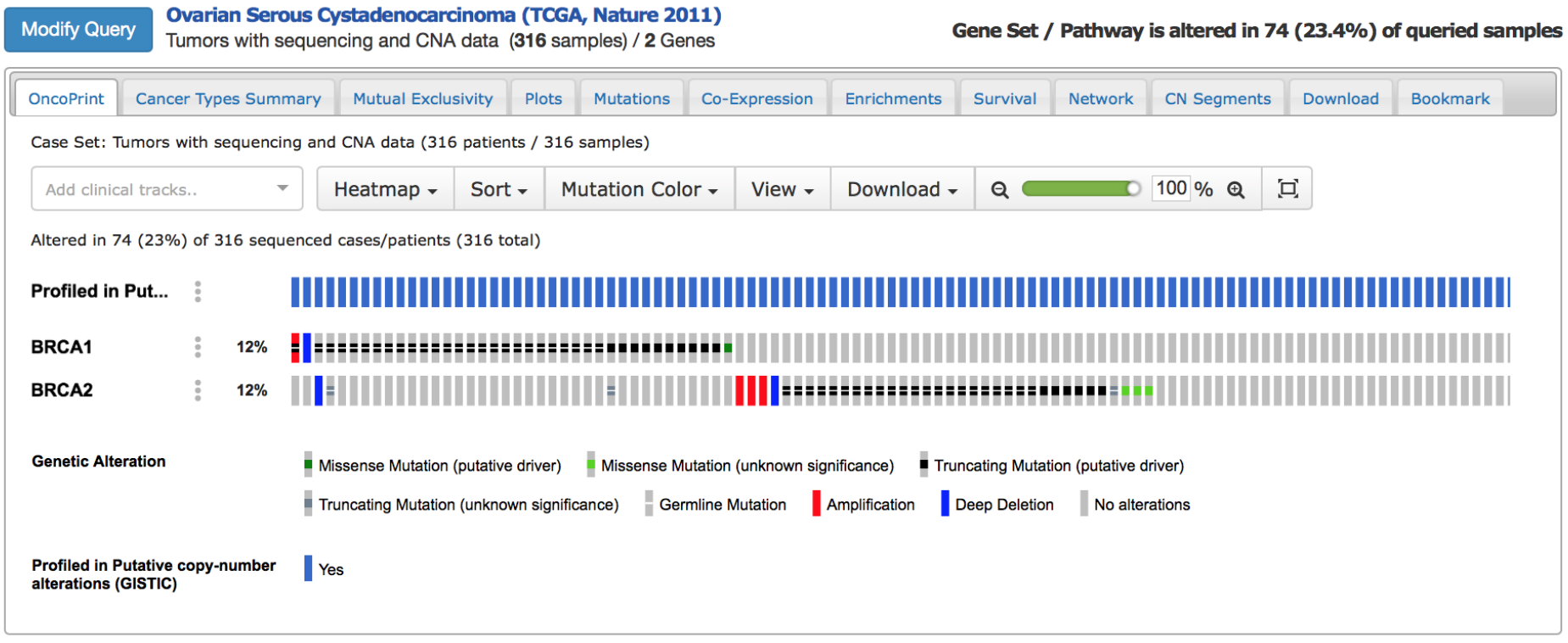


TCGA through cBioPortal: basic query

Each tab is rich with info and visualizations

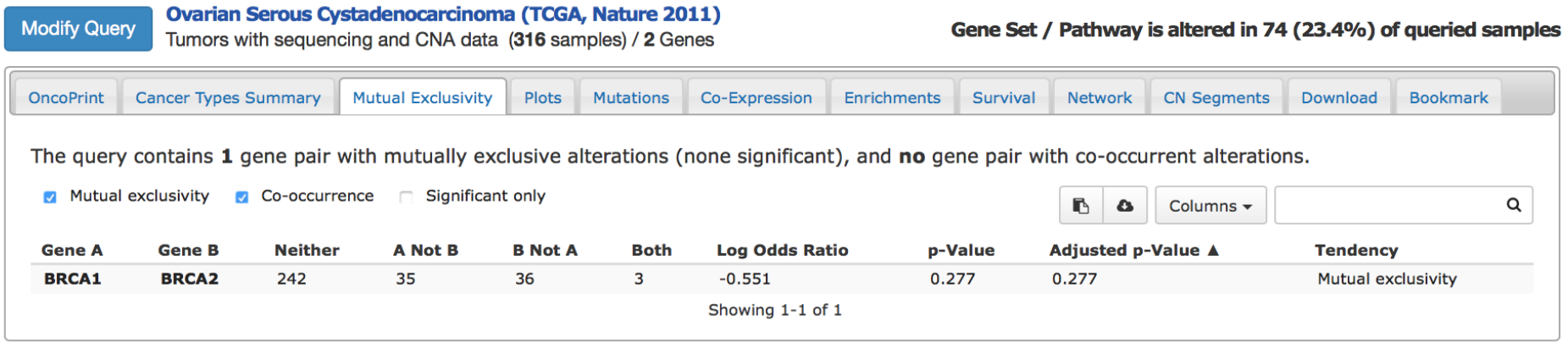
Note: the genomic profiles you select on the landing page determine how many patients have "altered" genes

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



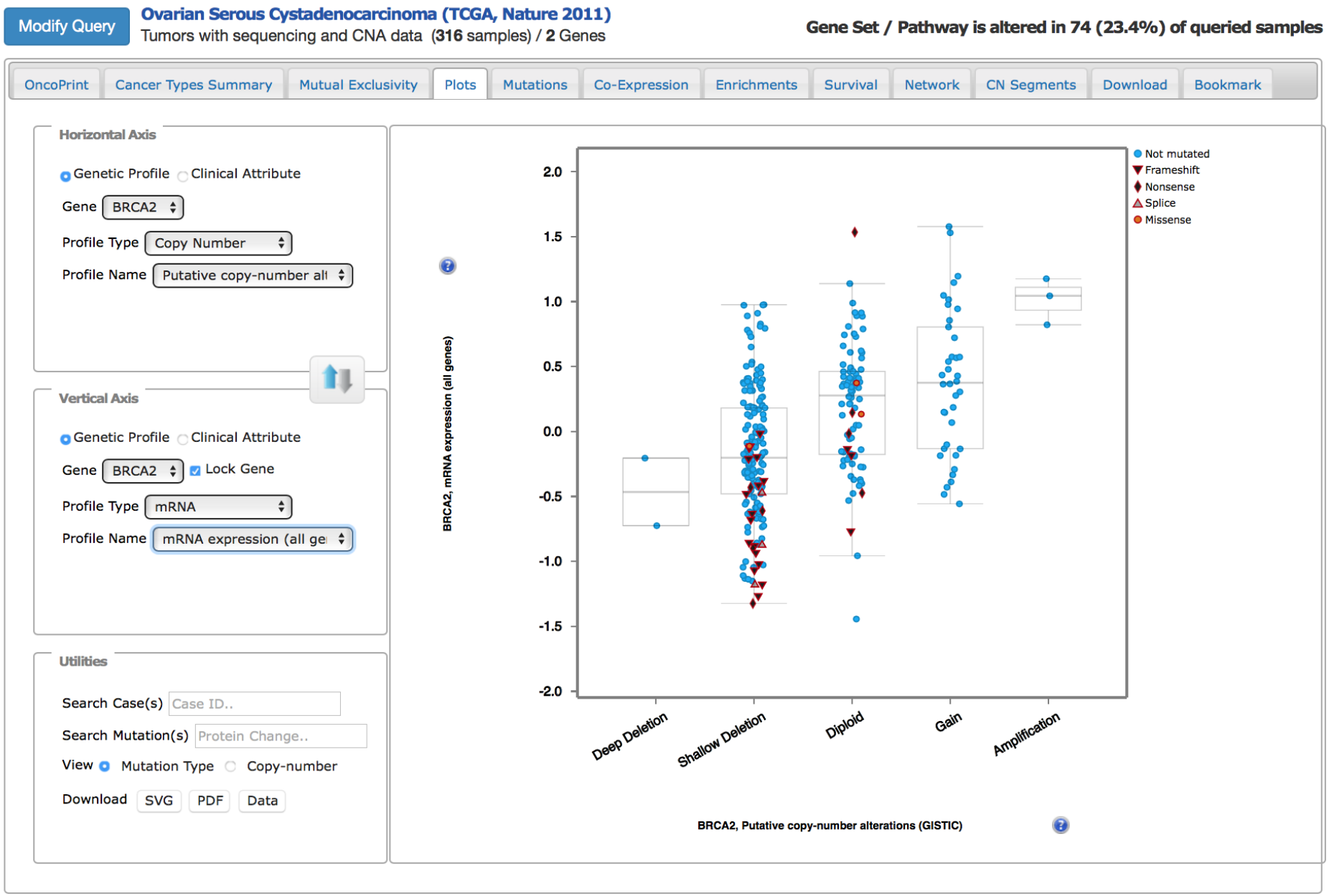
TCGA through cBioPortal: mutual exclusivity

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



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

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

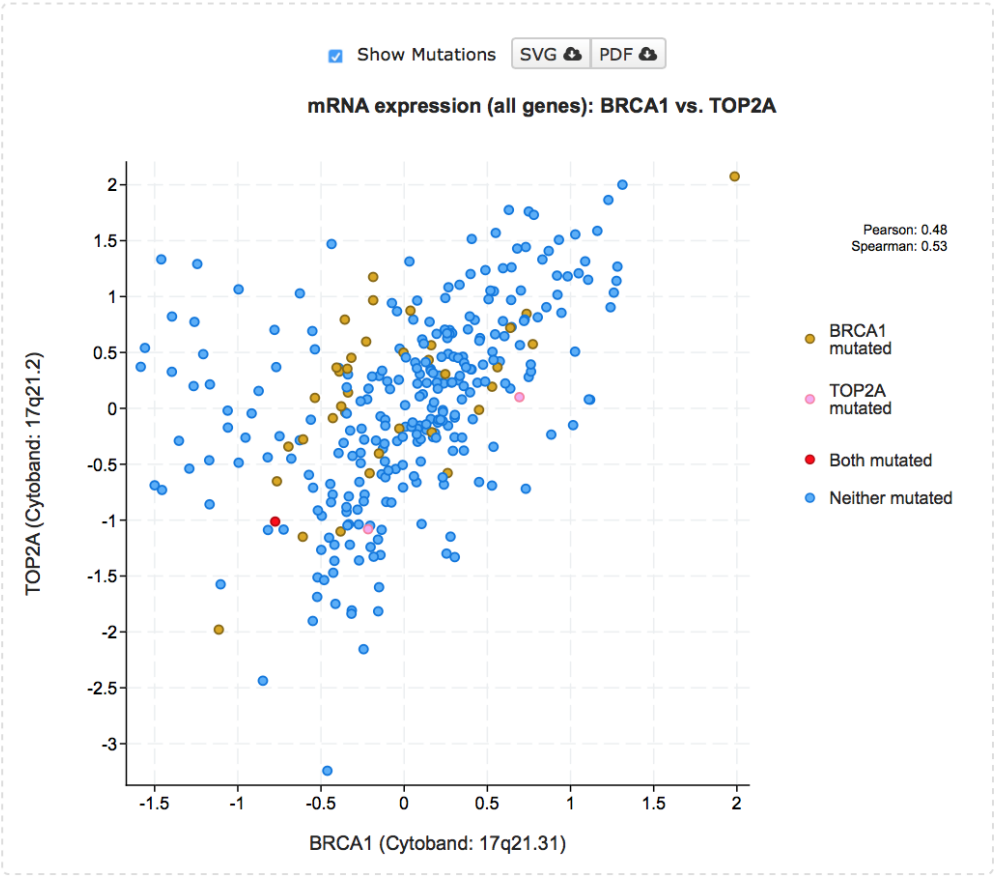
Data Set: mRNA expression (all genes) (316 samples)

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
RNFT2	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

Showing 1-25 of 18505 | Show more



TCGA through cBioPortal: survival

Compares survival among gene altered and non-altered patients

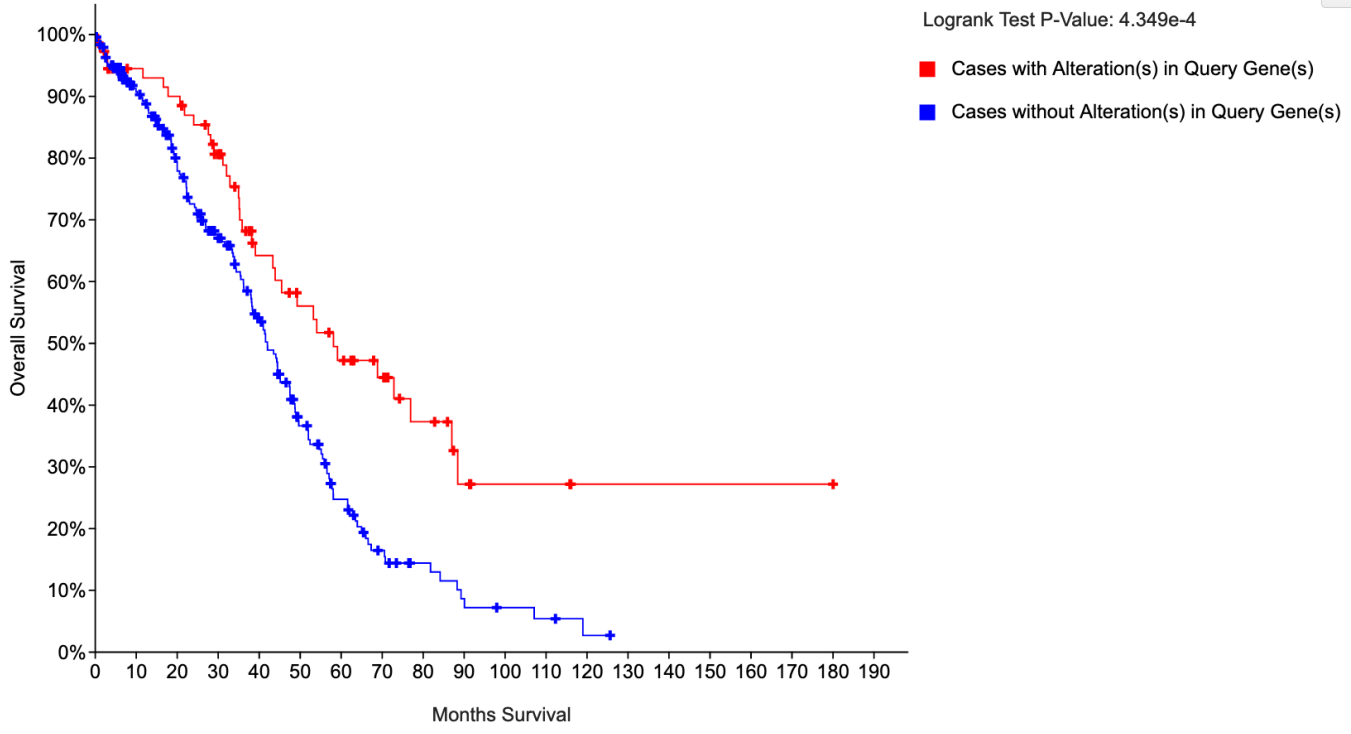
Modify Query

Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)
Samples with mutation and CNA data (**316** patients/samples) - BRCA1 & BRCA2

Queried genes are altered in 74 (23%) of queried patients/samples

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

Overall Survival Kaplan-Meier Estimate (Overall patient survival status.)

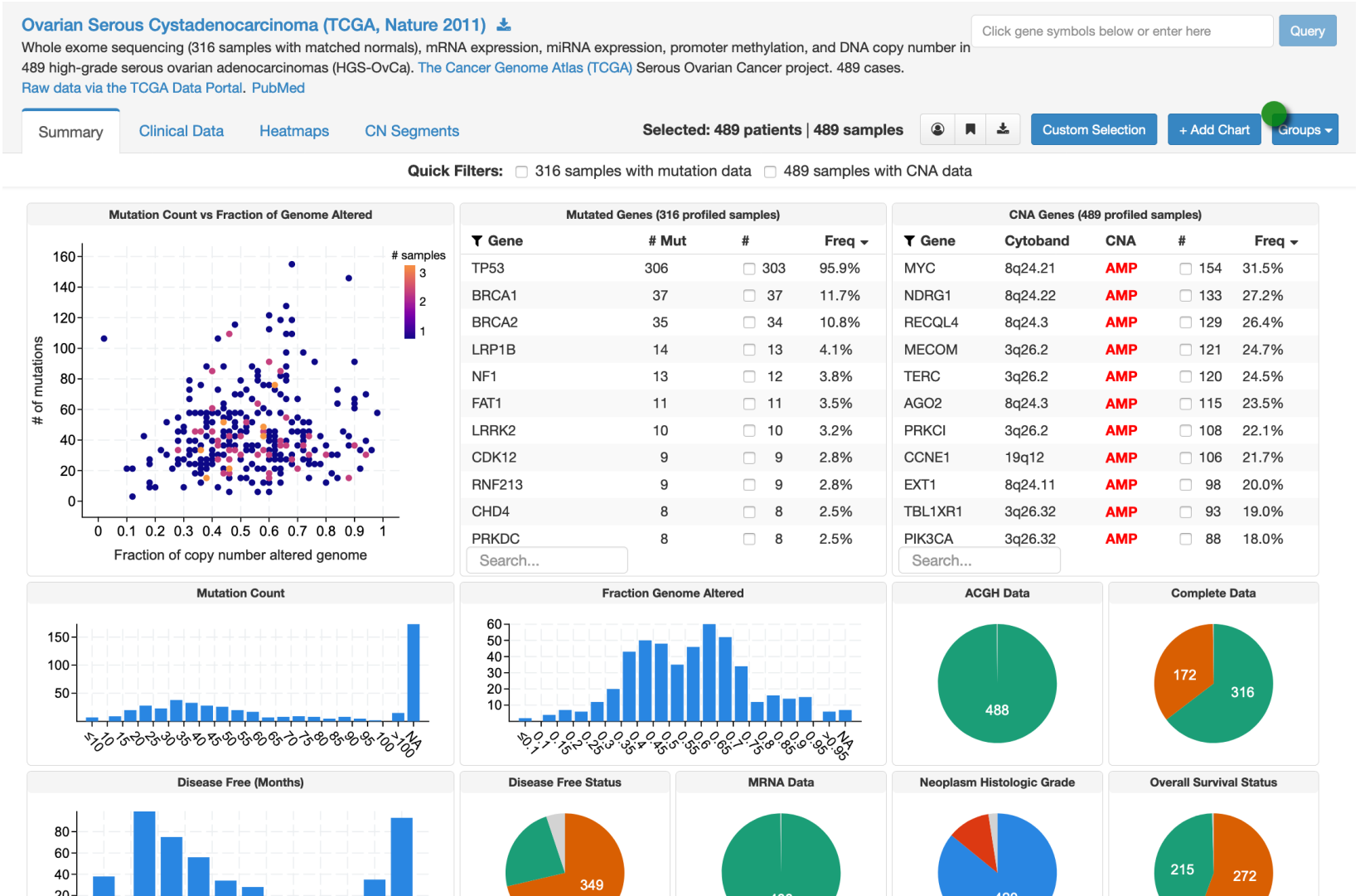


	Number of Cases, Total	Number of Cases, Deceased	Median Months Survival
Cases with Alteration(s) in Query Gene(s)	**73**	**35**	**58.12**
Cases without Alteration(s) in Query Gene(s)	**242**	**146**	**41.99**

TCGA through cBioPortal: study summaries

Access by clicking the study name wherever you see it

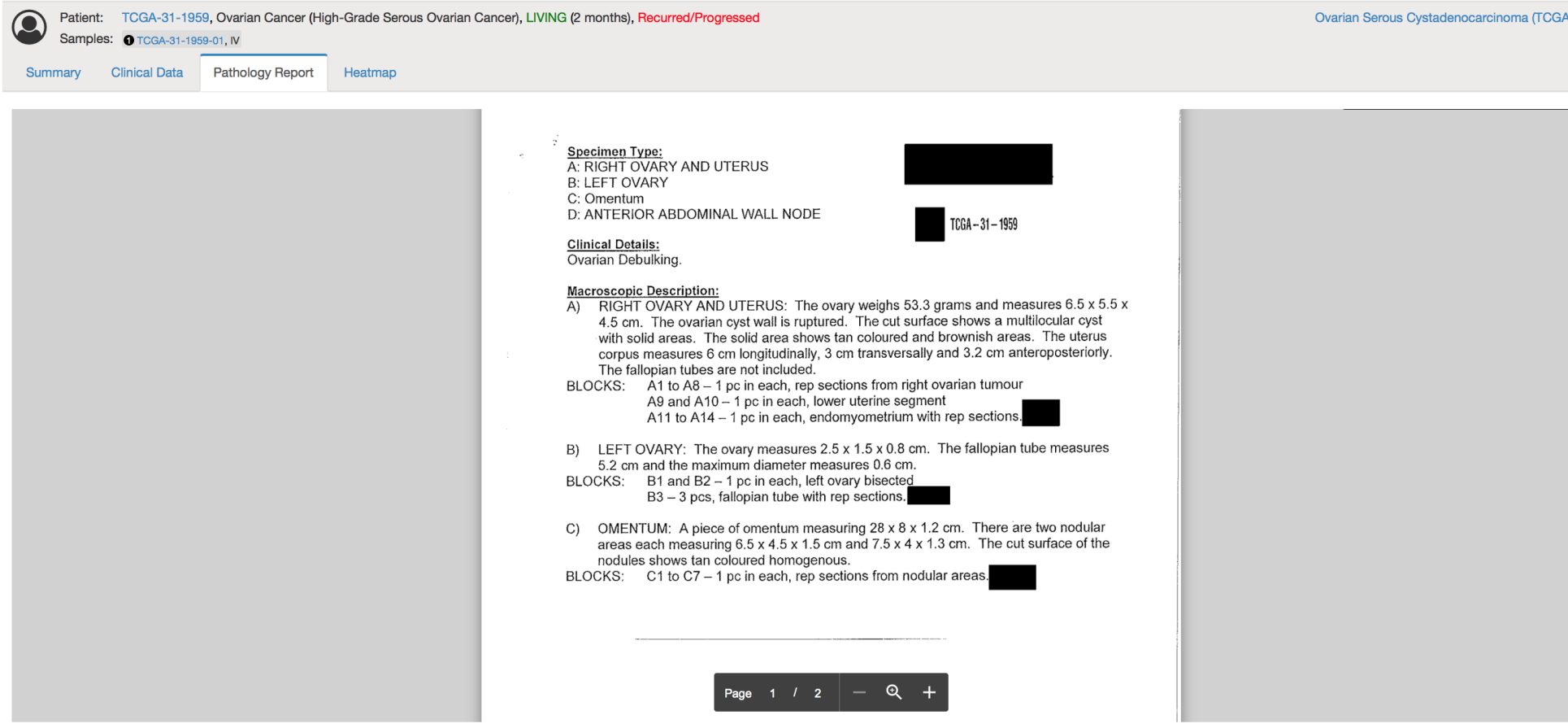
- Very useful high-level study summaries
- Click Clinical Data tab → DATA to download all clinical data*
- Merge with Download tab from gene queries for custom analyses



TCGA through cBioPortal: patient views

Access by clicking a patient ID wherever you see one

Includes detailed summary of mutations + clinical data



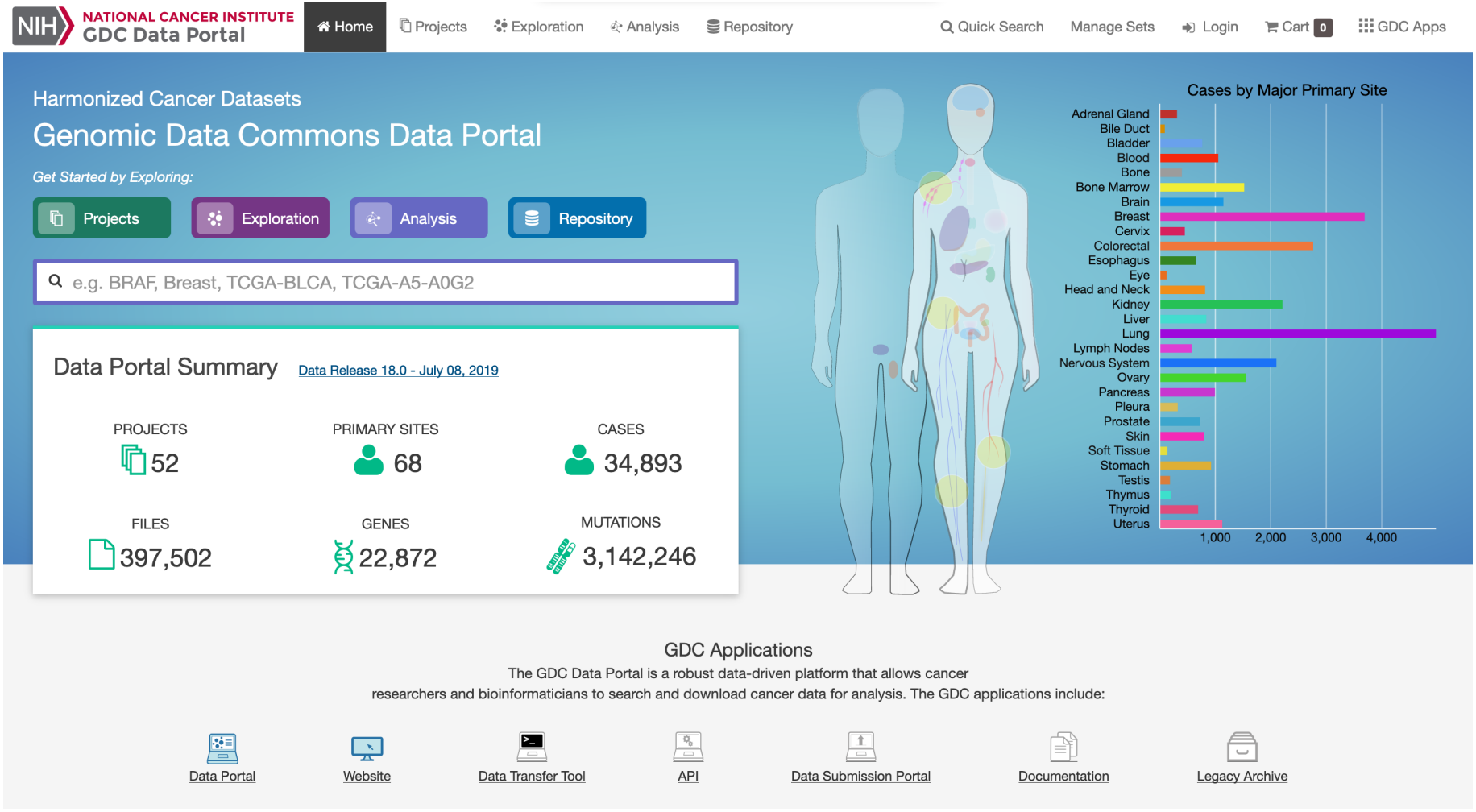
TCGA through GDC

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

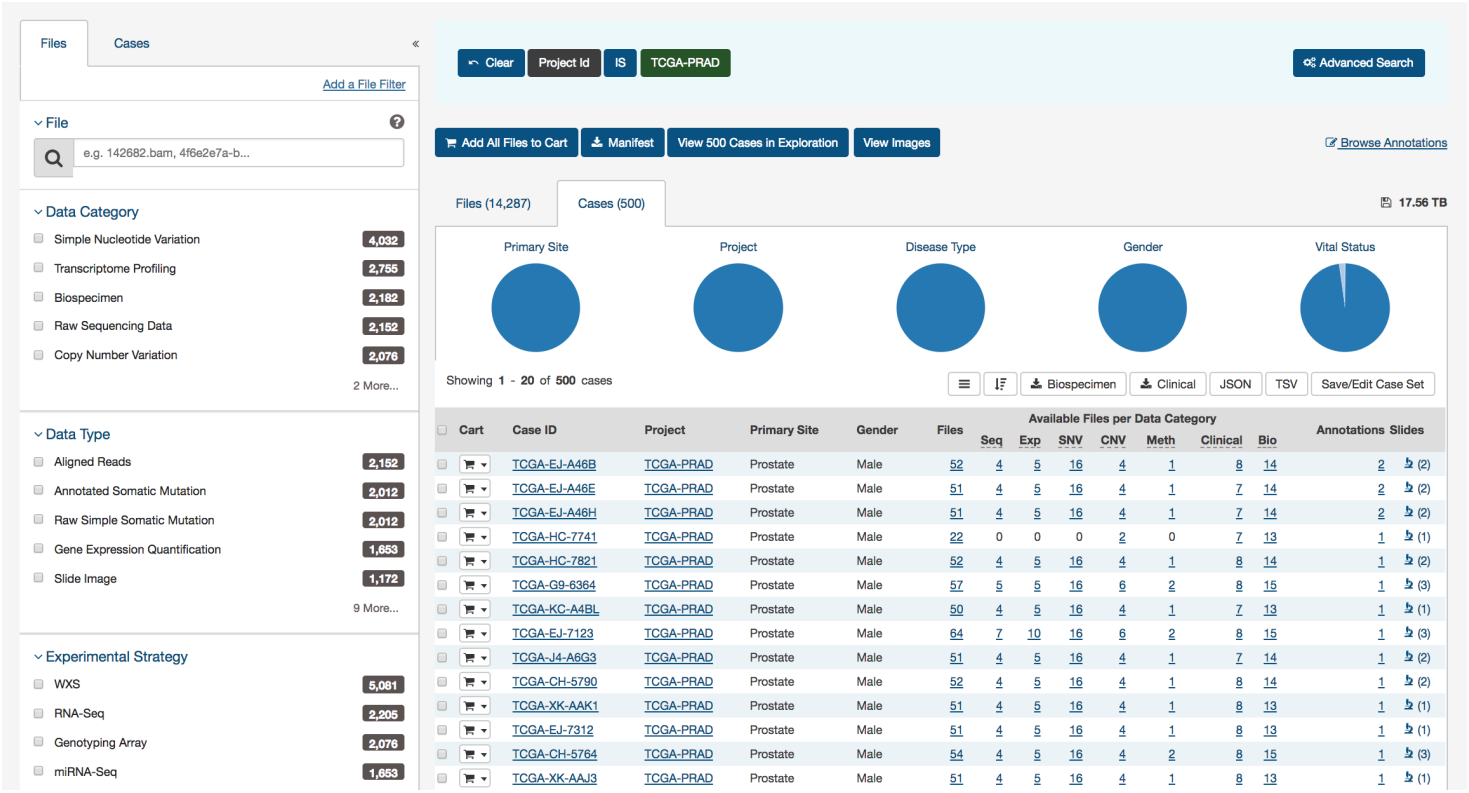
Also, analyses/UI within cBio are designed for small # of genes

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



TCGA through GDC: querying

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



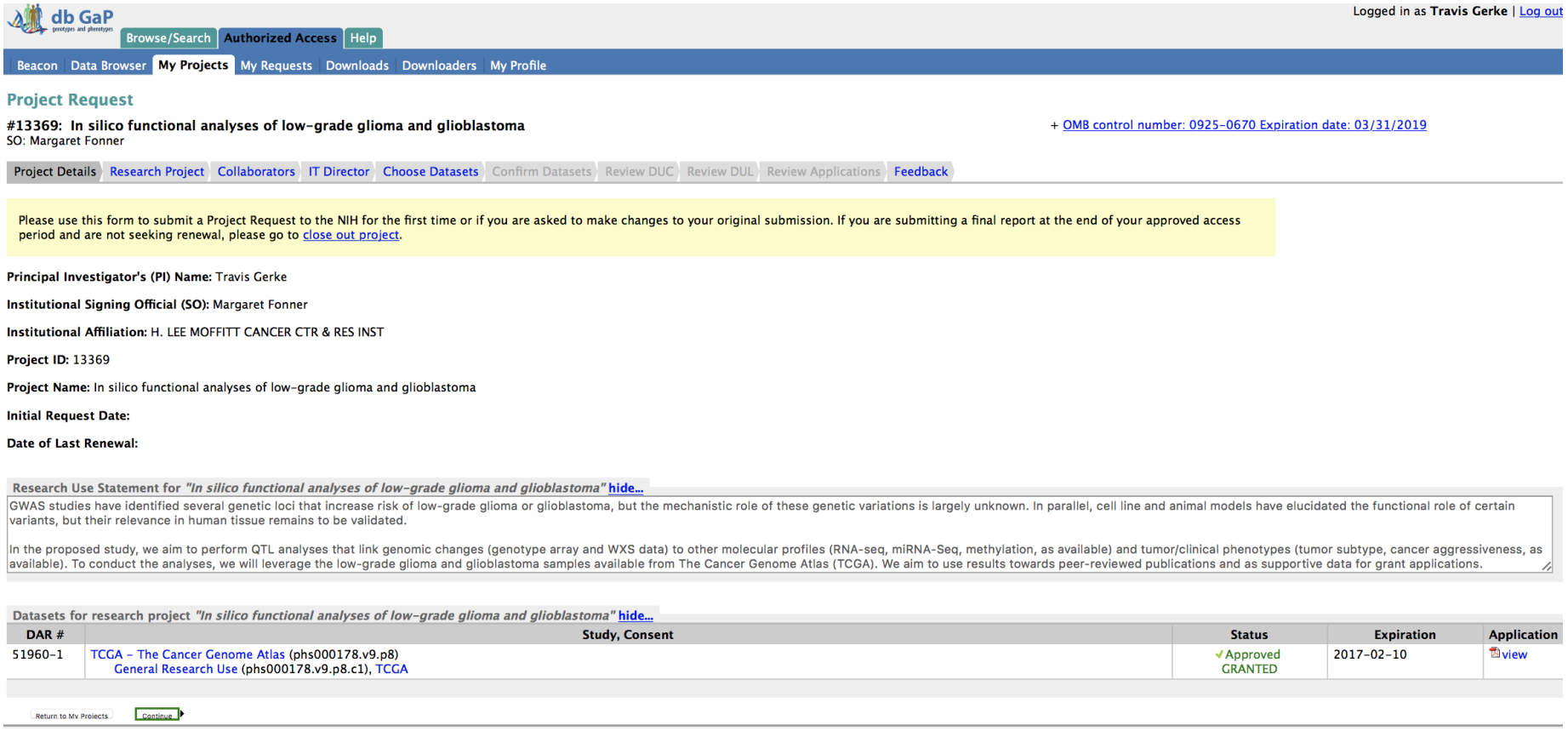
For larger downloads, use the manifest file and Data Transfer Tool [link]

- 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

Logged in as **Travis Gerke** | Log out

Browse/Search | Authorized Access | Help

Beacon | Data Browser | My Projects | My Requests | Downloads | Downloaders | 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 (SO): 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...

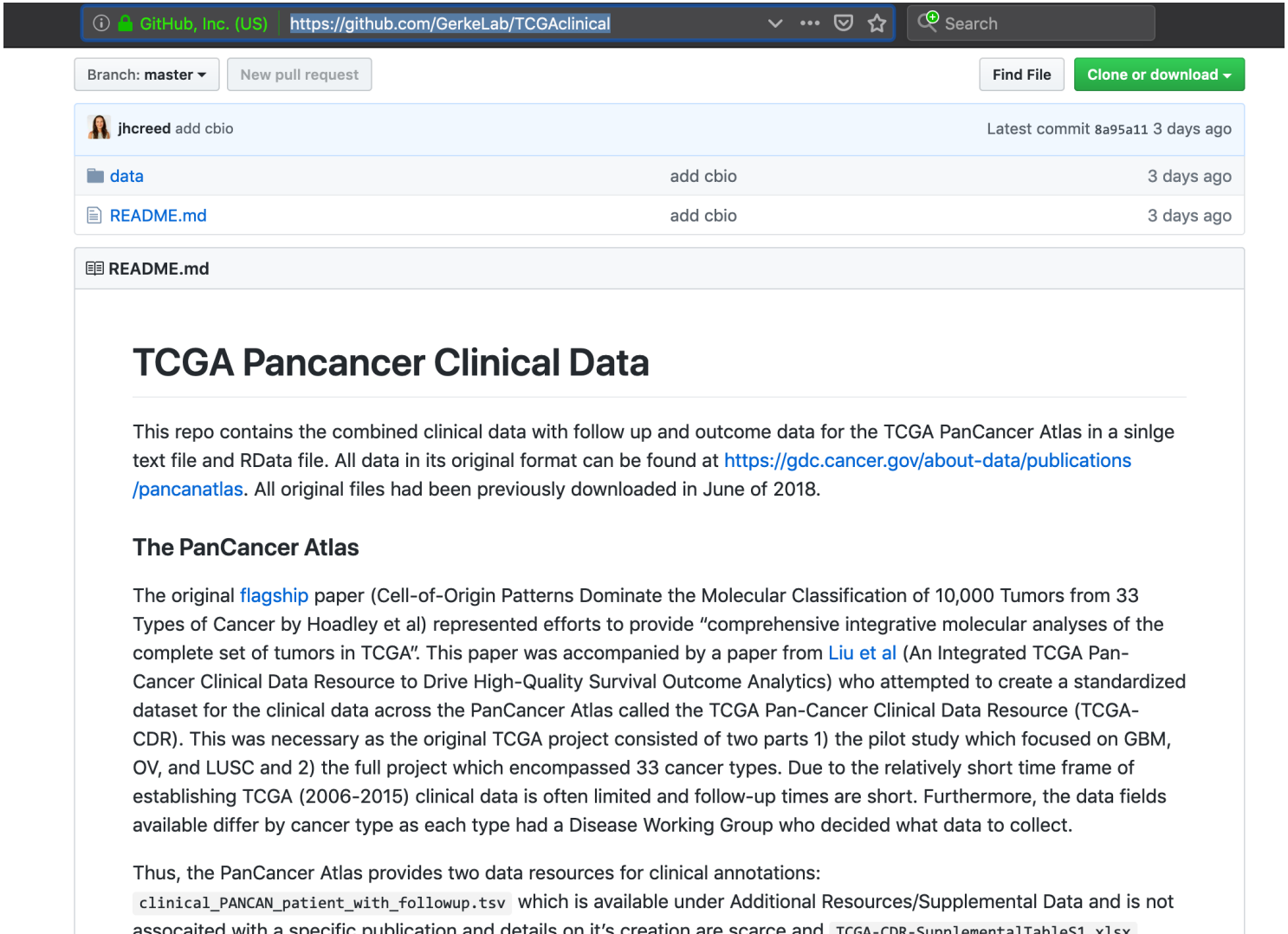
DAR #	Study, Consent	Status	Expiration	Application
51960-1	TCGA - The Cancer Genome Atlas (phs000178.v9.p8) General Research Use (phs000178.v9.p8.c1), TCGA	✓Approved GRANTED	2017-02-10	view

Return to My Projects | Continue

A final note about TCGA clinical data

Be aware of which version/source you use!

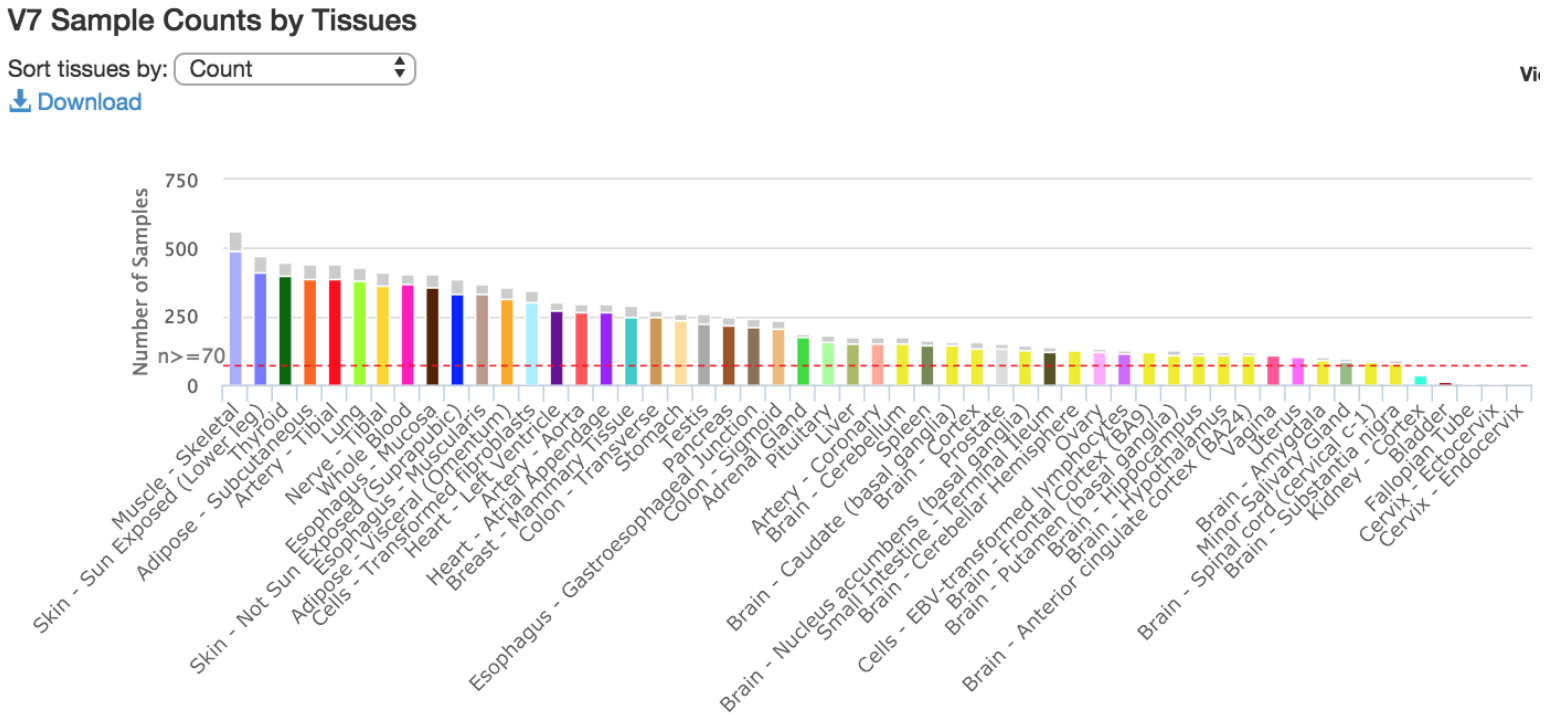
- Pan-Cancer group released a version with recommendations Liu Cell 2018
- Often Pan-Cancer $\neq$ Provisional $\neq$ Published $\neq$ cBio $\neq \cdots$
- E.g. we previously explored a list of such differences [link]



The Genotype-Tissue Expression project (GTEx)

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

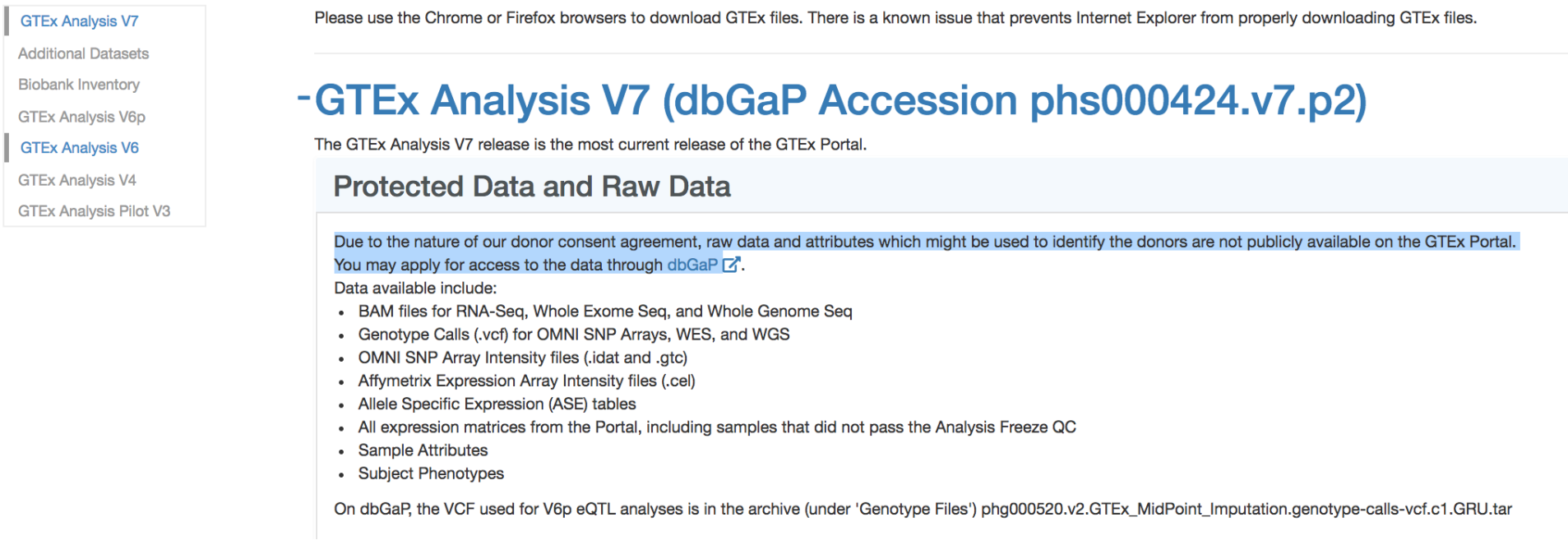
- Rapid autopsy program > 700 donors $\implies > 11,000$ samples
- *Not* cancer patients
- Can be useful to inform what genes are expressed in which tissues in a "normal" state



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.

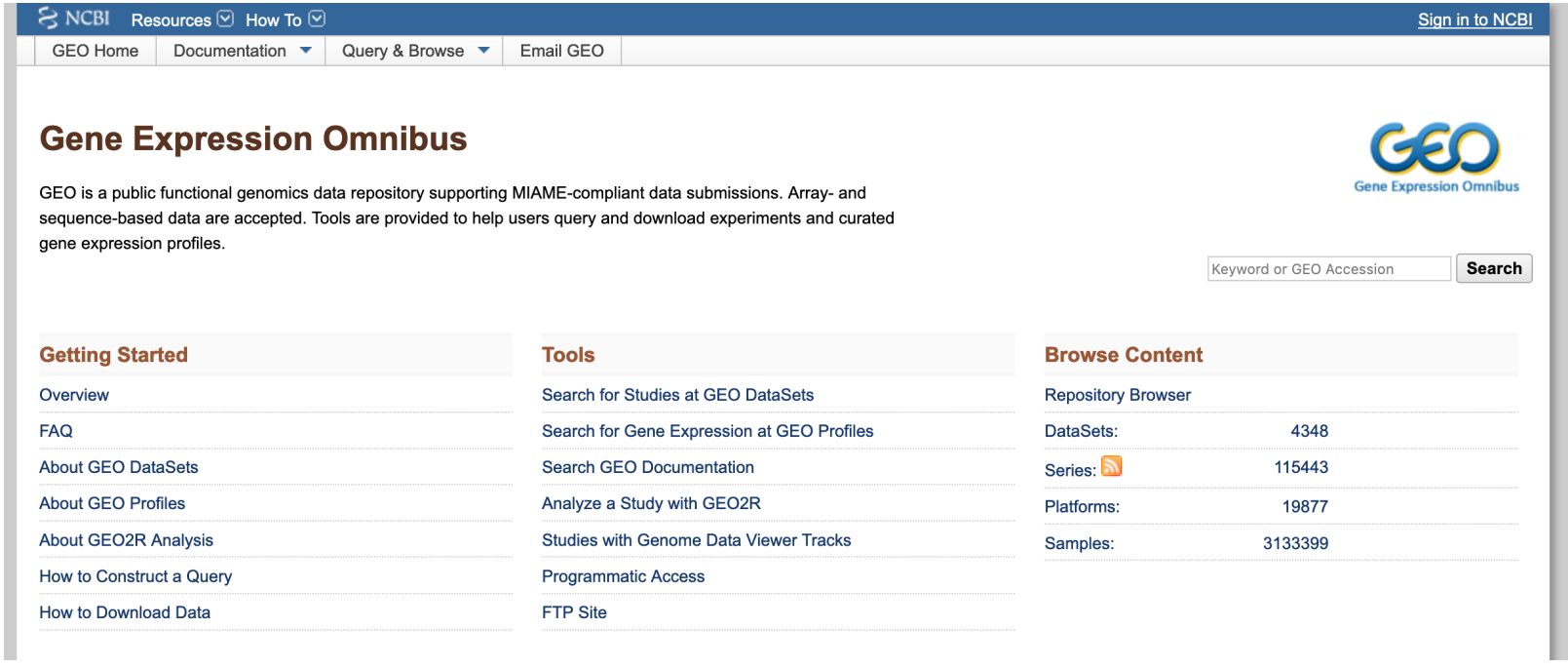
Data available include:

- 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 (.idat 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]

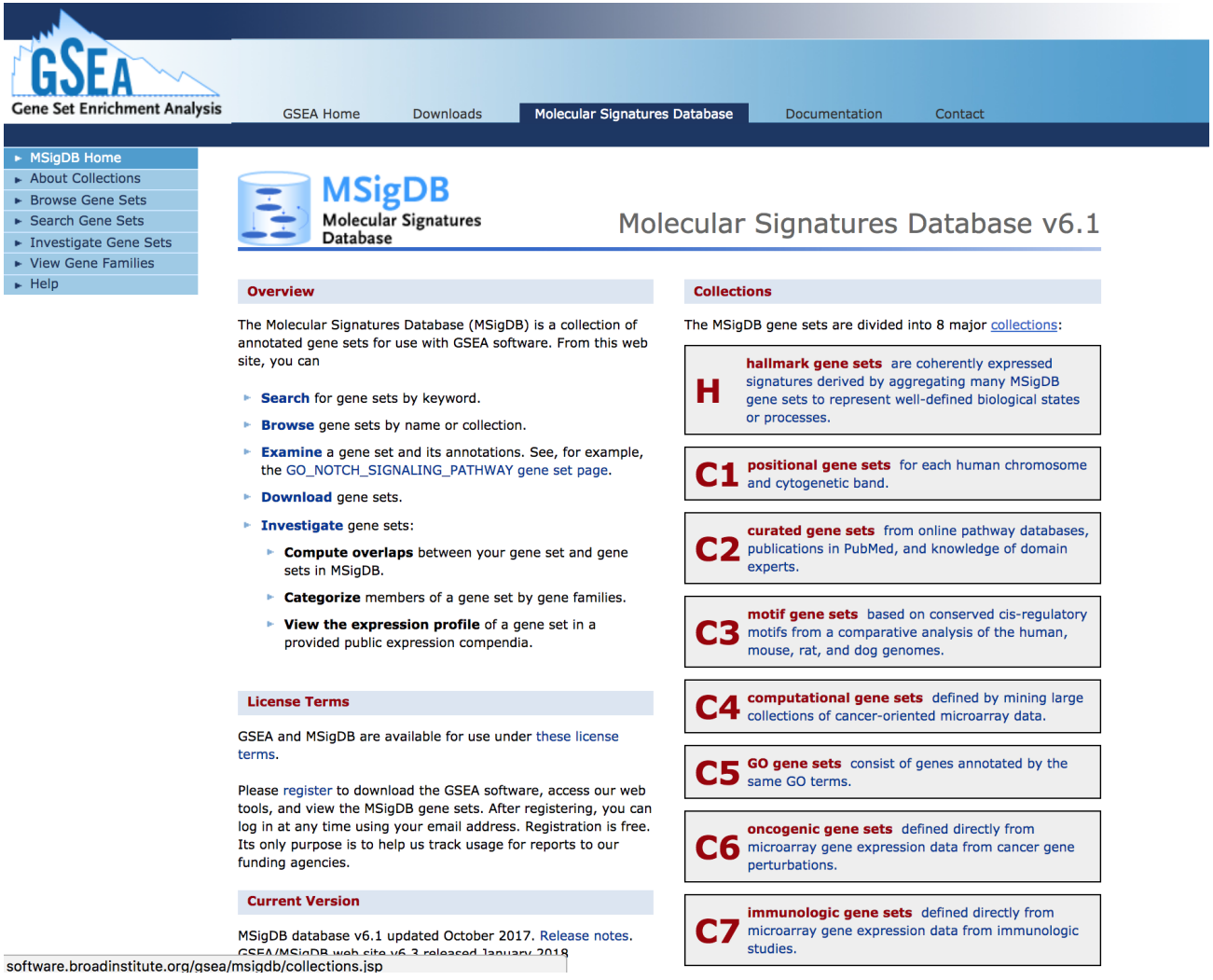
- Published studies of high-throughput data are very often publicly available
  - Typically, funding mechanism/journal required for reproducibility
  - When GEO is most useful: studies with long-term follow-up or epidemiologic data (that resources like TCGA/GTEx lack)
- Using this data requires programming (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]



Molecular Signatures Database (MSigDB) [link]

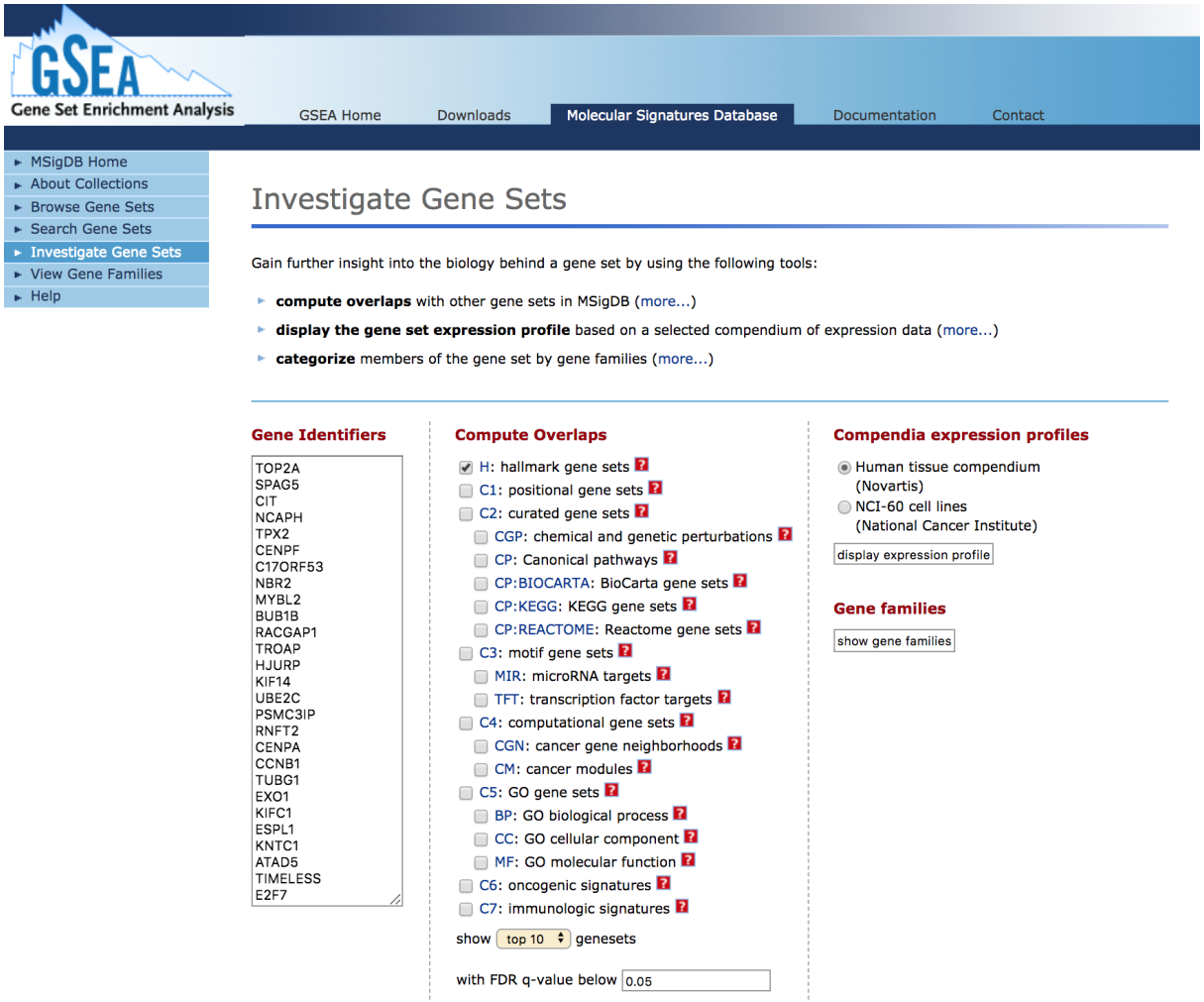
Rich compendium of annotated gene set collections

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



An MSigDB overlap query

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



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