The Cancer Genome Atlas (TCGA) & International Cancer Genome Consortium (ICGC)

Session 4 – Rebecca Poulos

Never Stand Still | Medicine | Prince of Wales Clinical School

Introductory bioinformatics for human genomics workshop, UNSW

20th – 21st April 2017

ADULT CANCER PROGRAM

LOWY CANCER RESEARCH CENTRE

tcrn bridging the gap translational cancer research network

NSW GOVERNMENT

cancer institute NSW

Facts on cancer

- An estimated 134,000 new cases of cancer will be diagnosed in Australia this year, with that number set to rise to 150,000 by 2020

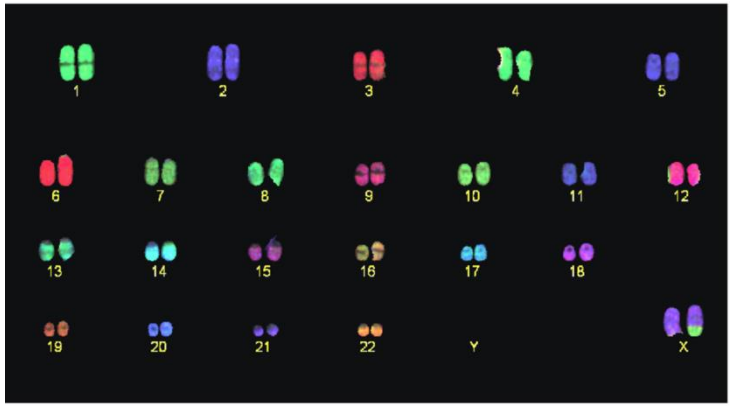
- Cancer is a leading cause of death in Australia. In 2014, > 44,000 people died from cancer, accounting for about 3 in every 10 deaths.

Source: Cancer Council Australia (2017)

Cancer is a disease of the genome

- Challenges in treating cancer:
  - Every patient is different
  - Every tumour is different, even in the same patient
  - Tumours can be highly heterogeneous
  - High rate of genomic abnormalities (few drivers, many passenger mutations)

Healthy 46 chromosomes



Example cancer 59 chromosomes

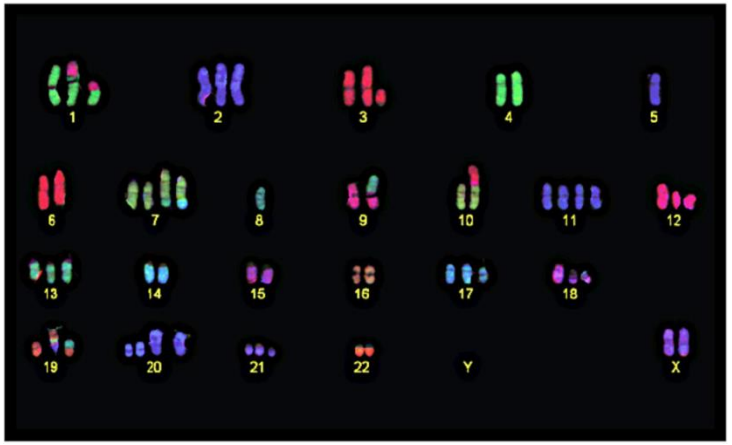


Image from Thompson & Compton *Chromosome Res* 2011.

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array; SNP array; WGS

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array; SNP array; WGS
DNA methylation	Methylation array; RRBS; WGBS

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array; SNP array; WGS
DNA methylation	Methylation array; RRBS; WGBS
mRNA expression changes	mRNA expression array; RNA-seq

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array; SNP array; WGS
DNA methylation	Methylation array; RRBS; WGBS
mRNA expression changes	mRNA expression array; RNA-seq
miRNA expression changes	miRNA expression array; miRNA-seq

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

What can go wrong in cancer genomes?

Types of changes	Some common technologies used to study these changes
DNA mutations - Point mutations - Insertions & deletions	WGS; WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array; SNP array; WGS
DNA methylation	Methylation array; RRBS; WGBS
mRNA expression changes	mRNA expression array; RNA-seq
miRNA expression changes	miRNA expression array; miRNA-seq
Protein expression	Protein arrays; mass spectrometry

WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

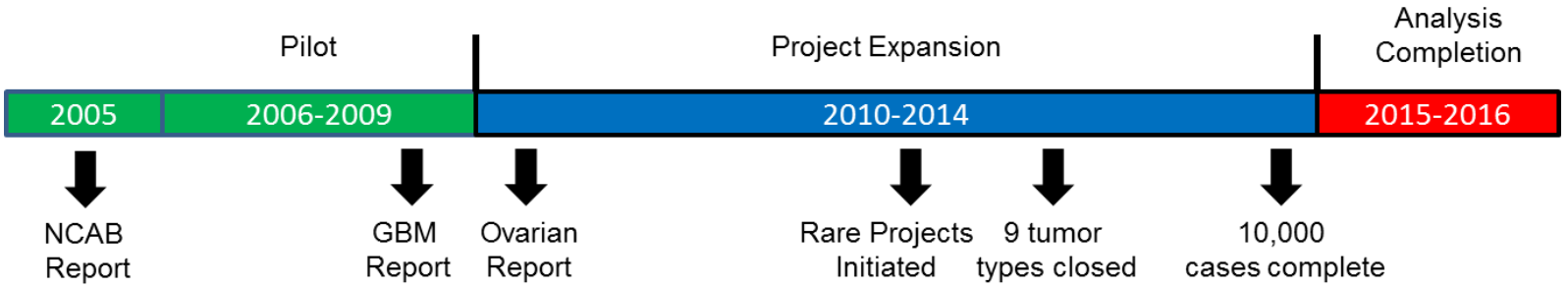
Goal of cancer genomics

- Identify changes in the genomes of tumors that drive cancer progression
- Understand how normal cells become cancerous
- Identify new targets for therapy
- Select drugs based on the genomics of the tumour – i.e. personalised therapy

Cancer Sequencing Projects

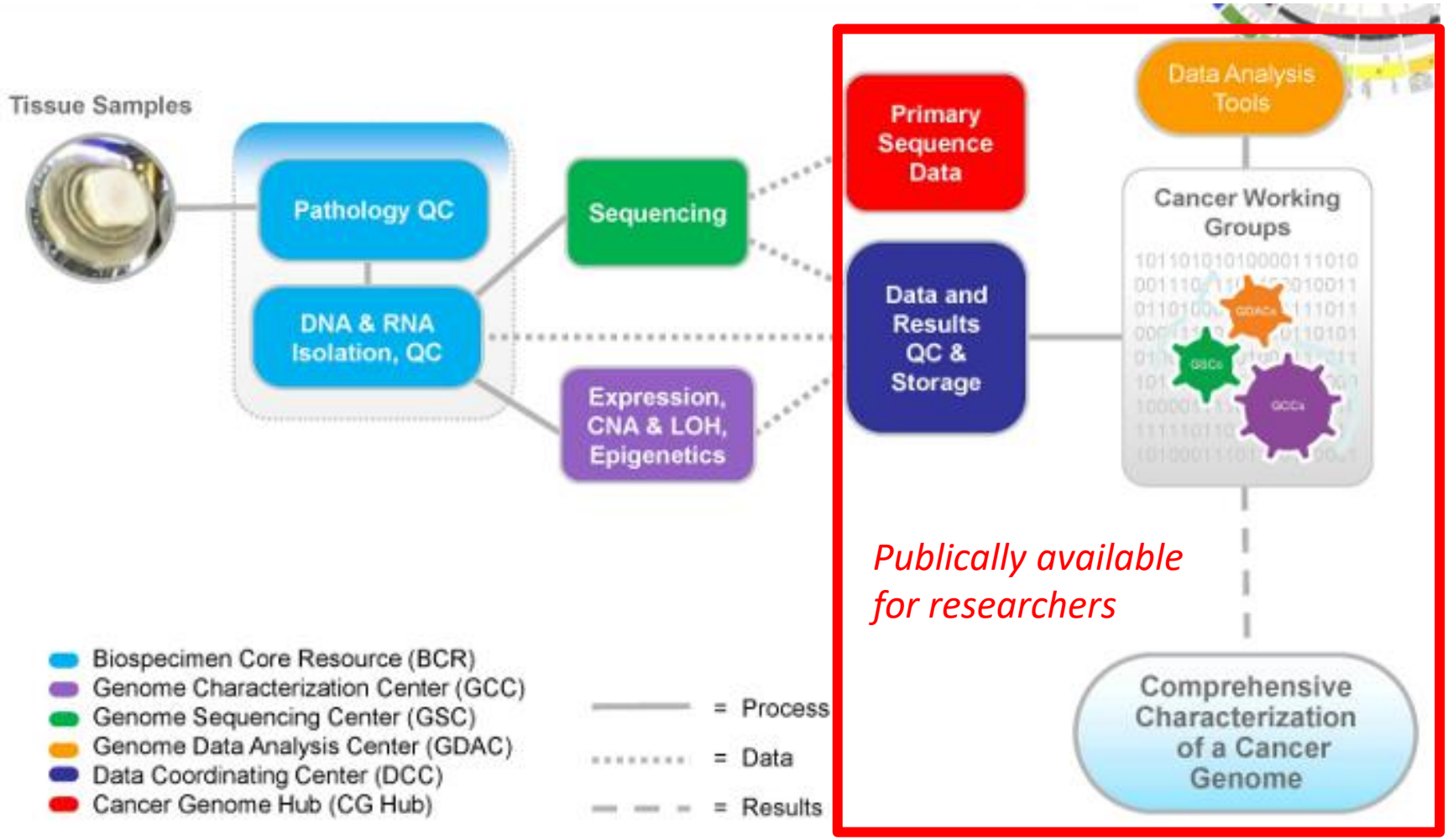
The Cancer Genome Atlas (TCGA)

- Led by NIH
- Initiated in 2006 (as a pilot program) and expanded in 2009



- **Aim:**
  To make the genomes of 20 cancers publically available
- **Update today:**
  33 cancer types & subtypes analysed (11,000 samples)

TCGA pipeline



Types of Cancers

- **Breast**
  - Ductal carcinoma
  - Lobular carcinoma
- **Central nervous system**
  - Glioblastoma multiforme
  - Lower grade glioma
- **Endocrine**
  - Adrenocortical carcinoma
  - Papillary thyroid carcinoma
  - Paraganglioma and pheochromocytoma
- **Gastrointestinal**
  - Cholangiocarcinoma
  - Colorectal Adenocarcinoma
  - Liver Hepatocellular Carcinoma
  - Pancreatic Ductal Adenocarcinoma
  - Stomach-Esophageal Cancer
- **Gynecological**
  - Cervical Cancer
  - Ovarian Serous Cystadenocarcinoma
  - Uterine Carcinosarcoma
  - Uterine Corpus Endometrial Carcinoma
- **Head and neck**
  - Squamous cell carcinoma
  - Uveal melanoma
- **Hematologic**
  - Acute myeloid leukemia
  - Thymoma
- **Skin**
  - Cutaneous melanoma
- **Soft tissue**
  - Sarcoma
- **Thoracic**
  - Lung Adenocarcinoma
  - Lung Squamous Cell Carcinoma
  - Mesothelioma
- **Urologic**
  - Chromophobe Renal Cell Carcinoma
  - Clear Cell Kidney Carcinoma
  - Papillary Kidney Carcinoma
  - Prostate Adenocarcinoma
  - Testicular Germ Cell Cancer
  - Urothelial Bladder Carcinoma

Datasets

Data types

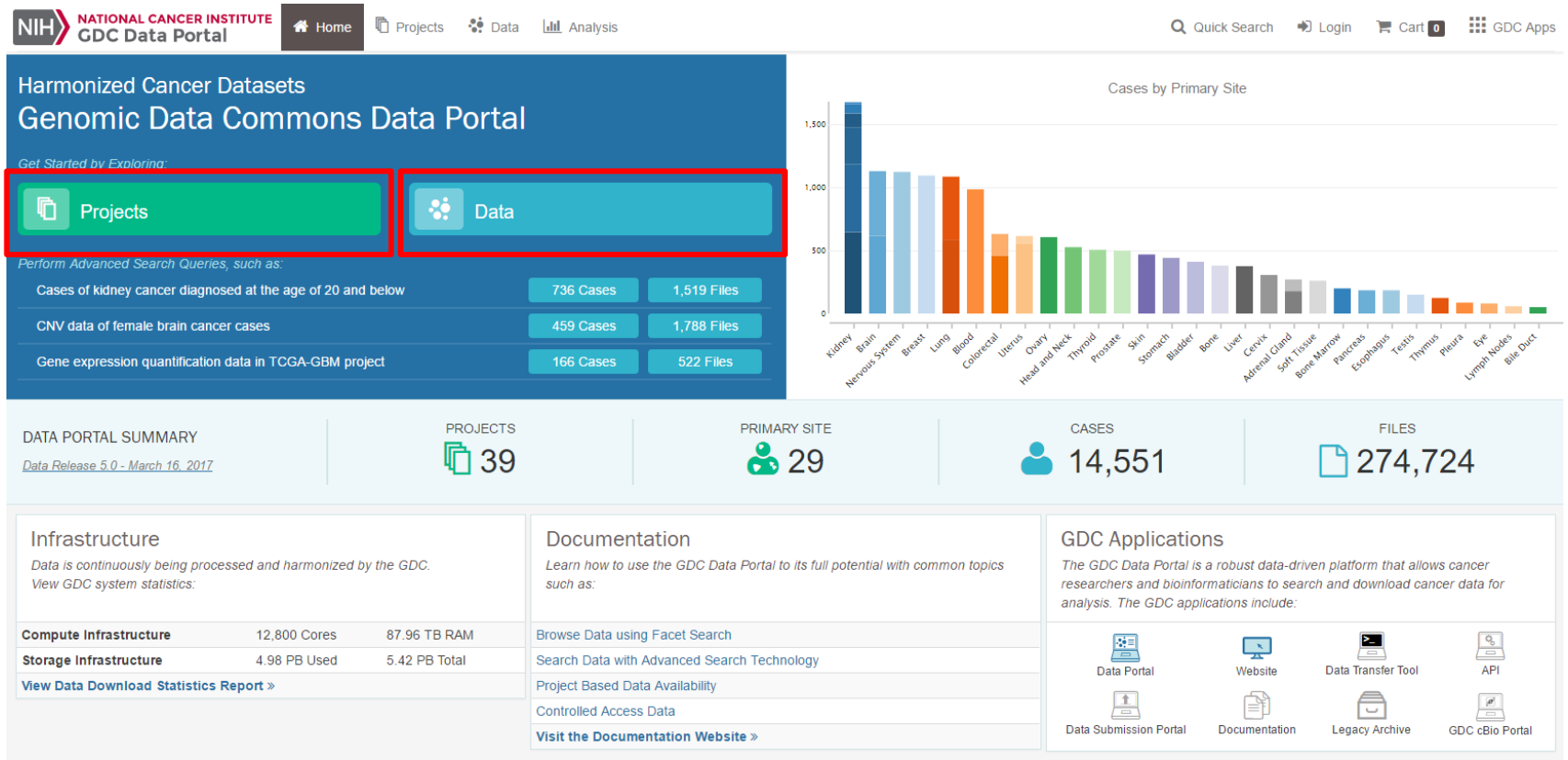
- Clinical data
- Images
- Microsatellite instability
- DNA sequencing
- miRNA sequencing
- Protein expression
- mRNA & RNA sequencing
- Array-based expression
- DNA methylation
- Copy number

Data access tiers

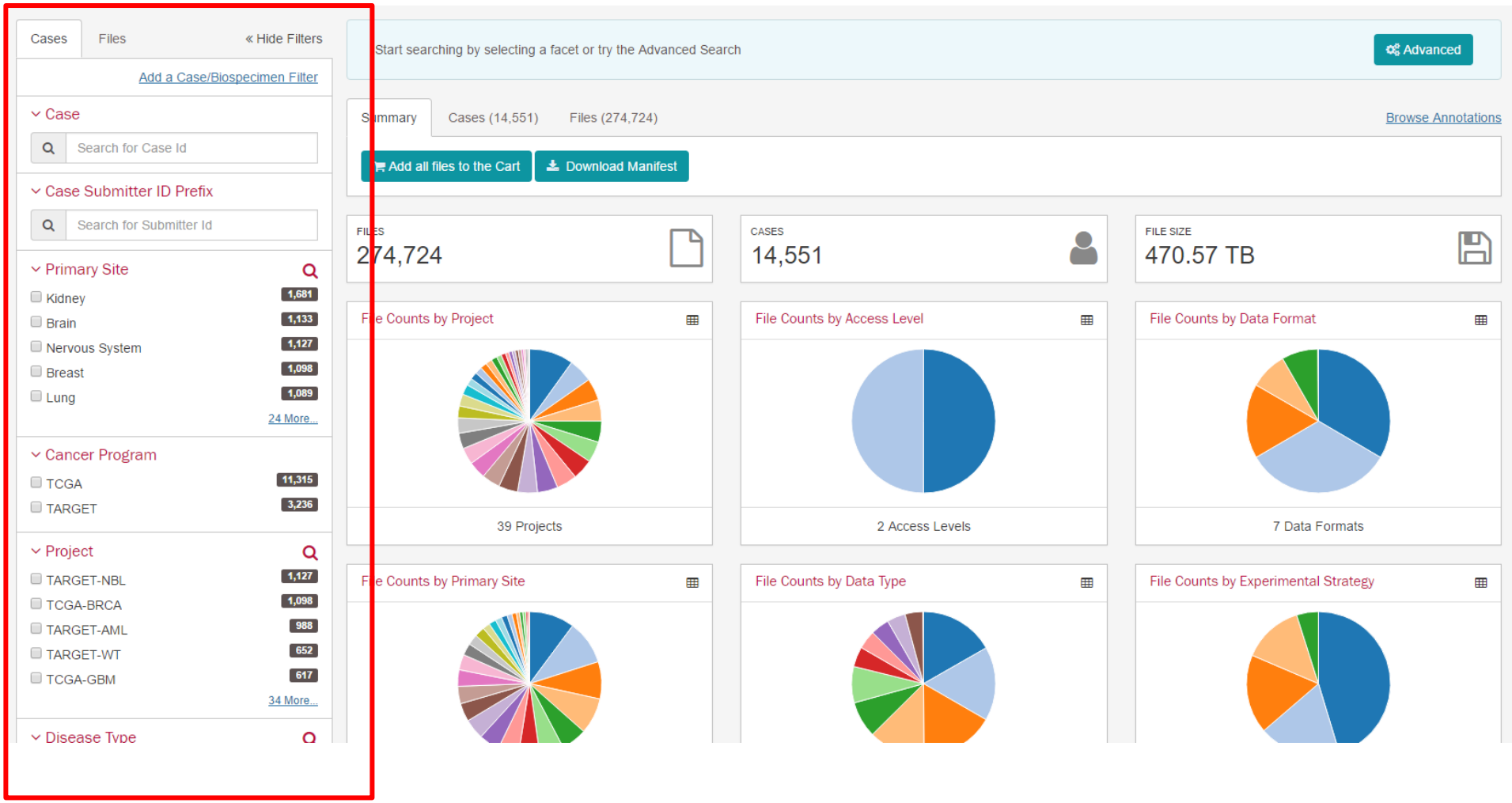
- Open access
  - De-identified
  - Requires no certification
- Controlled access
  - No direct identifiers
  - Must complete Data Access Request (DAR) form

Genomic Data Commons (GDC)

- TCGA data is stored on the Genomic Data Commons (GDC) data portal: https://portal.gdc.cancer.gov/

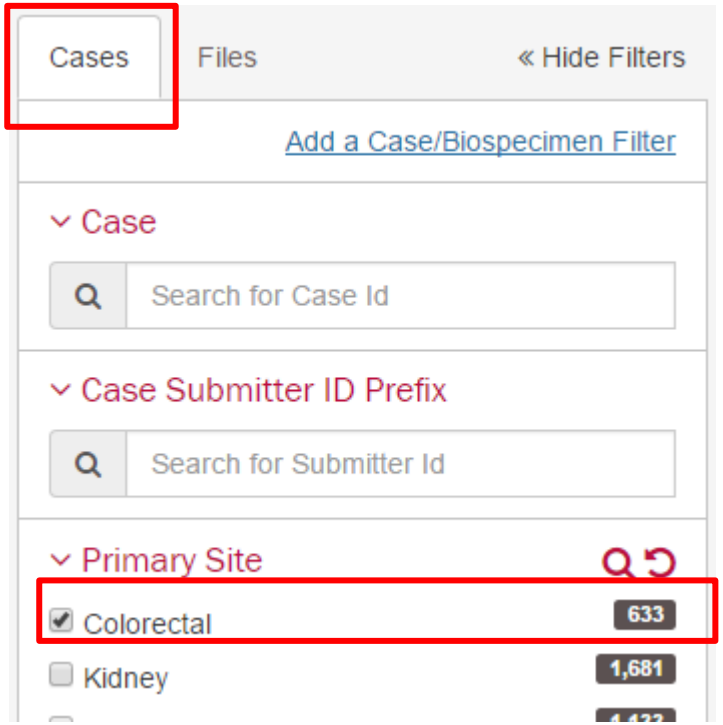


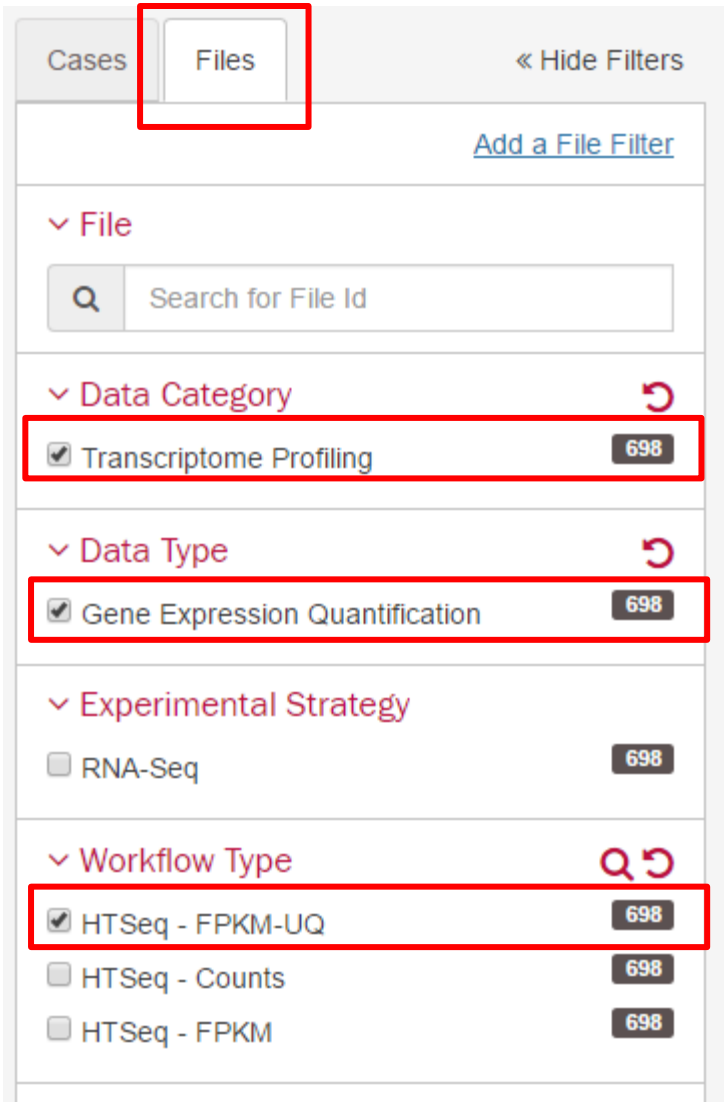
Exploring the “Data” option…



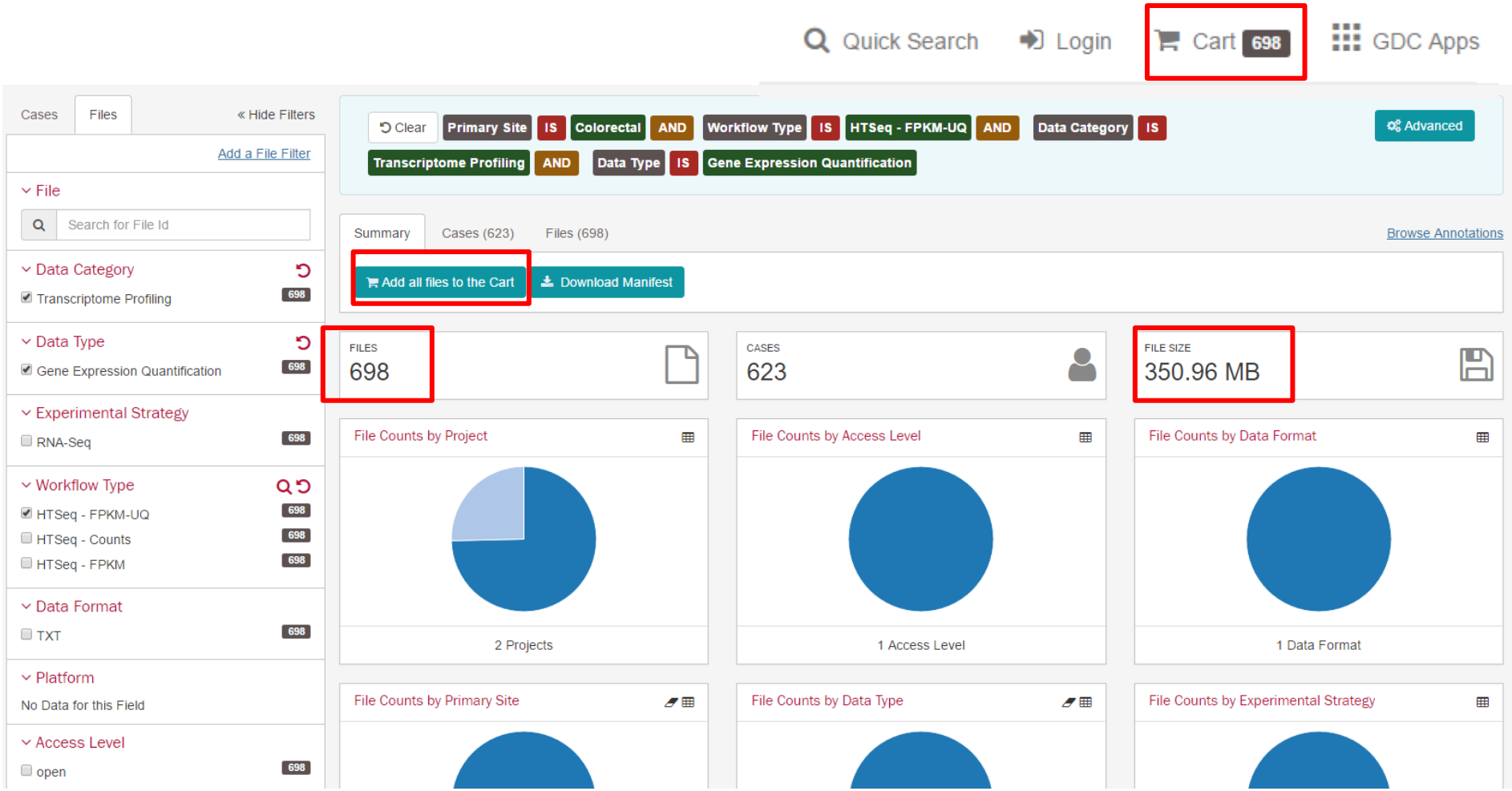
Search and filter files using this utility

Let's find all processed RNA-seq data for colorectal cancer…

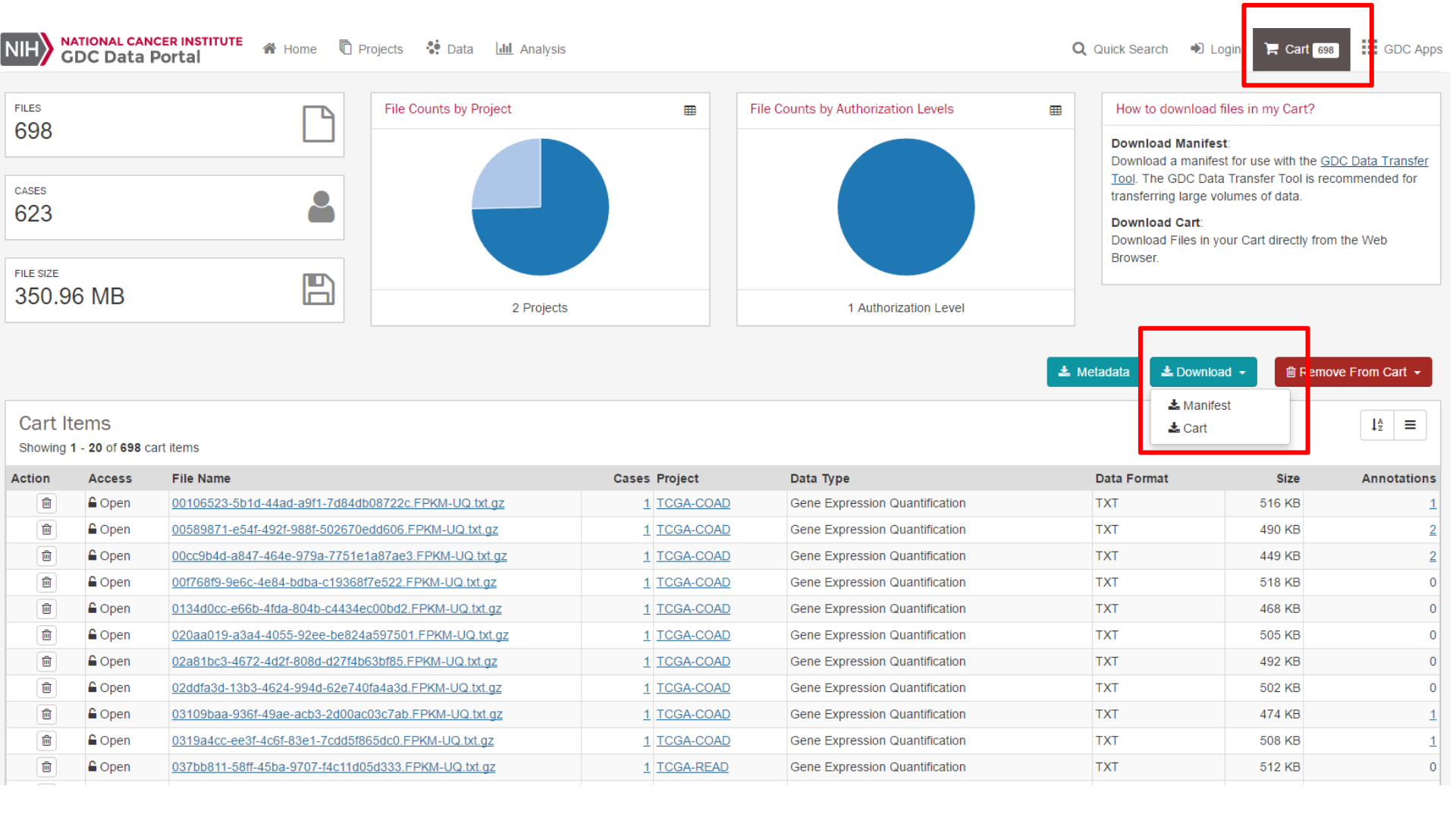




Let's find all processed RNA-seq data for colorectal cancer…



Let's find all processed RNA-seq data for colorectal cancer…



Action	Access	File Name	Cases	Project	Data Type	Data Format	Size	Annotations
	Open	00106523-5b1d-44ad-a9f1-7d84db08722c.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	516 KB	1
	Open	00589871-e54f-492f-988f-502670edd606.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	490 KB	2
	Open	00cc9b4d-a847-464e-979a-7751e1a87ae3.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	449 KB	2
	Open	00f768f9-9e6c-4e84-bdba-c19368f7e522.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	518 KB	0
	Open	0134d0cc-e66b-4fda-804b-c4434ec00bd2.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	468 KB	0
	Open	020aa019-a3a4-4055-92ee-be824a597501.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	505 KB	0
	Open	02a81bc3-4672-4d2f-808d-d27f4b63bf85.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	492 KB	0
	Open	02ddfa3d-13b3-4624-994d-62e740fa4a3d.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	502 KB	0
	Open	03109baa-936f-49ae-acb3-2d00ac03c7ab.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	474 KB	1
	Open	0319a4cc-ee3f-4c6f-83e1-7cdd5f865dc0.FPKM-UQ.txt.gz	1	TCGA-COAD	Gene Expression Quantification	TXT	508 KB	1
	Open	037bb811-58ff-45ba-9707-f4c11d05d333.FPKM-UQ.txt.gz	1	TCGA-READ	Gene Expression Quantification	TXT	512 KB	0

Genomic Data Commons (GDC)

- The GDC data portal is very user-friendly
- GDC is ideal for downloading data in large tab delimited format – perfect for a bioinformatician
- However, data portal files are difficult to use for the average biologist
- Fortunately there are some alternatives:
  - cBioPortal (*www.cbioportal.org/*)
  - ICGC data portal (*http://dcc.icgc.org/*)

cBioPortal (*www.cbioportal.org/*)

- A data analysis portal to TCGA data
- Provides functions for visualisation, analysis and download of data.
- Maintained by Memorial Sloan-Kettering Cancer Center

Features of cBioPortal

- Visualising frequency of mutations
- Correlation between occurrence of mutations
- Correlation of expression and CNV or methylation
- Visualisation of mutations
- Survival analysis
- Network analysis

Gao et al (2013) Sci. Signal

cBioPortal for Cancer Genomics

Visualize, analyze, discover.

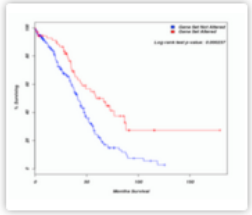
Memorial Sloan Kettering Cancer Center

HOME | DATA SETS | WEB API | R/MATLAB | TUTORIALS | FAQ | NEWS | TOOLS | ABOUT | VISUALIZE YOUR DATA

The cBioPortal for Cancer Genomics provides **visualization**, **analysis** and **download** of large-scale **cancer genomics** data sets.

Please adhere to the TCGA publication guidelines when using TCGA data in your publications.

Please cite Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012 when publishing results based on cBioPortal.



Query | **Download Data**

Select Cancer Study:

Search... No studies selected.

- All (121)
  - Adrenal Gland (1)
    - Adrenocortical Carcinoma (1)
      - **Adrenocortical Carcinoma (TCGA, Provisional)** *92 samples*
  - Biliary Tract (5)
    - Cholangiocarcinoma (4)
      - **Intrahepatic Cholangiocarcinoma (Johns Hopkins University, Nat Genet 2013)** *40 samples*
      - **Cholangiocarcinoma (National Cancer Centre of Singapore, Nat Genet 2013)** *15 samples*
      - **Cholangiocarcinoma (National University of Singapore, Nat Genet 2012)** *8 samples*

Select Data Type Priority: Mutation and CNA | Only Mutation | Only CNA

Enter Gene Set: Advanced: Onco Query Language (OQL)

User-defined List

Enter HUGO Gene Symbols or Gene Aliases

What's New

New Jobs available at Dana-Farber to work on cBioPortal

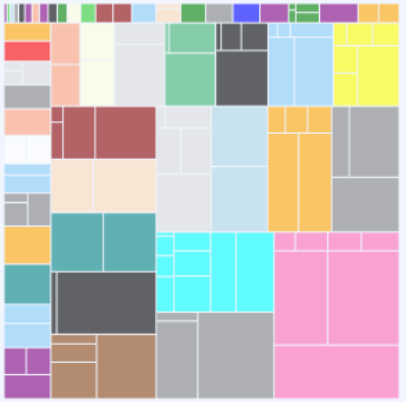
Sign up for low-volume email news alerts:

Subscribe

Or follow us @cbioportal on Twitter

Data Sets

The Portal contains **121 cancer studies.** [Details]



Example Queries

RAS/RAF alterations in colorectal cancer

BRCA1 and BRCA2 mutations in ovarian cancer

POLE hotspot mutations in endometrial cancer

TP53 and MDM2/4 alterations in GBM

PTEN mutations in GBM in text format

BRAF V600E mutations across cancer types

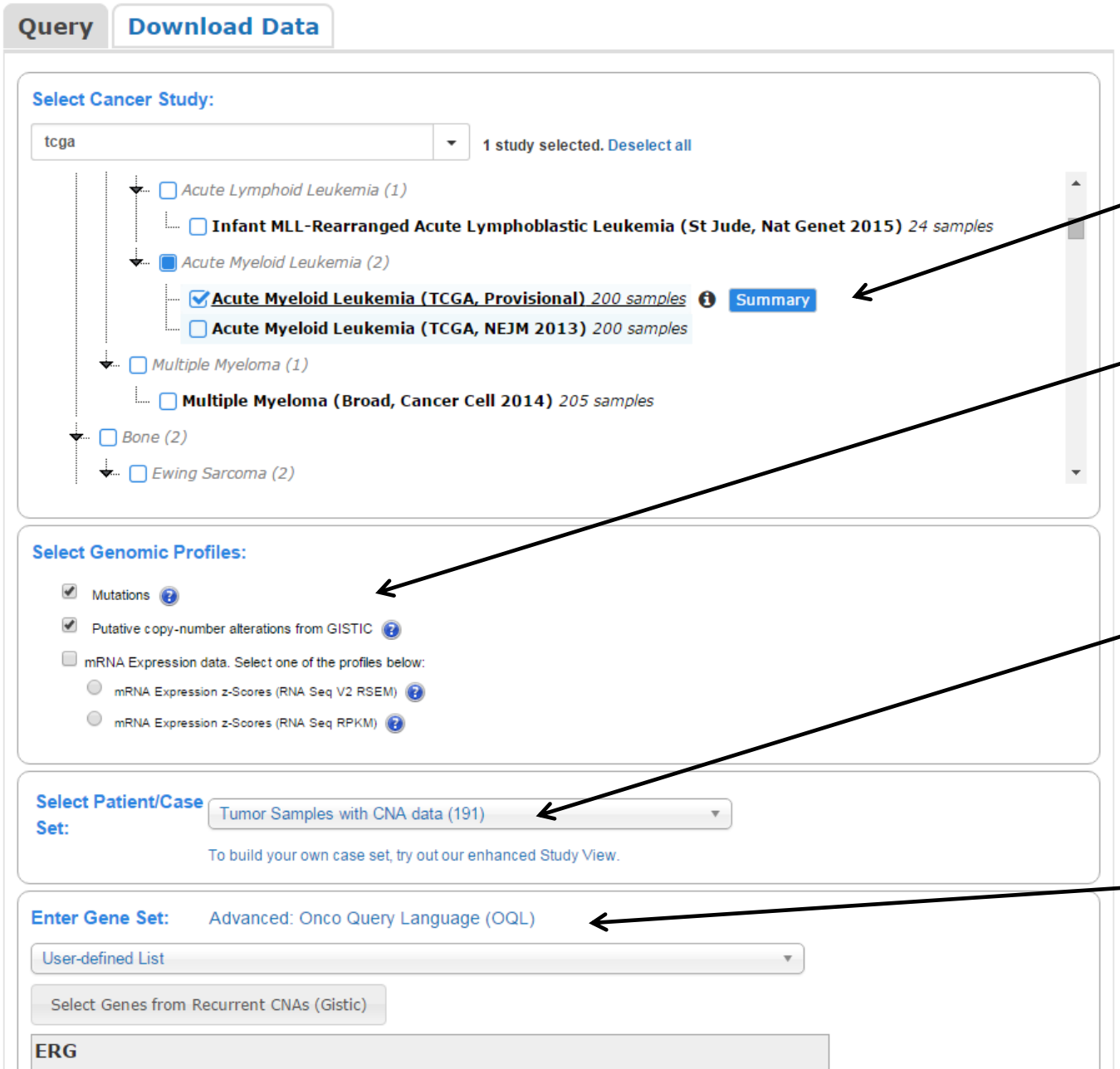
Patient view of an endometrial cancer case

What People are Saying

"I want to thank you for the nice, useful and user-friendly interface you have generated and shared with the

UNSW THE UNIVERSITY OF NEW SOUTH WALES

In this query, we are telling cBioPortal to perform an analyse comparing all AML samples with ERG mutation or CNA and those without ERG mutation nor CNA.



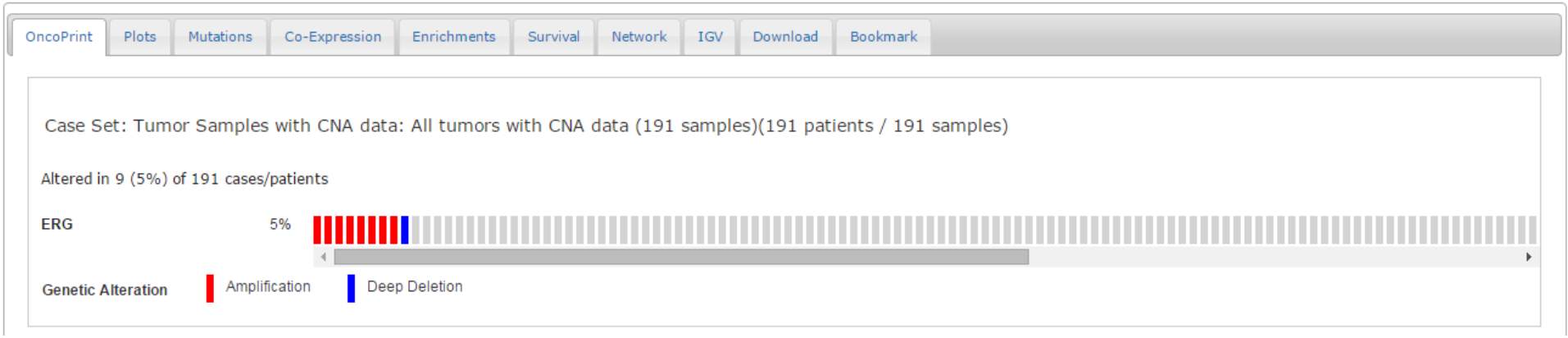
Select cancer study (AML, Provisional)

Select the type of aberration you are interested in (Mutations & CNA)

Select the sample set (Tumour samples with CAN data)

Type in gene - can accept any number. (For this example, we will look at ERG)

OncoPrint

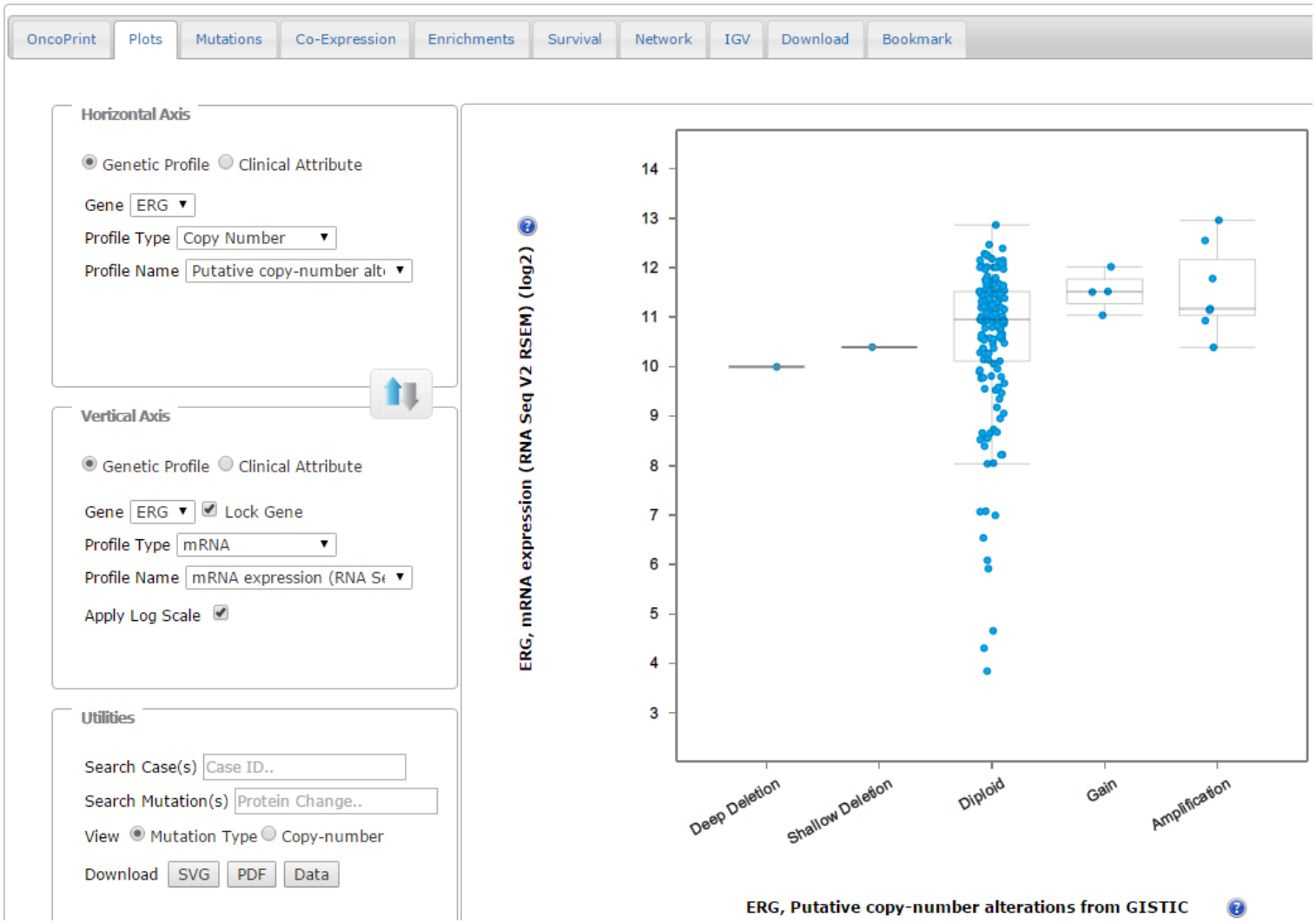


9 out of 191 samples have alteration in ERG:

- 8 samples have amplifications of ERG
- 1 sample has a deep deletion of ERG

Plots – correlation ERG expression with CNA

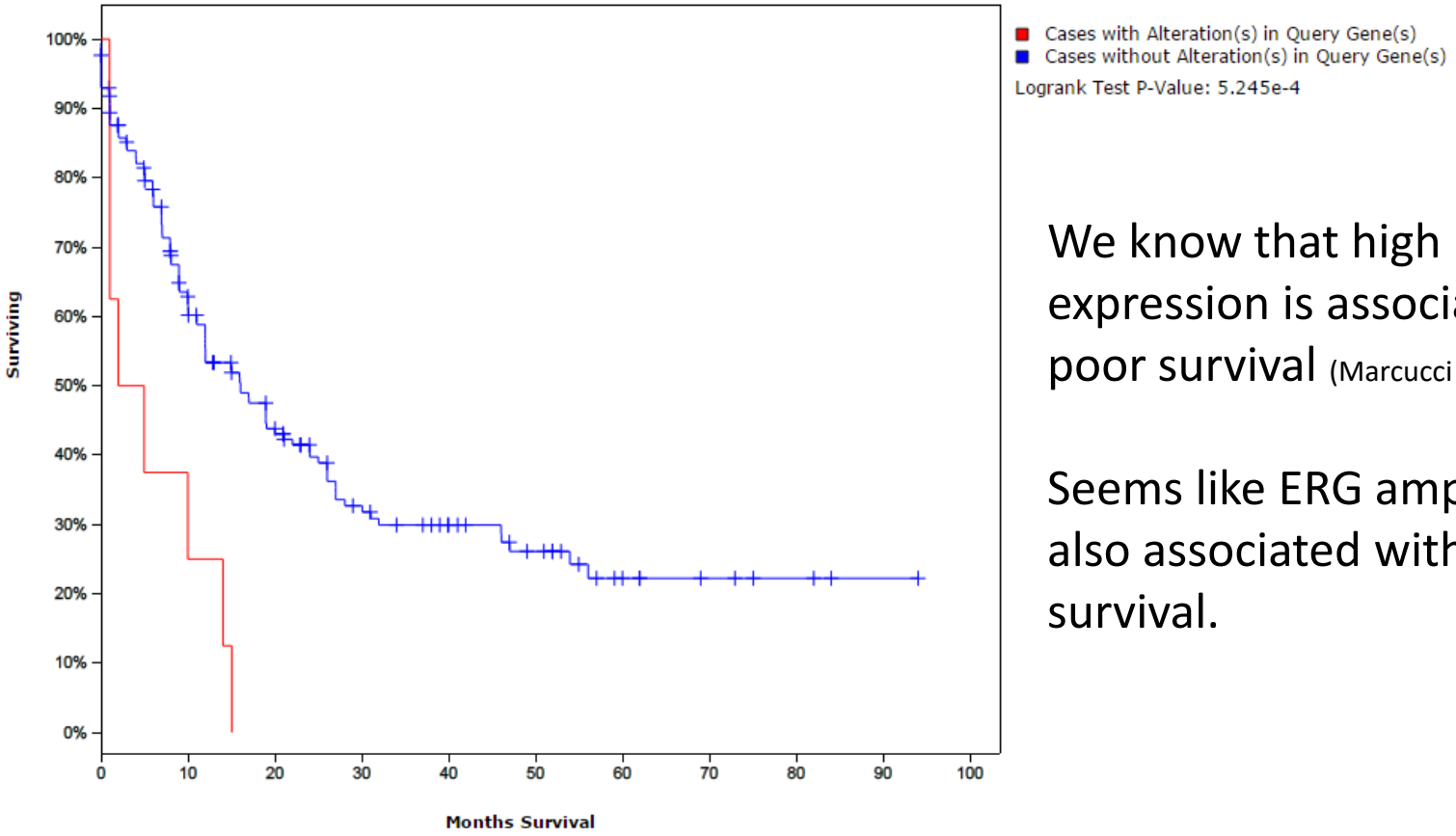
Samples with amplification possibly have higher expression



Survival analysis

OncoPrint | Plots | Mutations | Co-Expression | Enrichments | Survival | Network | IGV | Download | Bookmark

Overall Survival Kaplan-Meier Estimate SVG PDF Data

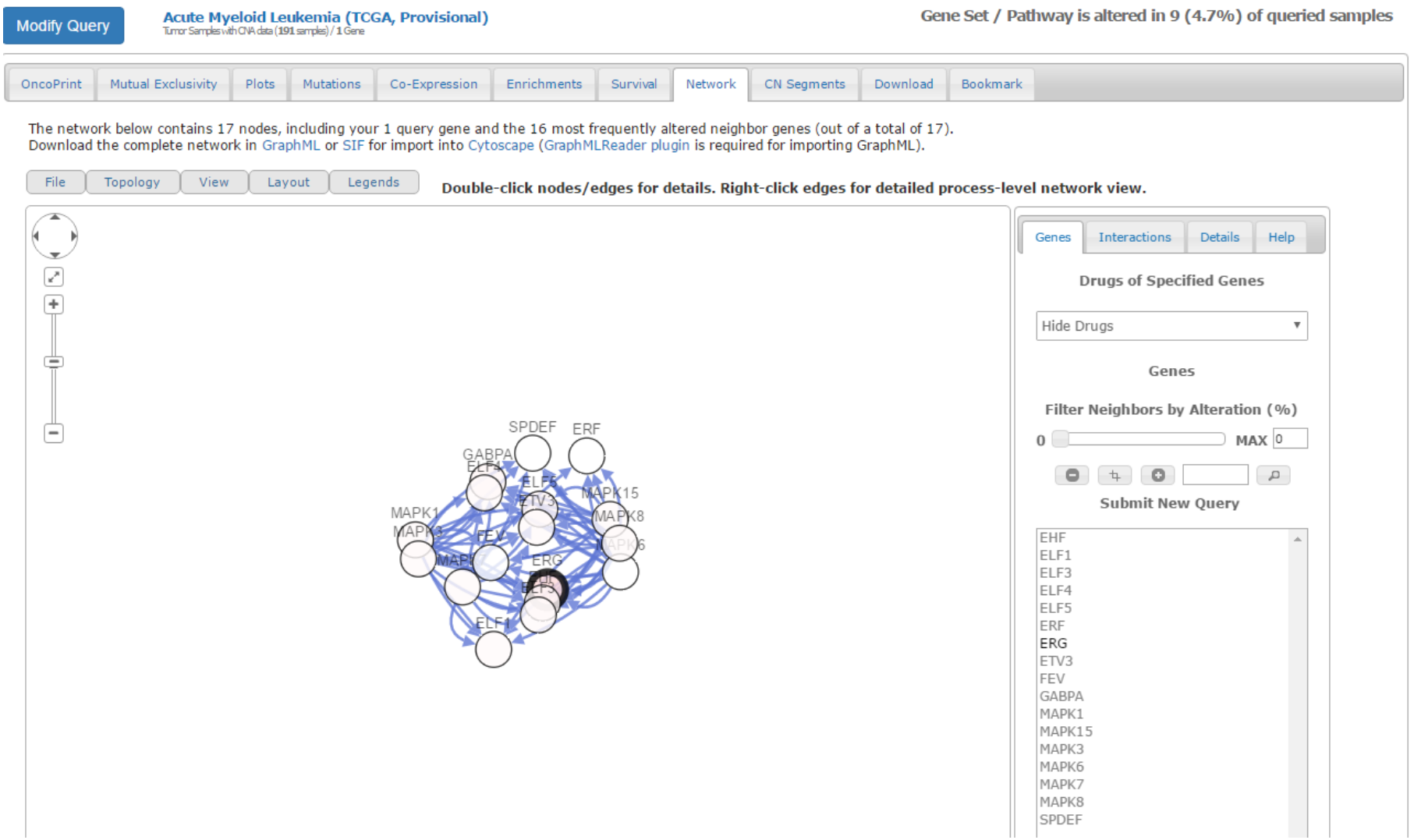


	#total cases	#cases deceased	median months survival
Cases with Alteration(s) in Query Gene(s)	8	8	2
Cases without Alteration(s) in Query Gene(s)	171	108	16.06

We know that high ERG expression is associated with poor survival (Marcucci *et al* JCO 2005).

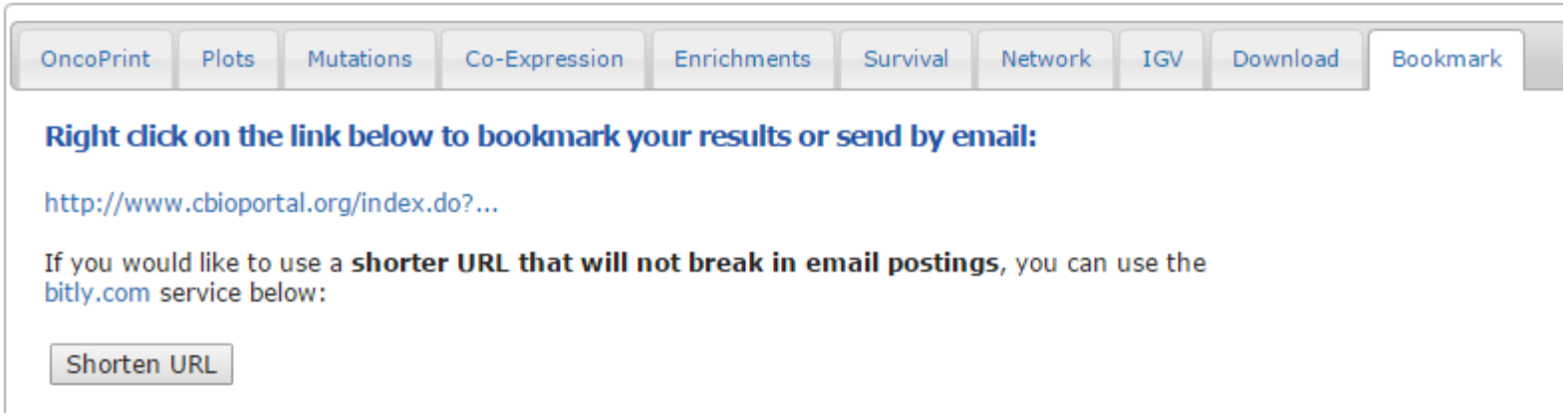
Seems like ERG amplification is also associated with poor survival.

Network analysis



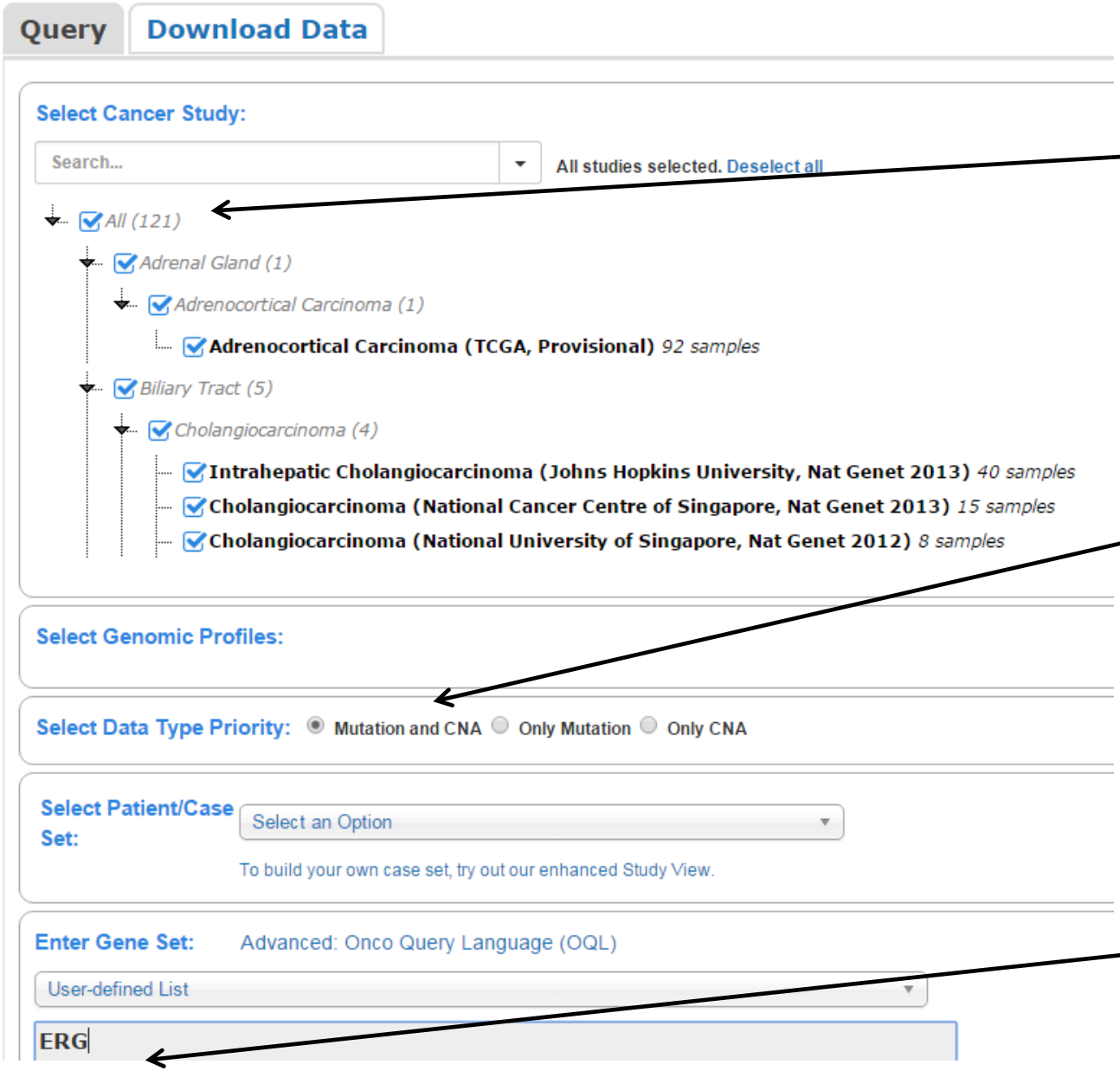
This network analysis is not that interesting, but it could be more useful with a larger input gene set.

Bookmark



You can make a URL to immediately share analysis with collaborators

Gene summaries across cancer types

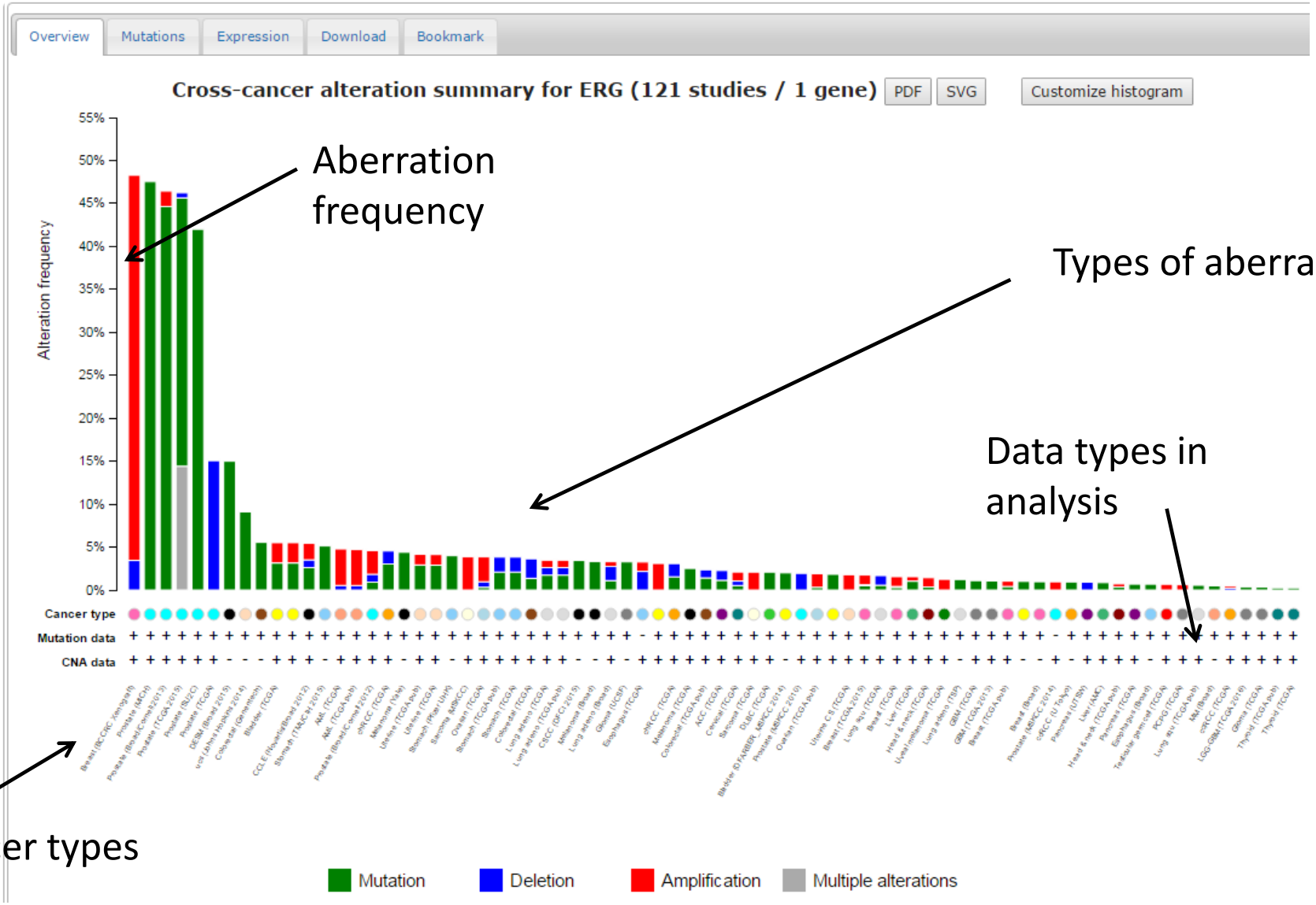


Select all cancers

Select the type of aberration you are interested in (Mutations & CNA)

Type in gene - can accept any number. (For this example, we will look at ERG)

Gene summaries across cancer types



Cancer Sequencing Projects

International Cancer Genome Consortium (ICGC)

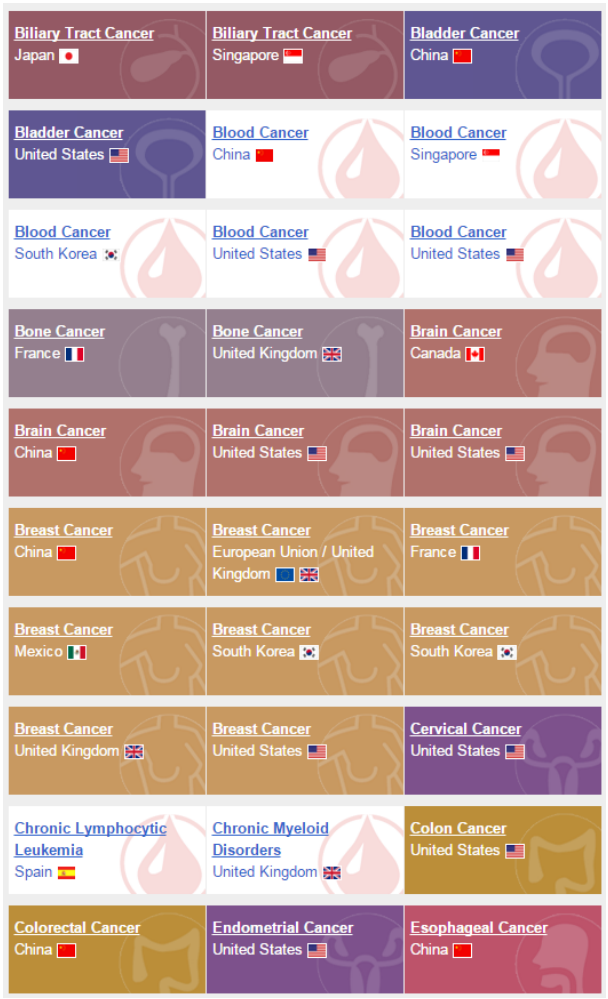
International Cancer Genome Consortium

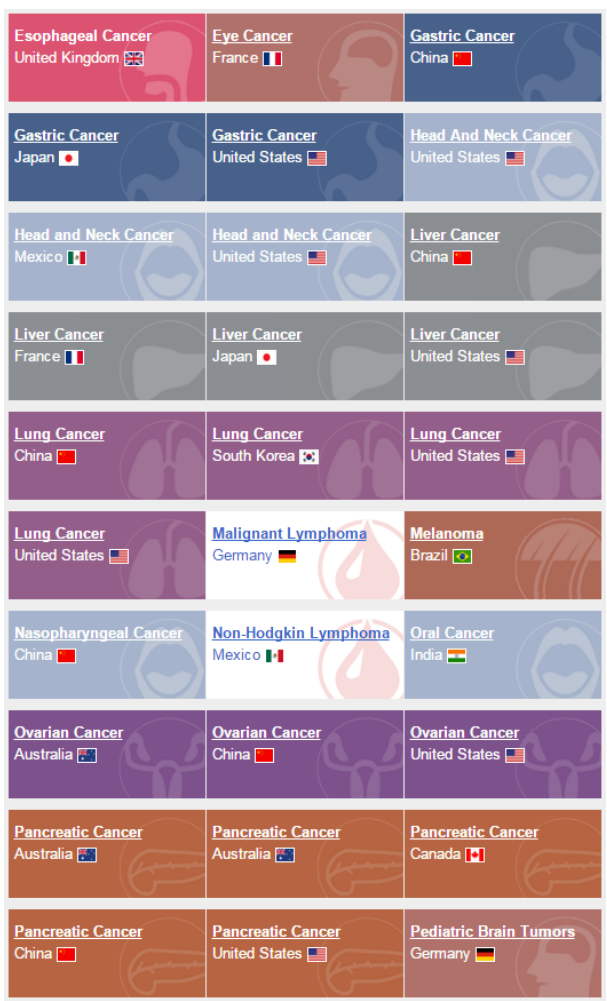
- Collaboration between 22 countries
- Initiated in 2007
- **Aim:**
  To catalogue genomic abnormalities in tumours from 50 different cancer types & subtypes
- **Update today:**
  70 projects, 21 primary sites, >16,246 tumour DNA data
- Uses data from TCGA and the Sanger Cancer Genome Project

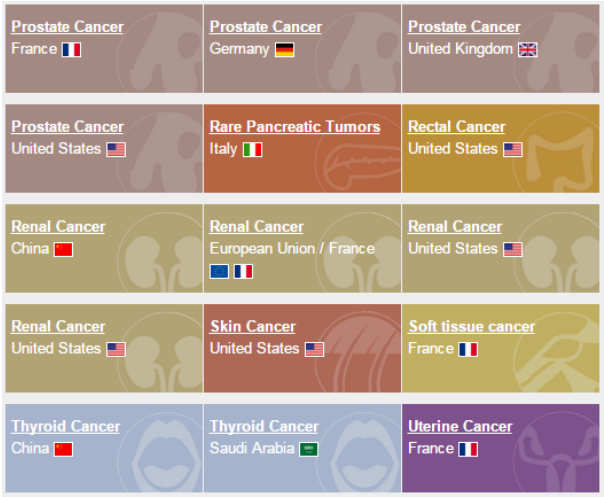
Working groups

INTERNATIONAL CANCER GENOME CONSORTIUM (ICGC) WORKING GROUPS						
Clinical and Pathology Issues	**Quality Standards of Samples**	**Genome Analyses**	**Informed Consent and Privacy Protections**	**Sample Size/Study Design**	**Data Management/ Databases and Coordination**	**Data Release, Data Tiers, Intellectual Property, and Publications**
Lynda Chin Jean-Yves Blay William Dalton Tony Green Stan Hamilton Timothy Ley Ed Liu Paul Mischel Kenneth Pienta Rajiv Sarin Daniel Tan	**Peter Lichter** Carolyn Compton Andy Futreal Youyong Lu Miguel Angel Piris	**Mike Stratton** Olli Kallioniemi Ed Liu Marco Marra John McPherson Brad Ozenberger Henk Stunnenberg Daniel Tan Brandon Wainwright Rick Wilson	**Bartha Knoppers** Martin Bobrow Wylie Burke Kazuto Kato Karen Kennedy Brad Ozenberger Daniel Tan Susan Wallace Henry Yang	**Eric Lander** Ron DePinho Doug Easton Gaddy Getz Partha P. Majumder	**Lincoln Stein** Cameron Brennan Arul Chinnaiyan Peter Good Joe Gray J Gowrishankar David Haussler David Housman Tim Hubbard Subha Madhaven Paul Spellman	**Mark Guyer** Daniela Gerhard Karen Kennedy Brad Ozenberger

ICGC Samples



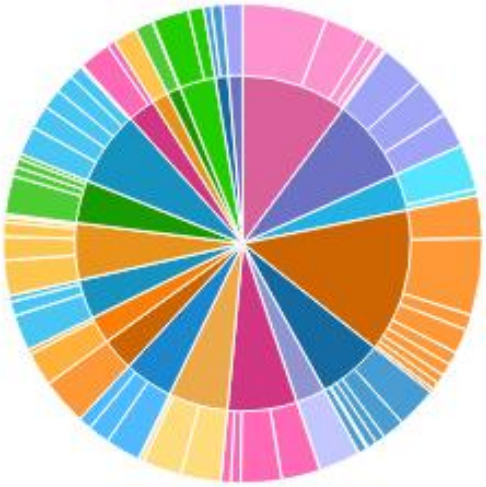




Data Release 23

Dec 7, 2016

Donor Distribution by Primary Site



Cancer projects	70
Cancer primary sites	21
Donors with molecular data in DCC	16,246
Total Donors	19,305
Simple somatic mutations	46,693,172
Mutated Genes	57,658

Data types

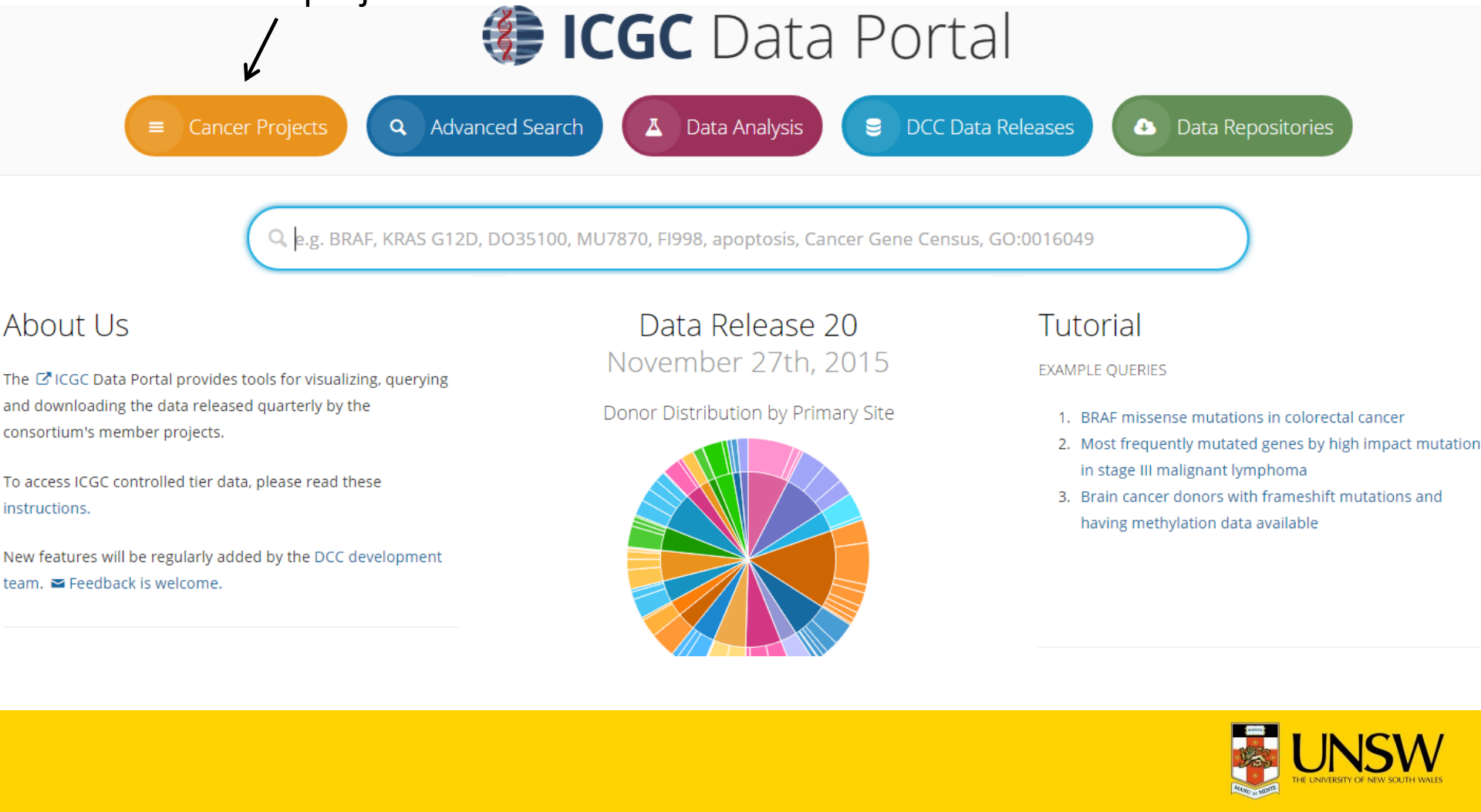
- Mandatory:
  - Genomic DNA analyses of tumors (and matching control DNA) are core elements of the project.
- Complementary (Recommended):
  - Additional studies of DNA methylation and RNA expression are recommended on the same samples that are used to find somatic mutations.
- Optional:
  - Proteomic analyses
  - Metabolomic analyses
  - Immunohistochemical analyses

Data access policy

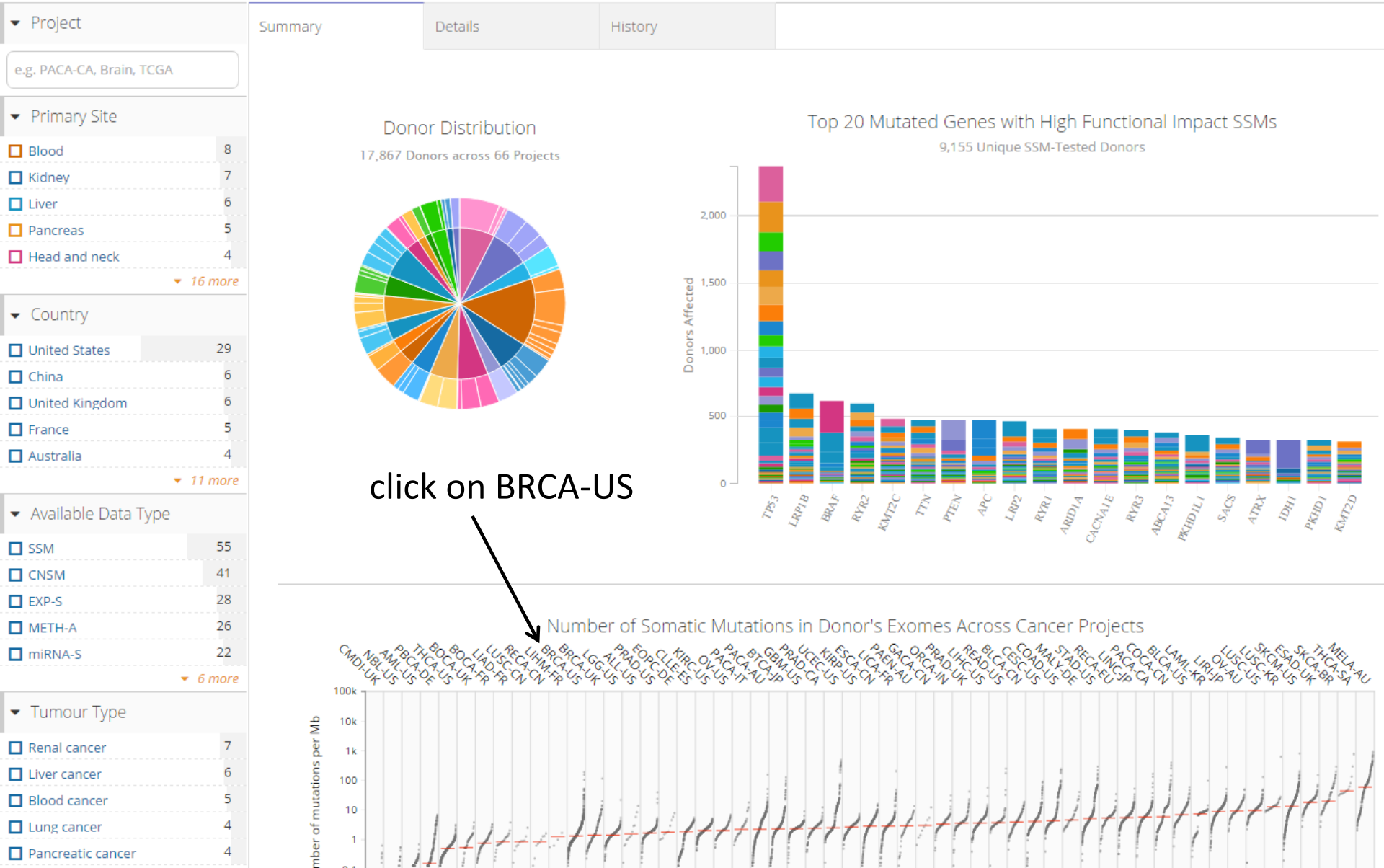
ICGC Open Access Datasets	ICGC Controlled Access Datasets
➢ Cancer Pathology Histologic type or subtype Histologic nuclear grade ➢ Patient/Person Gender Age range ➢ Gene Expression (normalized) ➢ DNA methylation ➢ Genotype frequencies ➢ Computed Copy Number and Loss of Heterozygosity ➢ Newly discovered somatic variants	➢ Detailed Phenotype and Outcome Data Patient demography Risk factors Examination Surgery/Drugs/Radiation Sample/Slide Specific histological features Protocol Analyte/Aliquot ➢ Gene Expression (probe-level data) ➢ Raw genotype calls ➢ Gene-sample identifier links ➢ Genome sequence files

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

Click on cancer projects



Cancer project view



BREAST CANCER - TCGA, US

Page Filters

Mutation Impact

- High
- Low
- Unknown

Summary

Code	BRCA-US
Name	Breast Cancer - TCGA, US
Primary Site	Breast
Tumour Type	Breast cancer
Tumour Subtype	Ductal & lobular
Countries	United States
Number of donors in PCAWG	92
Number of donors with molecular data in DCC	1,045
Total number of donors	1,099

Experimental Analyses

non-NGS	2107 samples from 1045 donors
WGS	19 samples from 19 donors
WXS	970 samples from 955 donors
RNA-Seq	1177 samples from 1041 donors
miRNA-Seq	1153 samples from 1026 donors

Download Sample Sheet

Raw data is available at

The Cancer Genome Atlas
Cancer Genomics Hub

An approved data access request is required.

Available Data Types

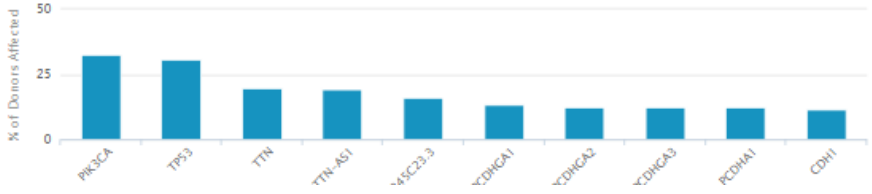
Clinical Data	1,099 donors
Simple Somatic Mutations (SSM)	955 donors
Copy Number Somatic Mutations (CNSM)	1,045 donors
Structural Somatic Mutations (StSM)	--
Simple Germline Variants (SGV)	--
Array-based DNA Methylation (METH-A)	1,013 donors
Sequence-based DNA Methylation (METH-S)	--
Array-based Gene Expression (EXP-A)	529 donors
Sequence-based Gene Expression (EXP-S)	1,041 donors
Protein Expression (PEXP)	298 donors
Sequence-based miRNA Expression (miRNA)	1,026 donors
Exon junction (JCN)	--

OPEN IN DCC Data Releases

Click on Genome Viewer

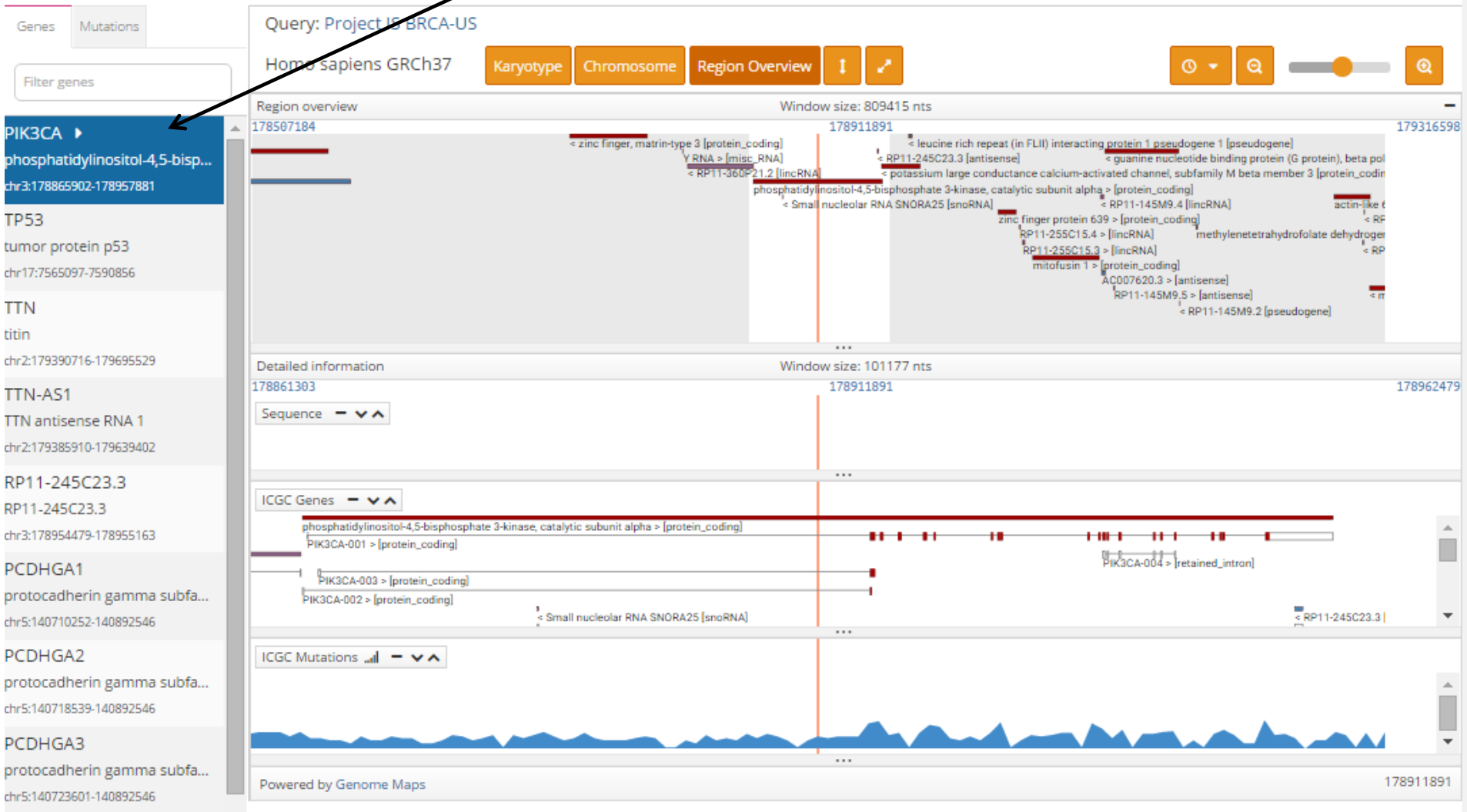
Most Frequently Mutated Genes

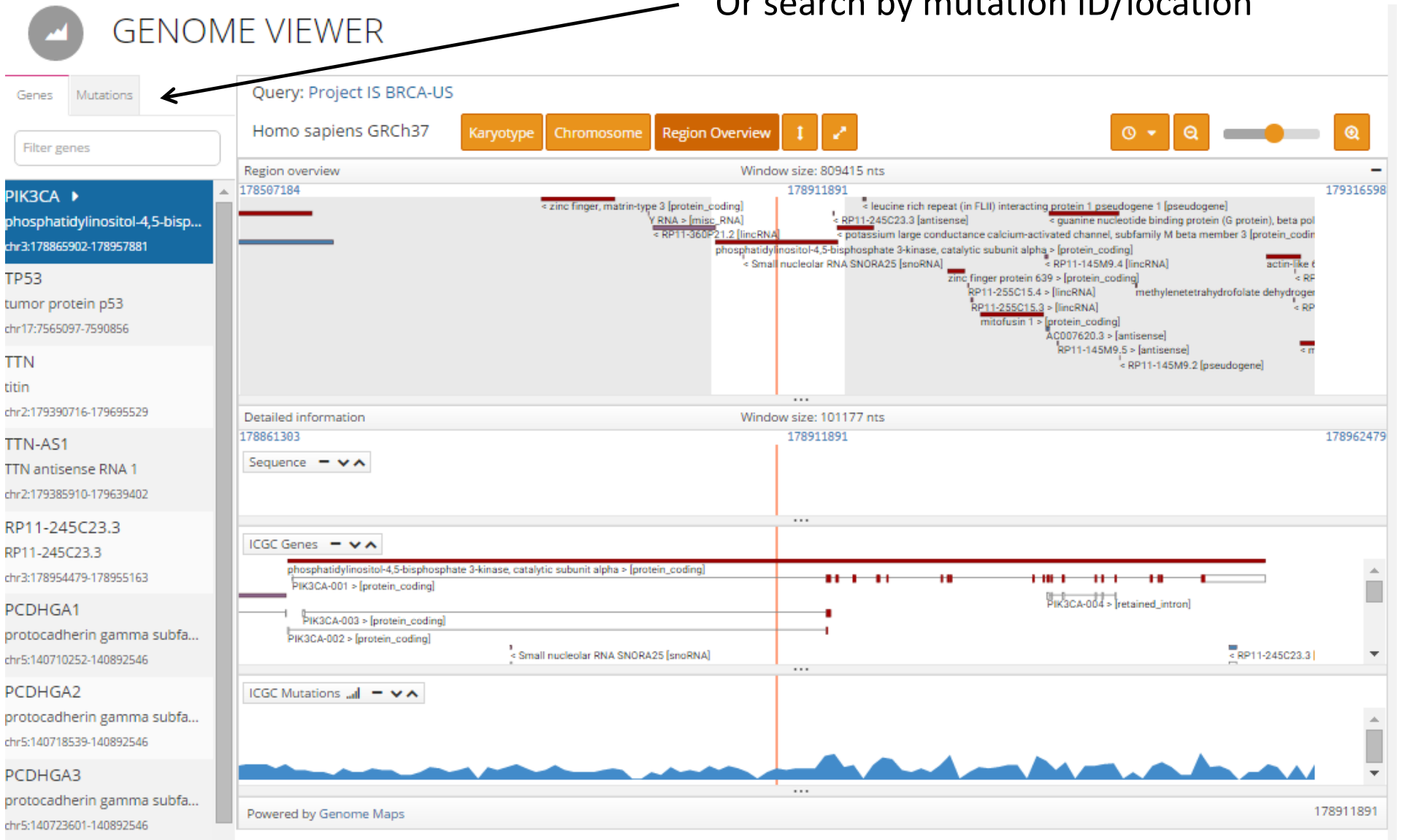
OPEN IN ADVANCED SEARCH | GENOME VIEWER



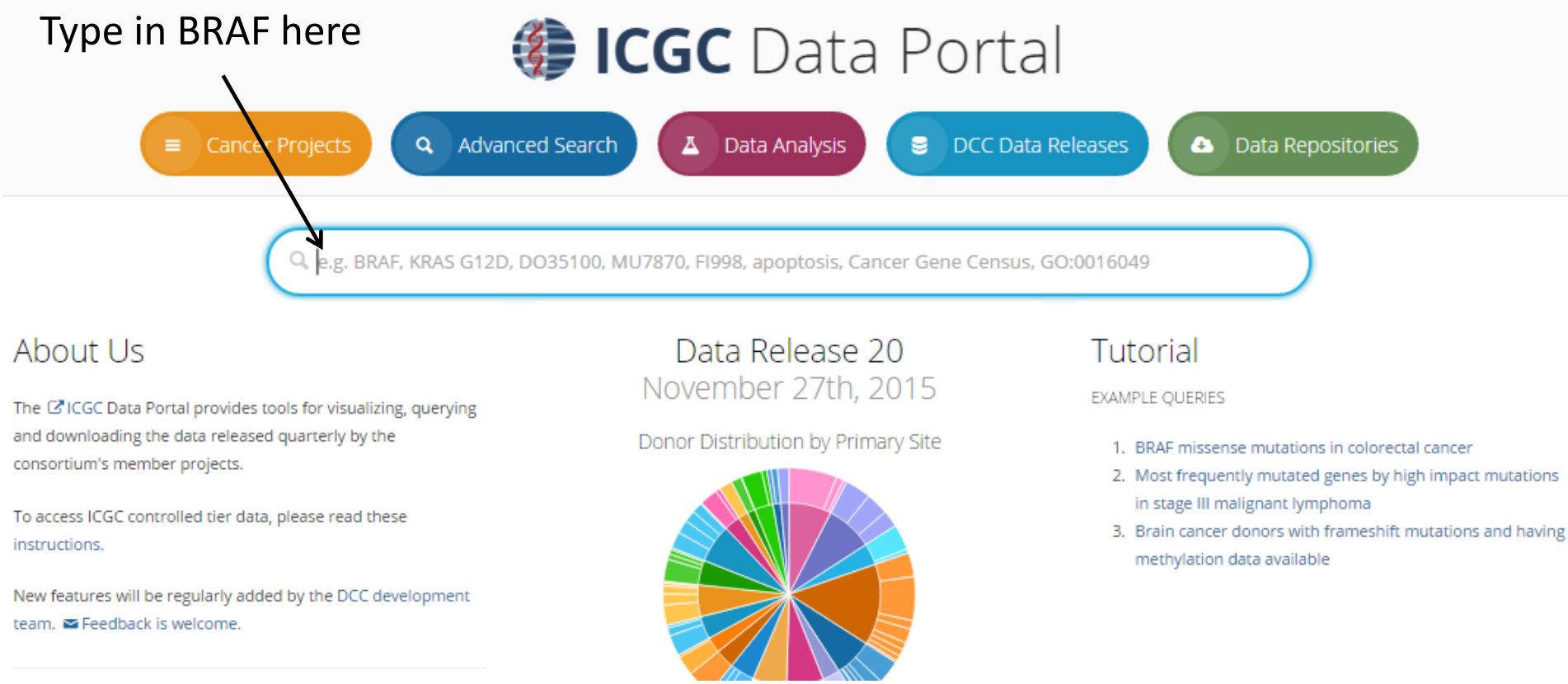
GENOME VIEWER

View top mutated genes

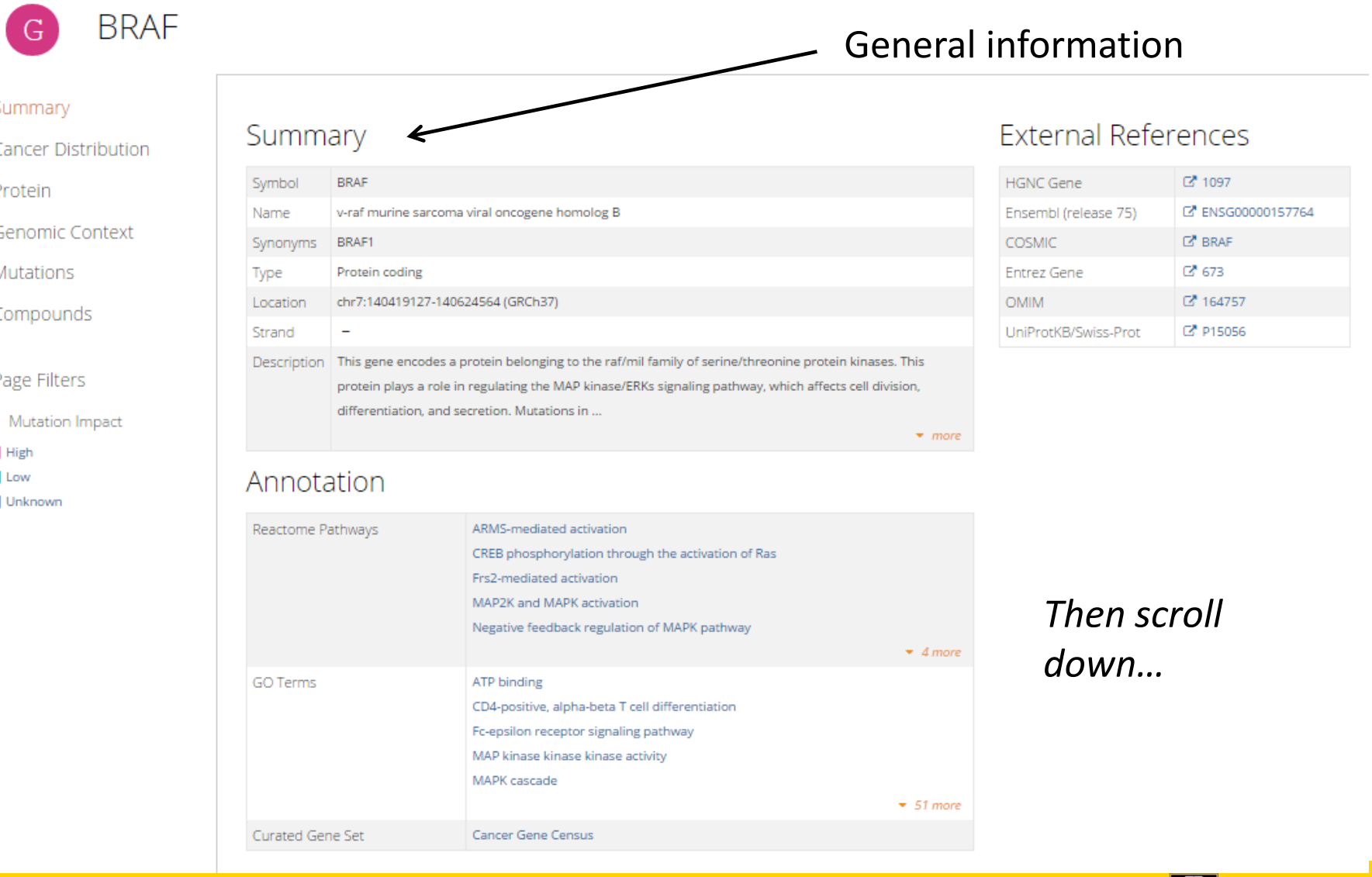



GENOME VIEWER
Or search by mutation ID/location
Genes
Mutations
Filter genes
PIK3CA
phosphatidylinositol-4,5-bisp...
chr3:178865902-178957881
TP53
tumor protein p53
chr17:7565097-7590856
TTN
titin
chr2:179390716-179695529
TTN-AS1
TTN antisense RNA 1
chr2:179385910-179639402
RP11-245C23.3
RP11-245C23.3
chr3:178954479-178955163
PCDHGA1
protocadherin gamma subfa...
chr5:140710252-140892546
PCDHGA2
protocadherin gamma subfa...
chr5:140718539-140892546
PCDHGA3
protocadherin gamma subfa...
chr5:140723601-140892546
Query: Project IS BRCA-US
Homo sapiens GRCh37
Karyotype
Chromosome
Region Overview
Region overview
Window size: 809415 nts
178507184
178911891
179316598
Detailed information
Window size: 101177 nts
178861303
178911891
178962479
Sequence
ICGC Genes
phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha > [protein_coding]
PIK3CA-001 > [protein_coding]
PIK3CA-004 > [retained_intron]
PIK3CA-003 > [protein_coding]
PIK3CA-002 > [protein_coding]
< Small nucleolar RNA SNORA25 [snoRNA]
ICGC Mutations
Powered by Genome Maps
178911891

Looking at mutations in specific genes



Gene centric view



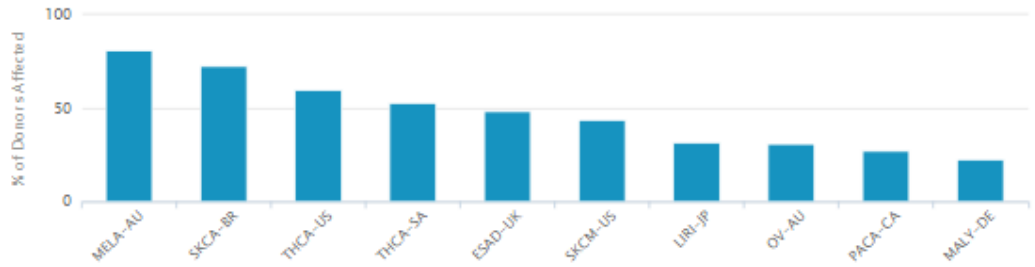
Then scroll down…

Gene centric view

Cancer Distribution

OPEN IN ADVANCED SEARCH

1,144 DONORS AFFECTED BY **1,244** MUTATIONS ACROSS **43** PROJECTS



Types of cancers with the mutation

Showing **43** projects

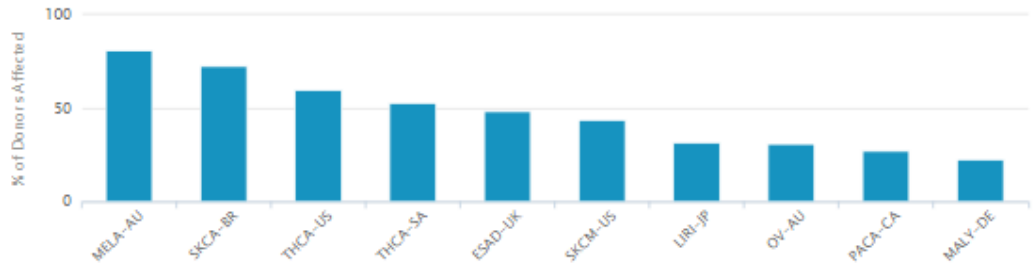
Project	Site	Tumour Type	Tumour Subtype	# Donors affected ▾	# Mutations
MELA-AU	Skin	Skin cancer	Melanoma	149 / 183 *(81.42%)*	501
SKCA-BR	Skin	Melanoma		48 / 66 *(72.73%)*	81
THCA-US	Head and neck	Head and Neck cancer	Thyroid carcinoma	238 / 400 *(59.50%)*	5
THCA-SA	Head and neck	Thyroid cancer	Papillary thyroid carcinoma	8 / 15 *(53.33%)*	2
ESAD-UK	Esophagus	Esophageal cancer	Esophageal adenocarcinoma	58 / 119 *(48.74%)*	93
SKCM-US	Skin	Skin cancer	Cutaneous melanoma	148 / 335 *(44.18%)*	22
LIRI-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	83 / 260 *(31.92%)*	131
OV-AU	Ovary	Ovarian cancer	Serous cystadenocarcinoma	29 / 93 *(31.18%)*	46
PACA-CA	Pancreas	Pancreatic cancer	Ductal adenocarcinoma	55 / 204 *(26.96%)*	75
MALY-DE	Blood	Malignant Lymphoma	Germinal center B-cell derived lymphomas	10 / 44 *(22.73%)*	19
PRAD-UK	Prostate	Prostate cancer	Adenocarcinoma	21 / 108 *(19.44%)*	21
RECA-EU	Kidney	Renal cancer	Renal cell carcinoma (Focus on but not limited to clear cell subtype)	18 / 95 *(18.95%)*	23
COAD-US	Colorectal	Colon cancer	Adenocarcinoma	33 / 216 *(15.28%)*	11
LINC-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	30 / 244 *(12.30%)*	34

Gene centric view

Cancer Distribution

OPEN IN ADVANCED SEARCH

1,144 DONORS AFFECTED BY **1,244** MUTATIONS ACROSS **43** PROJECTS



More detailed information

Showing **43** projects

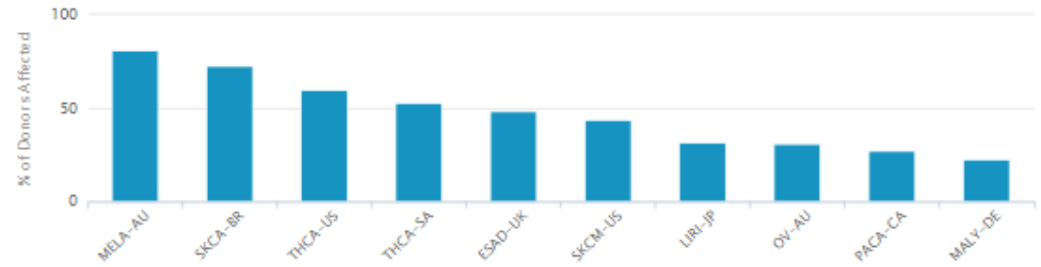
Project	Site	Tumour Type	Tumour Subtype	# Donors affected ▾	# Mutations
MELA-AU	Skin	Skin cancer	Melanoma	149 / 183 *(81.42%)*	501
SKCA-BR	Skin	Melanoma		48 / 66 *(72.73%)*	81
THCA-US	Head and neck	Head and Neck cancer	Thyroid carcinoma	238 / 400 *(59.50%)*	5
THCA-SA	Head and neck	Thyroid cancer	Papillary thyroid carcinoma	8 / 15 *(53.33%)*	2
ESAD-UK	Esophagus	Esophageal cancer	Esophageal adenocarcinoma	58 / 119 *(48.74%)*	93
SKCM-US	Skin	Skin cancer	Cutaneous melanoma	148 / 335 *(44.18%)*	22
LIRI-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	83 / 260 *(31.92%)*	131
OV-AU	Ovary	Ovarian cancer	Serous cystadenocarcinoma	29 / 93 *(31.18%)*	46
PACA-CA	Pancreas	Pancreatic cancer	Ductal adenocarcinoma	55 / 204 *(26.96%)*	75
MALY-DE	Blood	Malignant Lymphoma	Germinal center B-cell derived lymphomas	10 / 44 *(22.73%)*	19
PRAD-UK	Prostate	Prostate cancer	Adenocarcinoma	21 / 108 *(19.44%)*	21
RECA-EU	Kidney	Renal cancer	Renal cell carcinoma (Focus on but not limited to clear cell subtype)	18 / 95 *(18.95%)*	23
COAD-US	Colorectal	Colon cancer	Adenocarcinoma	33 / 216 *(15.28%)*	11
LINC-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	30 / 244 *(12.30%)*	34

Gene centric view

Cancer Distribution

1,144 DONORS AFFECTED BY **1,244** MUTATIONS ACROSS **43** PROJECTS

OPEN IN ADVANCED SEARCH

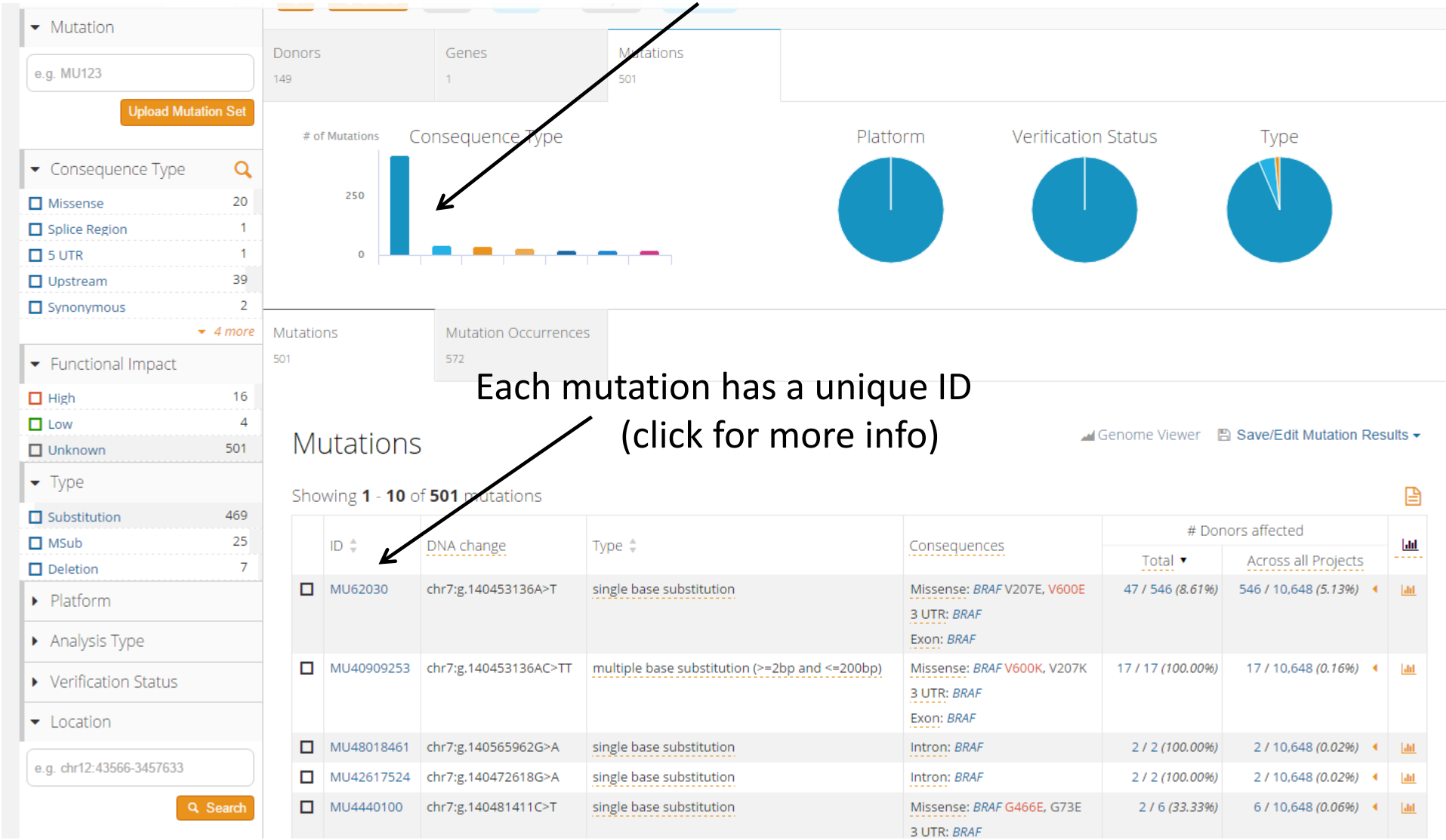


Click for more detail on the mutations

Showing **43** projects

Project	Site	Tumour Type	Tumour Subtype	# Donors affected ▾	# Mutations
MELA-AU	Skin	Skin cancer	Melanoma	149 / 183 *(81.42%)*	501
SKCA-BR	Skin	Melanoma		48 / 66 *(72.73%)*	81
THCA-US	Head and neck	Head and Neck cancer	Thyroid carcinoma	238 / 400 *(59.50%)*	5
THCA-SA	Head and neck	Thyroid cancer	Papillary thyroid carcinoma	8 / 15 *(53.33%)*	2
ESAD-UK	Esophagus	Esophageal cancer	Esophageal adenocarcinoma	58 / 119 *(48.74%)*	93
SKCM-US	Skin	Skin cancer	Cutaneous melanoma	148 / 335 *(44.18%)*	22
LIRI-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	83 / 260 *(31.92%)*	131
OV-AU	Ovary	Ovarian cancer	Serous cystadenocarcinoma	29 / 93 *(31.18%)*	46
PACA-CA	Pancreas	Pancreatic cancer	Ductal adenocarcinoma	55 / 204 *(26.96%)*	75
MALY-DE	Blood	Malignant Lymphoma	Germinal center B-cell derived lymphomas	10 / 44 *(22.73%)*	19
PRAD-UK	Prostate	Prostate cancer	Adenocarcinoma	21 / 108 *(19.44%)*	21
RECA-EU	Kidney	Renal cancer	Renal cell carcinoma (Focus on but not limited to clear cell subtype)	18 / 95 *(18.95%)*	23
COAD-US	Colorectal	Colon cancer	Adenocarcinoma	33 / 216 *(15.28%)*	11
LINC-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	30 / 244 *(12.30%)*	34

Hover over the section of the graph to see what region it represents



Each mutation has a unique ID (click for more info)

Mutations

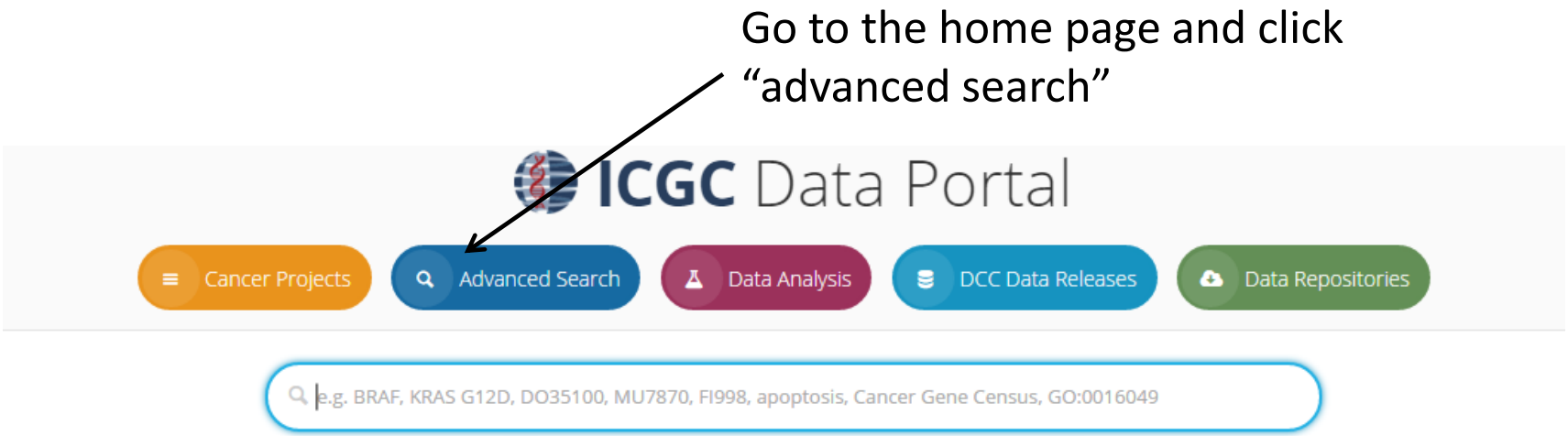
Genome Viewer | Save/Edit Mutation Results

Showing **1** - **10** of **501** mutations

	ID	DNA change	Type	Consequences	# Donors affected	
					Total	Across all Projects
☐	MU62030	chr7:g.140453136A>T	single base substitution	Missense: *BRAF* V207E, V600E 3 UTR: *BRAF* Exon: *BRAF*	47 / 546 *(8.61%)*	546 / 10,648 *(5.13%)*
☐	MU40909253	chr7:g.140453136AC>TT	multiple base substitution (>=2bp and <=200bp)	Missense: *BRAF* V600K, V207K 3 UTR: *BRAF* Exon: *BRAF*	17 / 17 *(100.00%)*	17 / 10,648 *(0.16%)*
☐	MU48018461	chr7:g.140565962G>A	single base substitution	Intron: *BRAF*	2 / 2 *(100.00%)*	2 / 10,648 *(0.02%)*
☐	MU42617524	chr7:g.140472618G>A	single base substitution	Intron: *BRAF*	2 / 2 *(100.00%)*	2 / 10,648 *(0.02%)*
☐	MU4440100	chr7:g.140481411C>T	single base substitution	Missense: *BRAF* G466E, G73E 3 UTR: *BRAF*	2 / 6 *(33.33%)*	6 / 10,648 *(0.06%)*

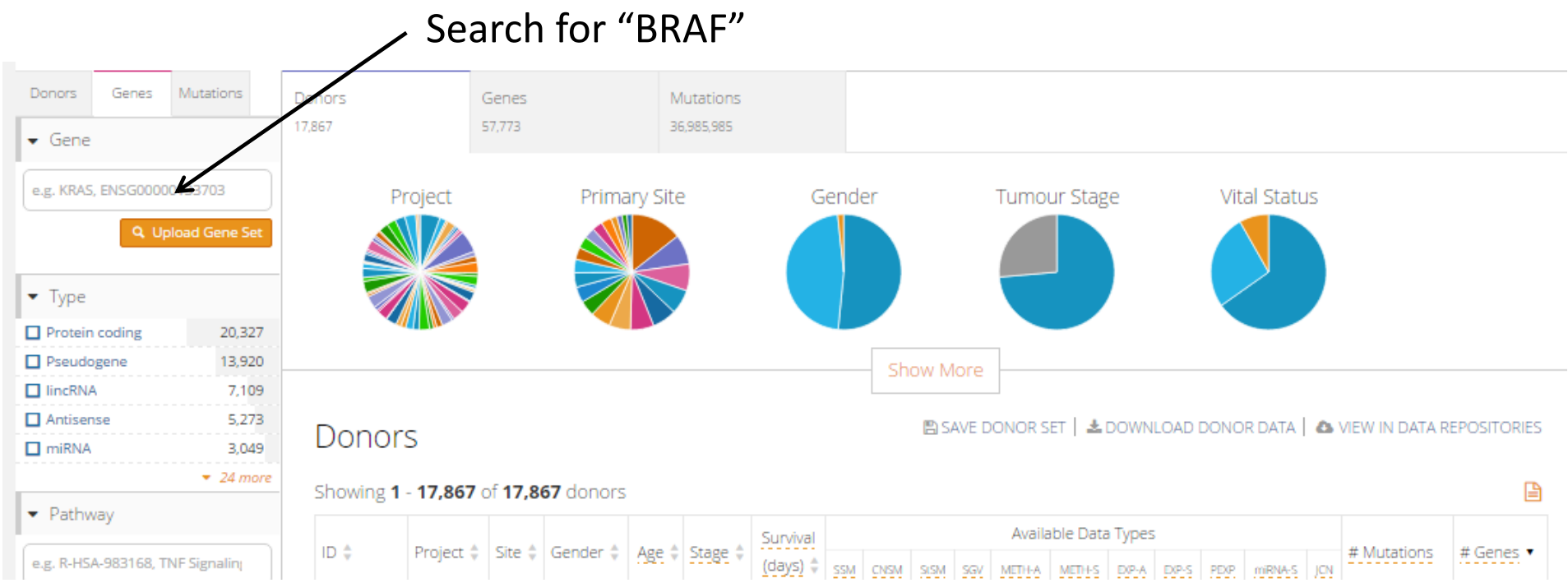
Advanced Search

Find out which cancers commonly have BRAF missense mutations



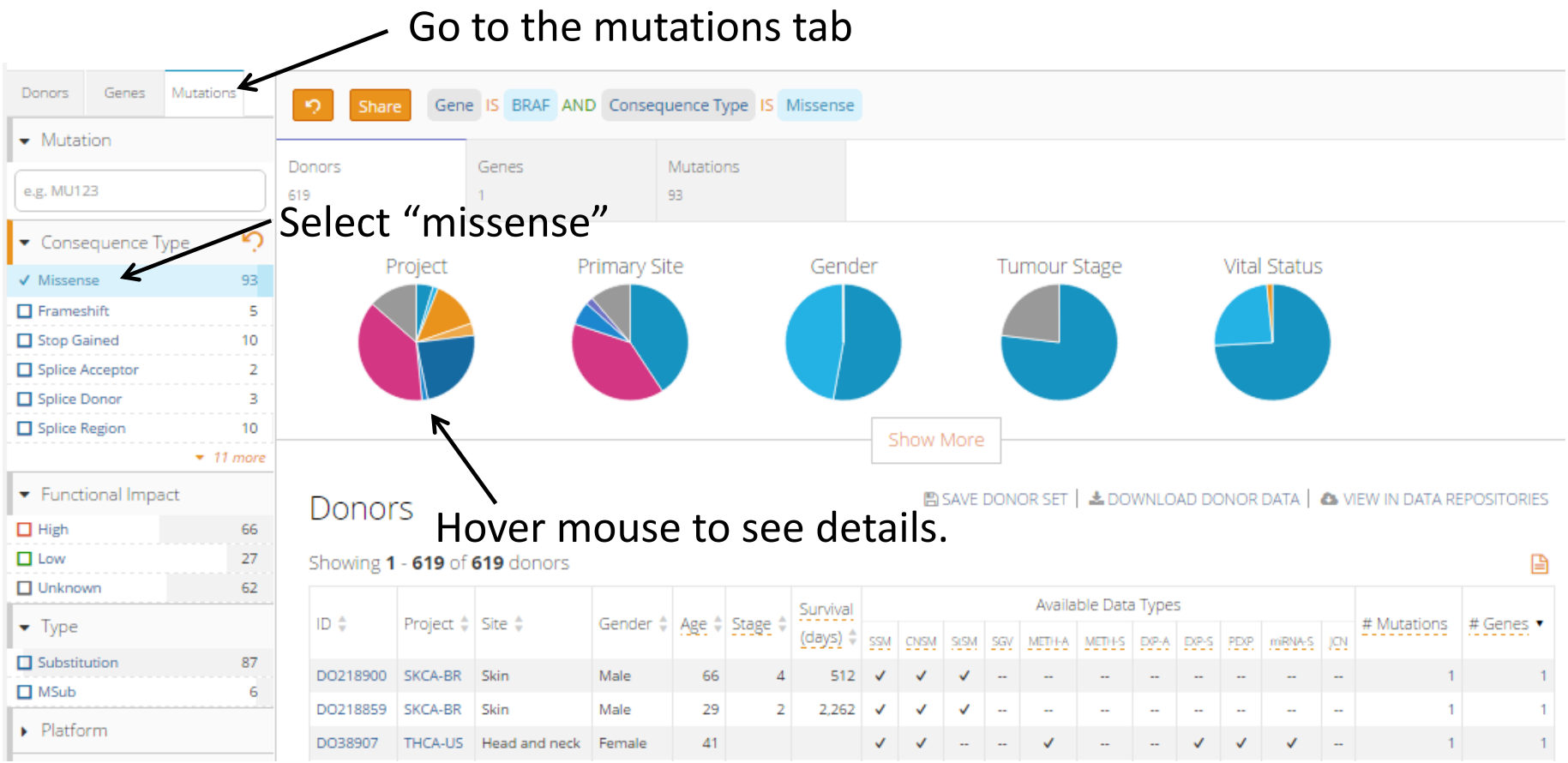
Advanced Search

Find out which cancers commonly have BRAF missense mutations



Advanced Search

Find out which cancers commonly have BRAF missense mutations



Most common cancers with BRAF missense mutations are thyroid cancer and melanoma.

UNSW THE UNIVERSITY OF NEW SOUTH WALES

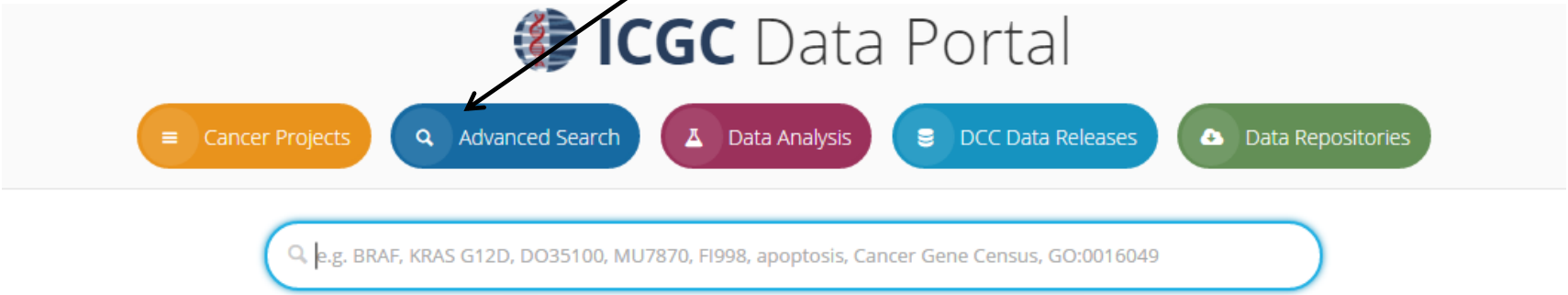
Limitations of data portal

- The data portal is mutation centric
  - i.e. All queries are related to retrieving tumours/samples with particular mutations in a particular gene

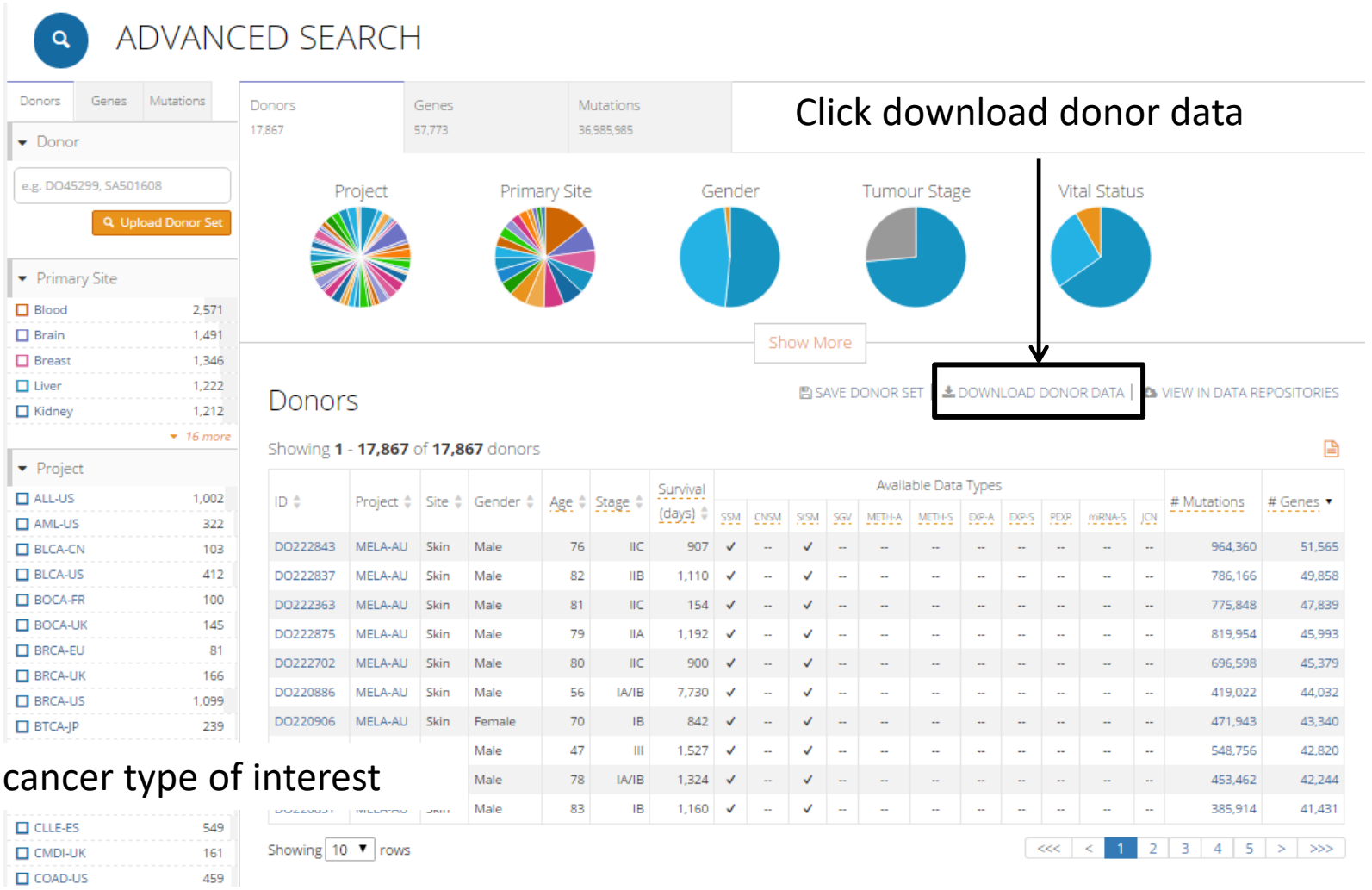
- If you just want expression/methylation data for a particular gene you still have to download the data

Downloading data from ICGC

Go to the home page and click "advanced search"



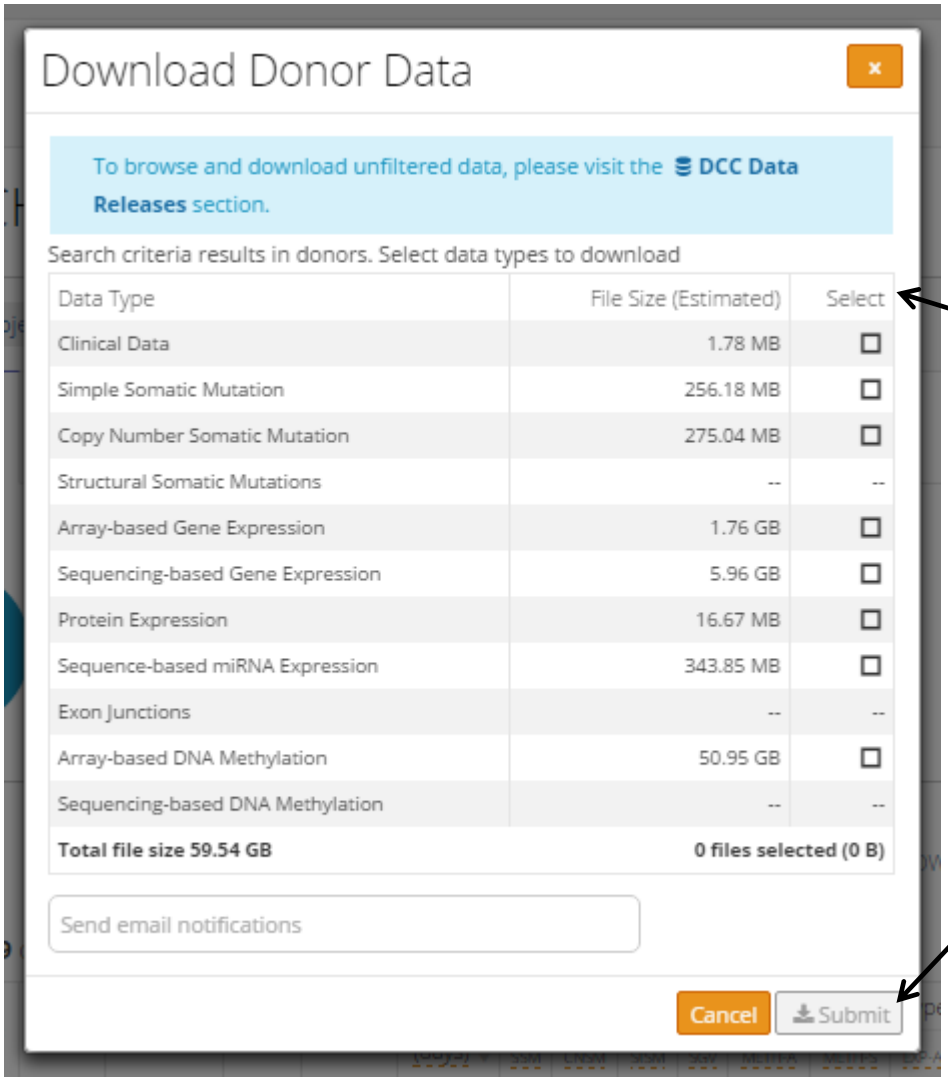
Downloading data from ICGC



Note: go back to Advance search on home page

UNSW THE UNIVERSITY OF NEW SOUTH WALES

Downloading data from ICGC

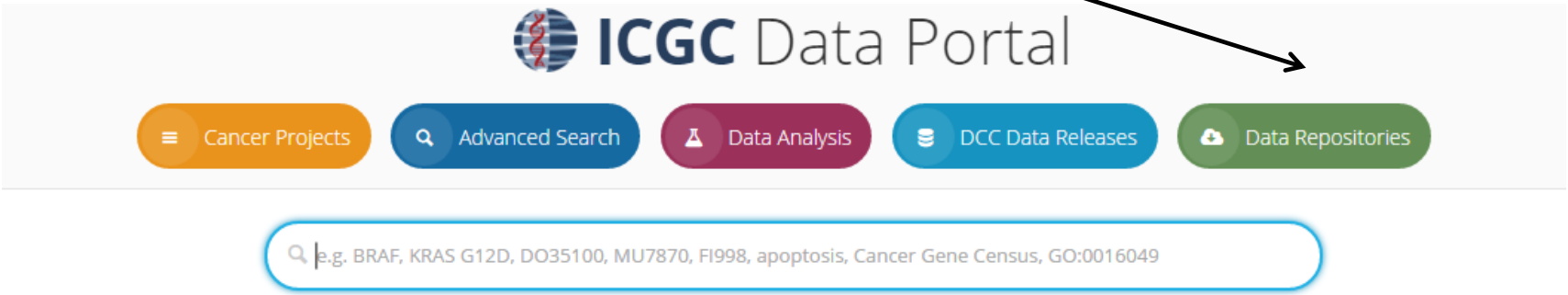


Select the data types of interest

Click "Submit"

Downloading data from ICGC

Or download from the data repository



DATA REPOSITORIES

Click through filters to choose what data you want

The ICGC DCC is periodically aggregating and indexing ICGC data hosted in multiple archives and cloud repositories worldwide. More information here.

Files | Donors

File

Enter File ID, Name or Object ID

Repository

AWS - Virginia	14,897
Collaboratory	8,762
PCAWG - Barcelona	4,820
PCAWG - Chicago (IC...	5,038
PCAWG - Chicago (T...	12,968

7 more

Data Type

Aligned Reads	76,380
SSM	12,912
Clinical	9,491
StSM	7,890
SGV	5,288

3 more

Experimental Strategy

WGS	44,253

Download manifests | **121,167** Files | **11,371** Donors | **1.26 PB**

Then download the data you selected

SAVE DONOR SET

Showing **1** - **25** of **121,167** files

File ID	Donor	Repository	Project	Study	Data Type	Strategy	Format	Size
FI99998	DO15366	CGHub	HNSC-US		Aligned Reads	RNA-Seq	BAM	6.89 GB
FI99996	DO14274	CGHub	HNSC-US		Aligned Reads	RNA-Seq	BAM	3.52 GB
FI99994	DO14462	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	97.27 MB
FI99992	DO14169	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	221.17 MB
FI99990	DO14158	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	435.70 MB
FI99988	DO16061	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	192.24 MB
FI99986	DO14686	CGHub	HNSC-US		Aligned Reads	WGS	BAM	183.91 GB
FI99984	DO14408	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	187.34 MB
FI99982	DO14286	CGHub	HNSC-US		Aligned Reads	RNA-Seq	BAM	7.27 GB
FI99980	DO14246	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	385.00 MB
FI9998	DO217962	PCAWG - Heidelberg	BRCA-EU	PCAWG	CNSM	WGS	VCF	1.83 KB
FI99978	DO14886	CGHub	HNSC-US		Aligned Reads	RNA-Seq	BAM	5.41 GB
FI99976	DO14214	CGHub	HNSC-US		Aligned Reads	RNA-Seq	BAM	6.69 GB
FI99974	DO14218	CGHub	HNSC-US		Aligned Reads	miRNA-Seq	BAM	192.05 MB

The advantage of ICGC is that data for all samples is in a single file so it is easier to work with in Excel (if file is small) or Galaxy (if file is big).

COSMIC database

http://cancer.sanger.ac.uk/cosmic

Select "Cancer Gene Census"

COSMIC
Catalogue of somatic mutations in cancer

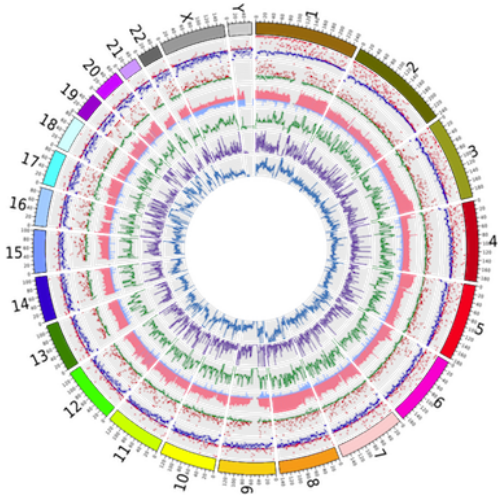
Home ▼ | Resources ▼ | Curation ▼ | Tools ▼ | Data ▼ | News ▼ | Help ▼ | About ▼ | Search COSMIC... | Login ▼

COSMIC v80, released ~~13-FEB-17~~

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below, or by browsing a region of the human genome using the map to the right.

eg: Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH



Browse the genomic landscape of cancer

R Resources

Key COSMIC resources

Cell Lines Project
COSMIC
Cancer Gene Census
Drug Sensitivity
Mutational Signatures
GRCh37 Cancer Archive

T Tools

Additional tools to explore COSMIC

Cancer Browser
Genome Browser
GA4GH Beacon
COSMIC-3D Beta
COSMIC In BigQuery New
CONAN

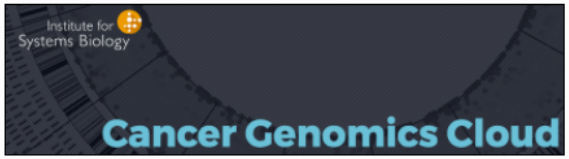
C Expert Curation

High quality curation by expert postdoctoral

D Data

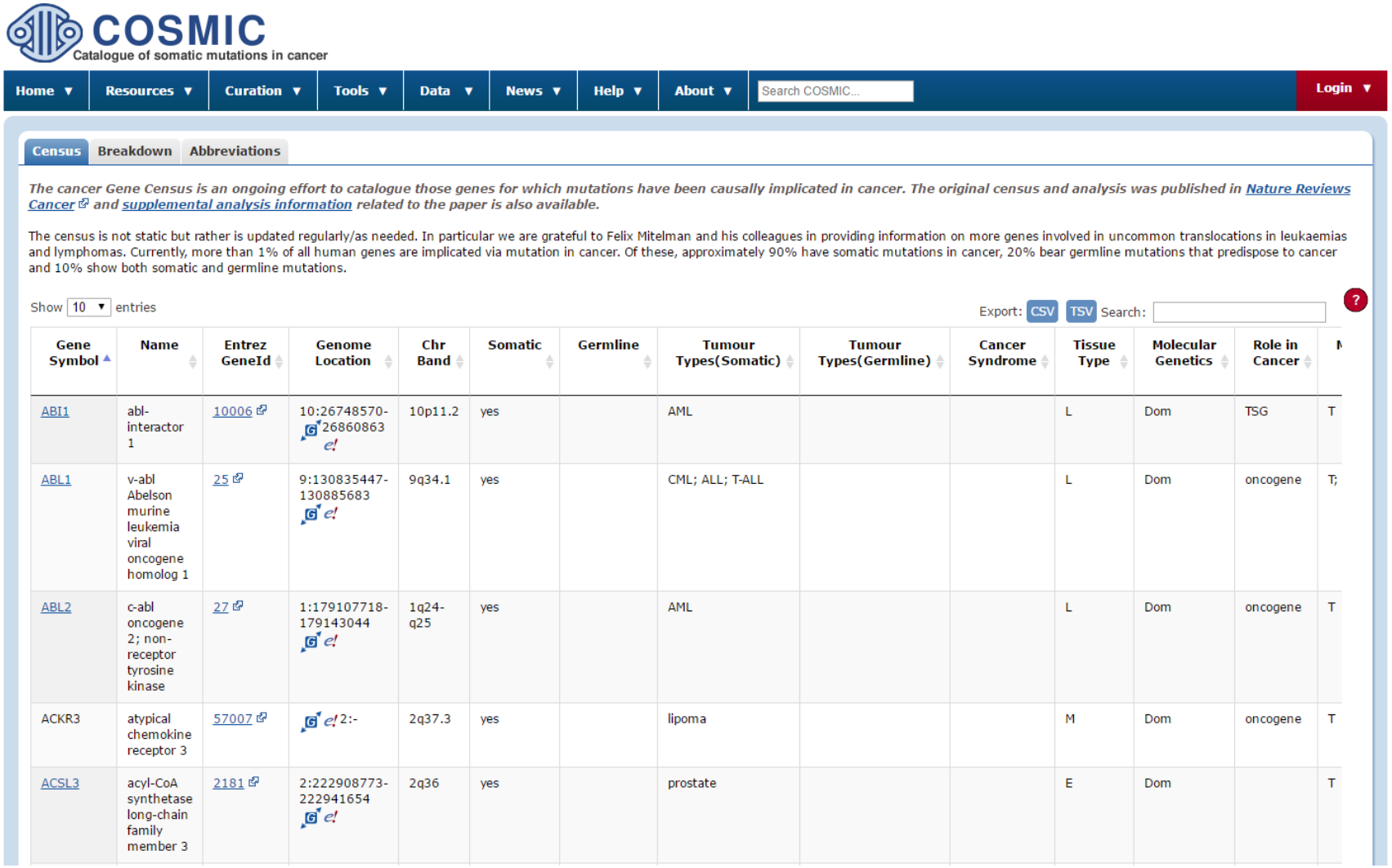
Further details on using COSMIC's content

COSMIC in BigQuery hosted by ISB-CGC



UNSW THE UNIVERSITY OF NEW SOUTH WALES

Cancer Gene Census



Gene Symbol	Name	Entrez GeneId	Genome Location	Chr Band	Somatic	Germline	Tumour Types(Somatic)	Tumour Types(Germline)	Cancer Syndrome	Tissue Type	Molecular Genetics	Role in Cancer	M
ABI1	abl-interactor 1	10006	10:26748570-26860863	10p11.2	yes		AML			L	Dom	TSG	T
ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	25	9:130835447-130885683	9q34.1	yes		CML; ALL; T-ALL			L	Dom	oncogene	T;
ABL2	c-abl oncogene 2; non-receptor tyrosine kinase	27	1:179107718-179143044	1q24-q25	yes		AML			L	Dom	oncogene	T
ACKR3	atypical chemokine receptor 3	57007	2:-	2q37.3	yes		lipoma			M	Dom	oncogene	T
ACSL3	acyl-CoA synthetase long-chain family member 3	2181	2:222908773-222941654	2q36	yes		prostate			E	Dom		T

Summary

- There are global cancer genome sequencing projects with publically available data

- TCGA data can downloaded or easily viewed through cBioPortal

- ICGC data can be downloaded or viewed from the user interface

- COSMIC database allows you to easily select cancer-associated genes

Exercises

1. Download patient clinical annotations for AML (TCGA dataset) using GDC data portal and then using the ICGC data portal.

2. Using the ICGC data portal:
   a. What is the cancer with most frequent RUNX1 mutations?
   b. Which cancer has the most RUNX1 frameshift mutations?

3. Using cBioPortal and COSMIC:
   a. Do kidney renal papillary cell carcinoma patients with BAP1 mutations have worse survival than those without?
   b. Is this gene listed in the Cancer Gene Census and, if so, what is its role in cancer?