Bioinformatics for cancer research

Bing Zhang, Ph.D.

Professor of Molecular and Human Genetics

Lester & Sue Smith Breast Center

Baylor College of Medicine

bing.zhang@bcm.edu

What is bioinformatics

Bio Bioinformatics informatics

- Hypotheses
- Questions
- Samples
- Experiments

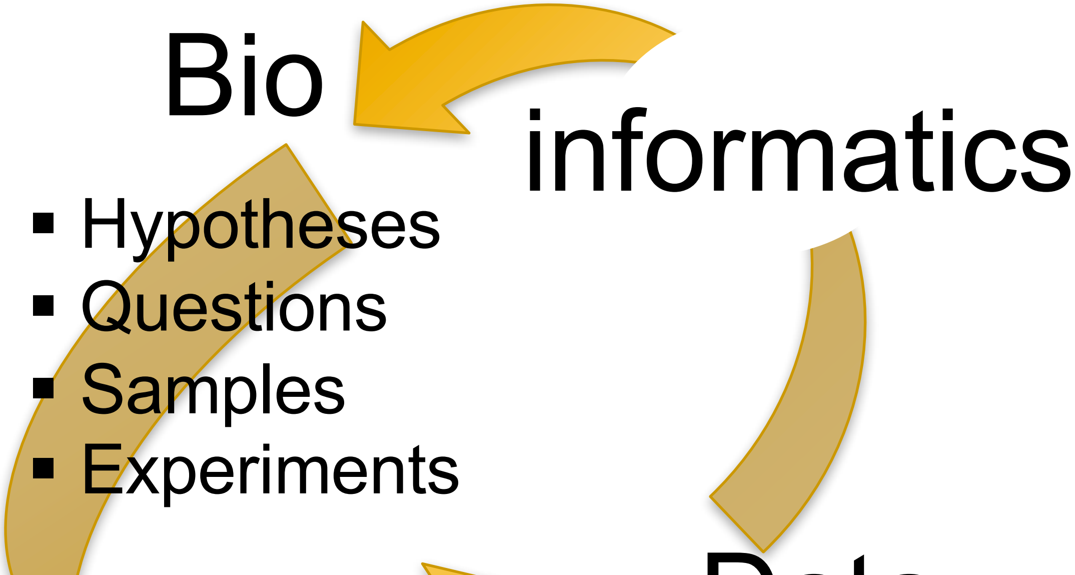
- Storage/retrieval
- Visualization
- Computational methods
- Statistical methods

Data

- DNA
- RNA
- Protein
- Metabolite
- Phenotype

- Sequence
- Expression
- Structure
- Interaction

Why now?



Bio

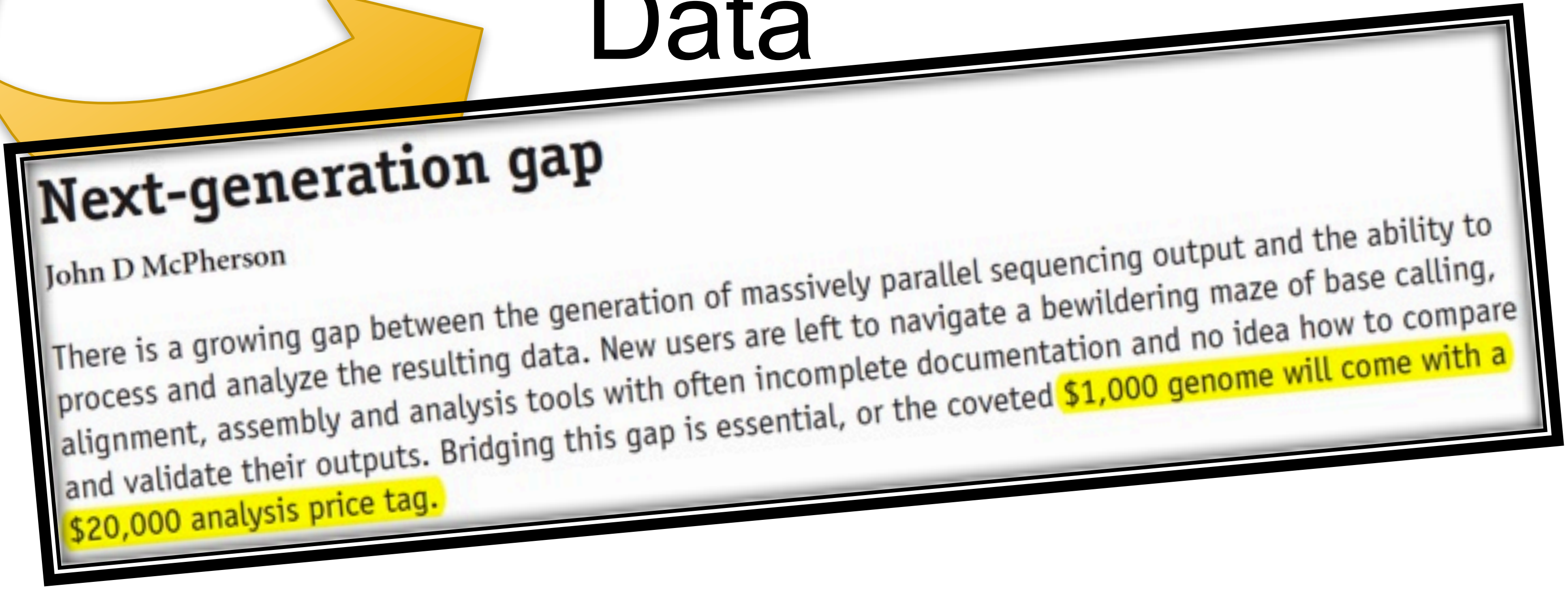
- Hypotheses
- Questions
- Samples
- Experiments

informatics

Data

informatics

- Storage/retrieval
- Visualization
- Computational methods
- Statistical methods



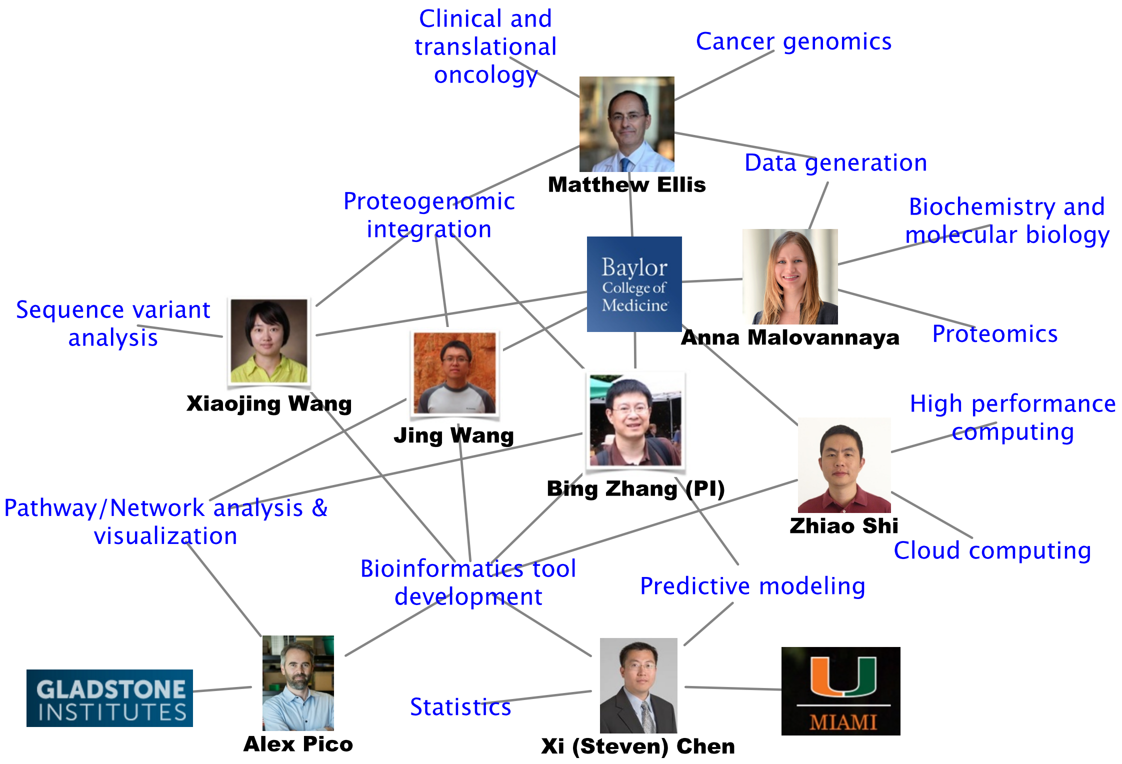
Next-generation gap

John D McPherson

There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag.

Translational Breast Cancer Research, 2016

Roles for different investigators in bioinformatics


Clinical and translational oncology

Cancer genomics

Matthew Ellis

Data generation

Proteogenomic integration

Biochemistry and molecular biology

Anna Malovannaya

Sequence variant analysis

Proteomics

Xiaojing Wang

Jing Wang

High performance computing

Bing Zhang (PI)

Zhiao Shi

Pathway/Network analysis & visualization

Cloud computing

Bioinformatics tool development

Predictive modeling

Statistics

Alex Pico

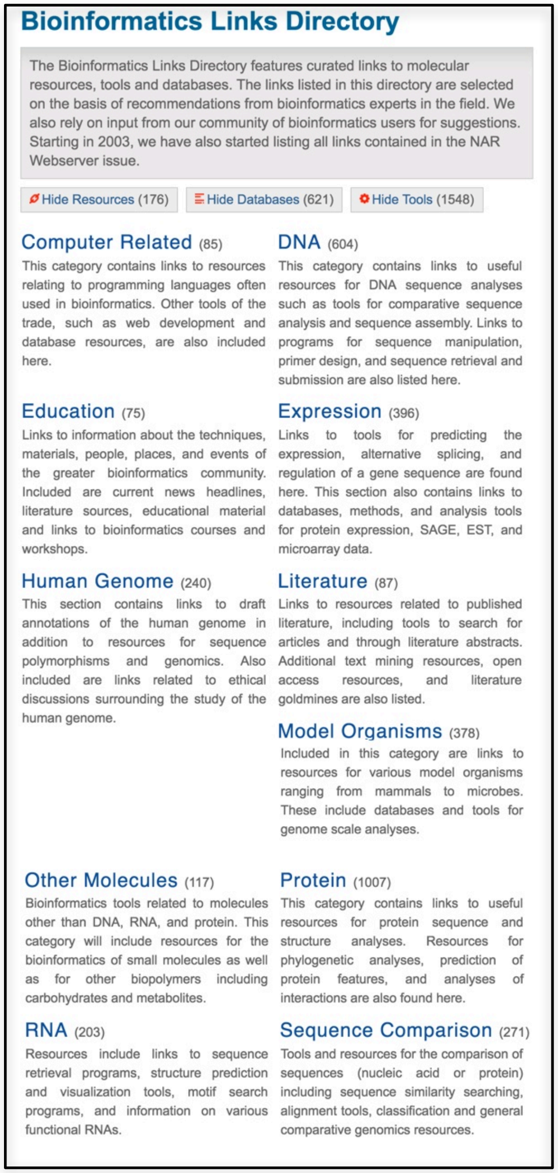
Xi (Steven) Chen

iPGDAC team

- Algorithm developer
  - Statisticians
  - Mathematicians
  - Computer scientists
- Tool developer
  - Bioinformaticians
- Data provider/consumer
  - Biologists

Translational Breast Cancer Research, 2016

Comprehensive list of bioinformatics resources



Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Computer Related (85)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

Education (75)

Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

Human Genome (240)

This section contains links to draft annotations of the human genome in addition to resources for sequence polymorphisms and genomics. Also included are links related to ethical discussions surrounding the study of the human genome.

DNA (604)

This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Expression (396)

Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data.

Literature (87)

Links to resources related to published literature, including tools to search for articles and through literature abstracts. Additional text mining resources, open access resources, and literature goldmines are also listed.

Model Organisms (378)

Included in this category are links to resources for various model organisms ranging from mammals to microbes. These include databases and tools for genome scale analyses.

Other Molecules (117)

Bioinformatics tools related to molecules other than DNA, RNA, and protein. This category will include resources for the bioinformatics of small molecules as well as for other biopolymers including carbohydrates and metabolites.

RNA (203)

Resources include links to sequence retrieval programs, structure prediction and visualization tools, motif search programs, and information on various functional RNAs.

Protein (1007)

This category contains links to useful resources for protein sequence and structure analyses. Resources for phylogenetic analyses, prediction of protein features, and analyses of interactions are also found here.

Sequence Comparison (271)

Tools and resources for the comparison of sequences (nucleic acid or protein) including sequence similarity searching, alignment tools, classification and general comparative genomics resources.

- October 2016
  - ❑ 176 Resources
  - ❑ 621 Databases
  - ❑ 1548 Tools

http://bioinformatics.ca/links_directory/

Translational Breast Cancer Research, 2016

Sequence and structure databases

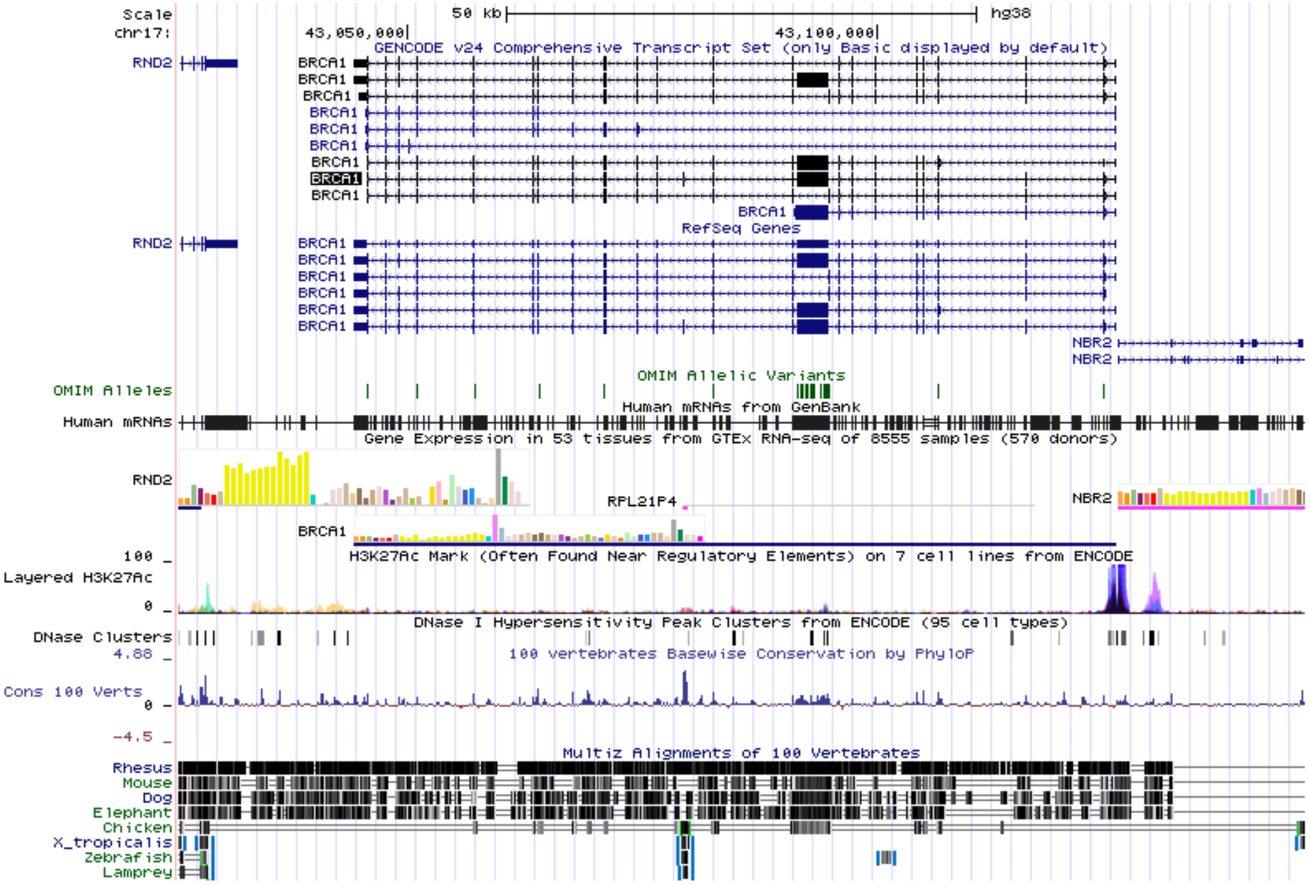
- Genbank: http://www.ncbi.nlm.nih.gov/genbank/
  - Annotated collection of all publicly available **DNA sequences**
  - 220,731,315,250 bases in 197,390,691 sequences as of October 2016
  - Whole Genome Sequencing (WGS) data: ftp://ftp.ncbi.nih.gov/ncbi-asn1/wgs ftp://ftp.ncbi.nih.gov/genbank/wgs
  - WGS: 1,676,238,489,250 bases in 363,213,315 sequences as of October 2016
- UniProt: http://www.uniprot.org/
  - Comprehensive resource for **protein sequences** and functional information
  - 552,259 reviewed entries as of October 2016
- PDB: http://www.rcsb.org/
  - **3D structures** of large biological molecules, including proteins, nucleic acids, and complex assemblies
  - 123,870 structures as of October 2016
- Pfam: http://pfam.xfam.org/
  - Collection of **protein families**, each represented by multiple sequence alignments and hidden Markov models (HMMs)
  - 16,306 families as of October 2016

Genome browsers

Graph interface for browsing and visualizing genome-wide sequence and annotation data.

- UCSC genome browser
  - http://genome.ucsc.edu/cgi-bin/hgGateway
- Ensembl genome browser
  - http://www.ensembl.org/index.html
- Integrative Genomics Viewer (IGV)
  - http://software.broadinstitute.org/software/igv/

UCSC genome browser screenshot



Scale
chr17:
50 kb
hg38
43,050,000
43,100,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
RND2
BRCA1
RefSeq Genes
NBR2
OMIM Alleles
OMIM Allelic Variants
Human mRNAs
Human mRNAs from GenBank
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
RPL21P4
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
4.88
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey

Base position

GenCode annotation

RefSeq annotation

OMIM Alleles
Human mRNAs

Tissue expression

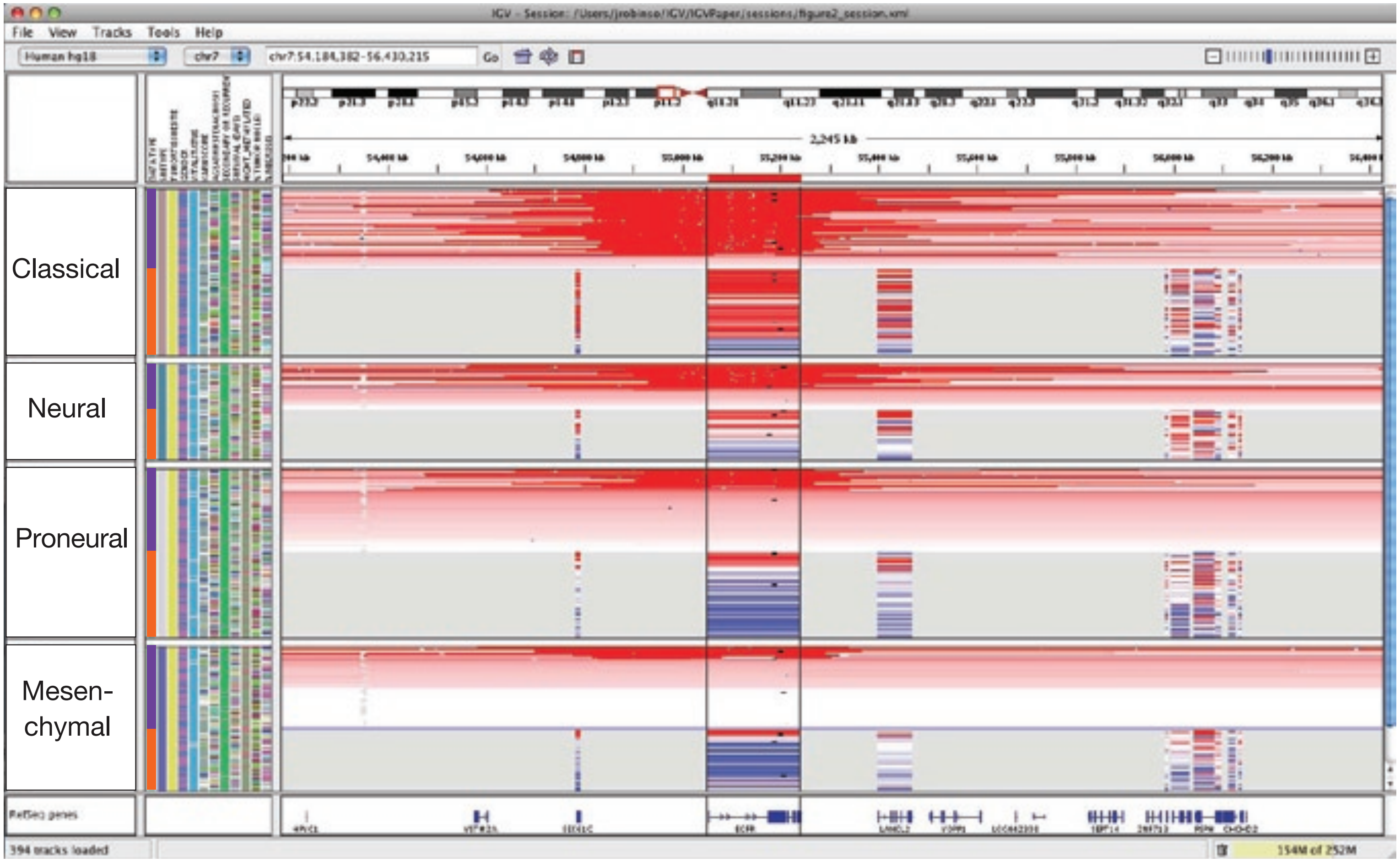
Regulatory elements

Comparative genomics

Translational Breast Cancer Research, 2016

Genome browsers

IGV: copy number, expression and mutation data grouped by tumor subtype



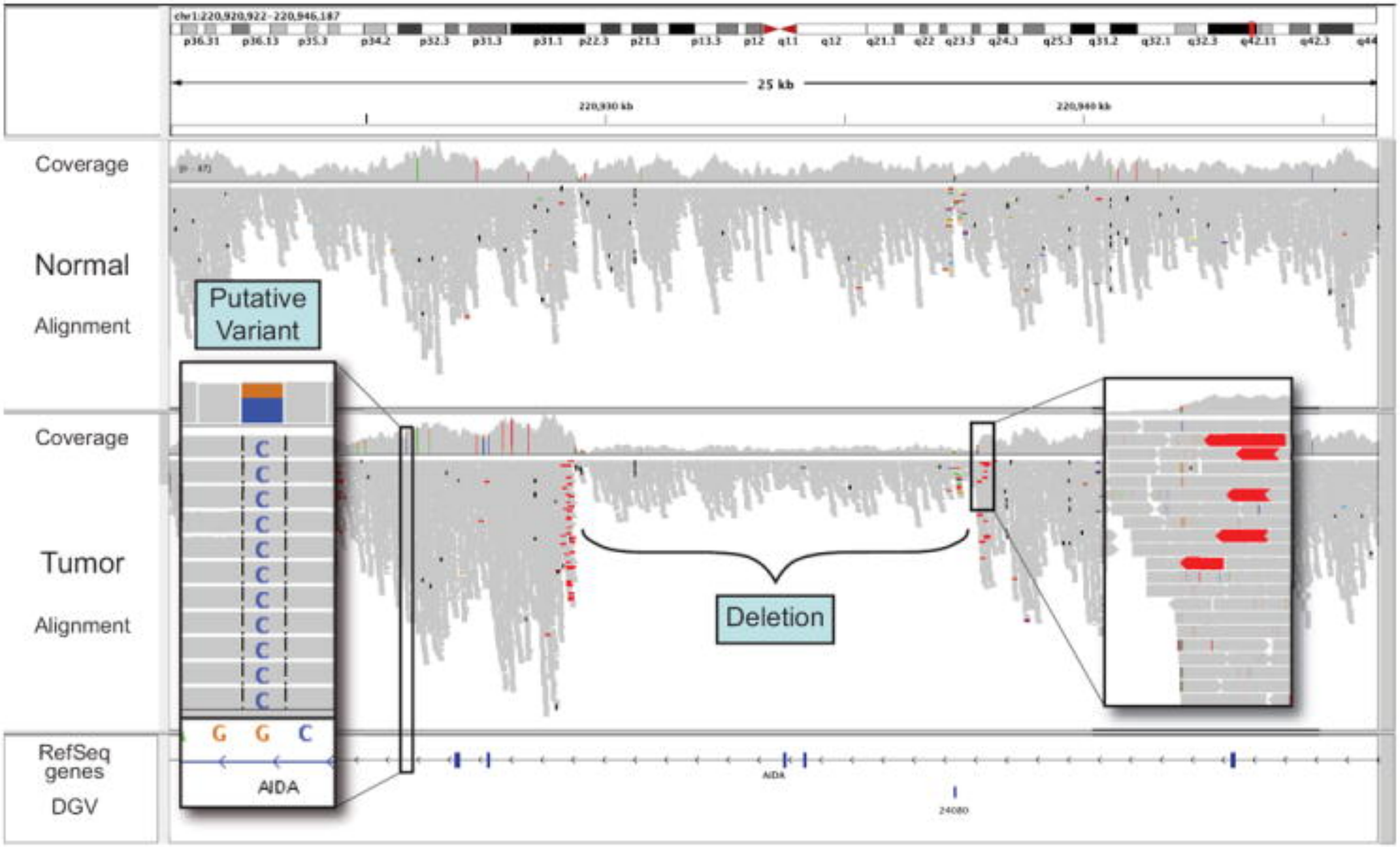
EGFR

Robinson et al. Nat Biotechnol, 2011

Translational Breast Cancer Research, 2016

Genome browsers

IGV: view of aligned reads at 20Kb resolution



Robinson et al. Nat Biotechnol, 2011

Gene-centric databases

- Entrez Gene
  - http://www.ncbi.nlm.nih.gov/gene
  - NCBI/NIH
  - All completely sequenced genomes
  - **One gene per page**
- Ensembl BioMart
  - http://www.ensembl.org/biomart/martview
  - EMBL-EBI and Sanger Institute
  - Vertebrates and other selected eukaryotic species
  - **Batch information retrieval**

Gene/protein expression data repositories

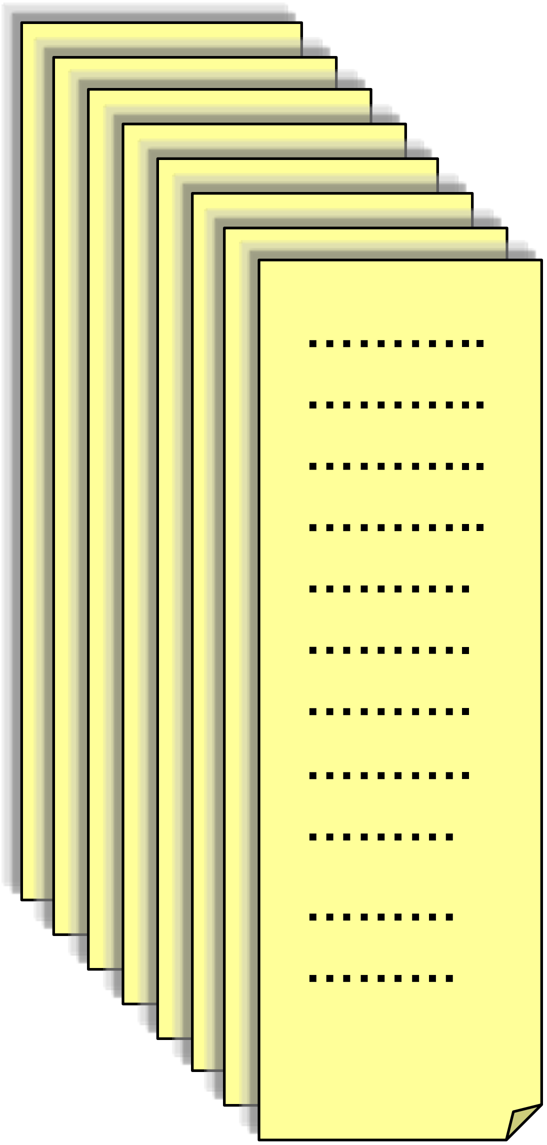
- Gene Expression Omnibus (GEO)
  - http://www.ncbi.nlm.nih.gov/geo/
- ArrayExpress
  - http://www.ebi.ac.uk/arrayexpress/
- PRIDE
  - https://www.ebi.ac.uk/pride/archive/

Pathway and network databases

- Gene Ontology (GO): http://www.geneontology.org/
- Pathway databases
  - KEGG: http://www.genome.jp/kegg/pathway.html
  - Reactome: http://www.reactome.org/
  - WikiPathways: http://www.wikipathways.org/
- Protein-protein interaction databases
  - DIP: http://dip.doe-mbi.ucla.edu/
  - MINT: http://mint.bio.uniroma2.it/mint/
  - BioGRID: http://www.thebiogrid.org/
  - HPRD: http://www.hprd.org
  - iRef: http://wodaklab.org/iRefWeb
- Protein-DNA interaction database
  - Transfac: http://www.gene-regulation.com
  - Jaspar: http://jaspar.genereg.net/

Pathway and network analysis: motivation

- Genomics
  - Genome Wide Association Study (GWAS)
  - Whole genome or exome sequencing
  - Copy number analysis
- Epigenomics
  - DNA methylation
- Transcriptomics
  - mRNA profiling
    - Microarray
    - RNA-Seq
  - Protein-DNA interaction
    - Chromatin immunoprecipitation (ChIP)-Seq
- Proteomics
  - Protein profiling
    - LC-MS/MS
  - Protein-protein interaction
    - Yeast two hybrid
    - Affinity pull-down/LC-MS/MS



Pathway and network analysis: tools

- Pathway analysis
  - WebGestalt: http://www.webgestalt.org
  - DAVID: https://david.ncifcrf.gov/
  - GSEA: http://software.broadinstitute.org/gsea
- Network analysis
  - Cytoscape: http://www.cytoscape.org/
  - NetGestalt: http://www.netgestal.org
  - STRING: http://string-db.org
  - GeneMANIA: http://genemania.org/
  - Gene2Net: http://www.gene2net.org

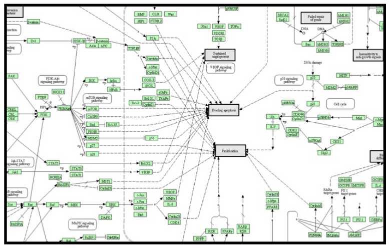
WebGestalt: http://www.webgestalt.org

Gene list

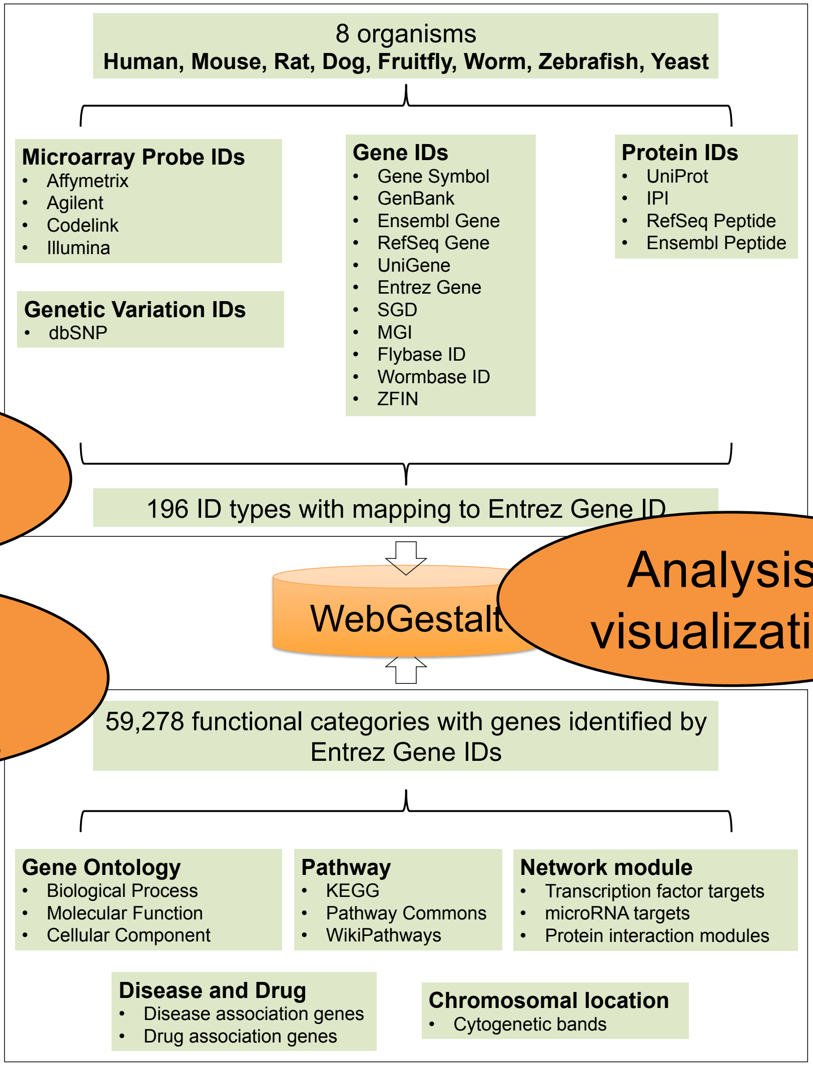
92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
……

~200
ID types

~60K
Functional
categories



Pathways/
functional categories



8 organisms
Human, Mouse, Rat, Dog, Fruitfly, Worm, Zebrafish, Yeast

Microarray Probe IDs
- Affymetrix
- Agilent
- Codelink
- Illumina

Genetic Variation IDs
- dbSNP

Gene IDs
- Gene Symbol
- GenBank
- Ensembl Gene
- RefSeq Gene
- UniGene
- Entrez Gene
- SGD
- MGI
- Flybase ID
- Wormbase ID
- ZFIN

Protein IDs
- UniProt
- IPI
- RefSeq Peptide
- Ensembl Peptide

196 ID types with mapping to Entrez Gene ID

WebGestalt

Analysis/
visualization

59,278 functional categories with genes identified by Entrez Gene IDs

Gene Ontology
- Biological Process
- Molecular Function
- Cellular Component

Pathway
- KEGG
- Pathway Commons
- WikiPathways

Network module
- Transcription factor targets
- microRNA targets
- Protein interaction modules

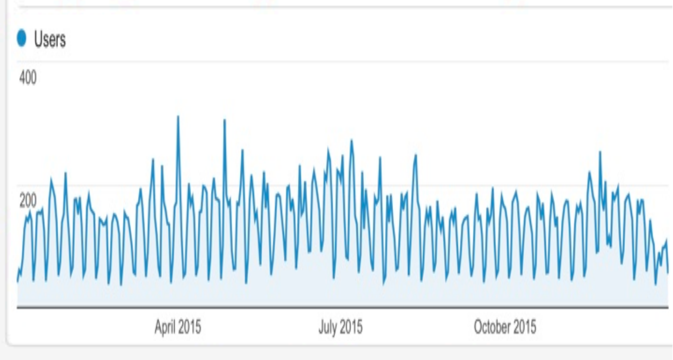
Disease and Drug
- Disease association genes
- Drug association genes

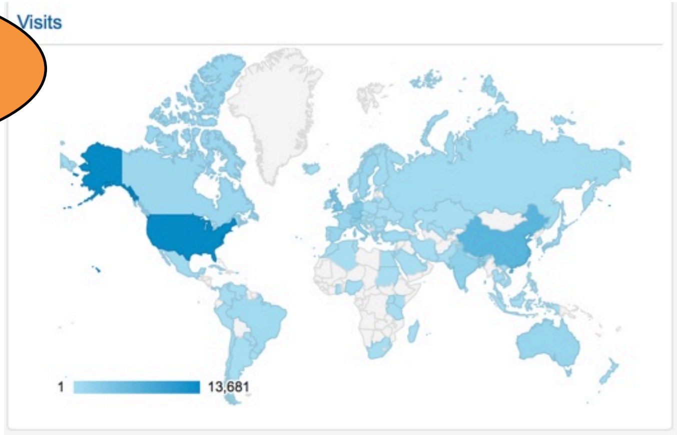
Chromosomal location
- Cytogenetic bands

Zhang et.al. Nucleic Acids Res. 33:W741, 2005
Wang et al. Nucleic Acids Res. 41:W77, 2013

http://www.webgestalt.org

Daily Unique Visitors (Jan 2015 – Dec 2015)

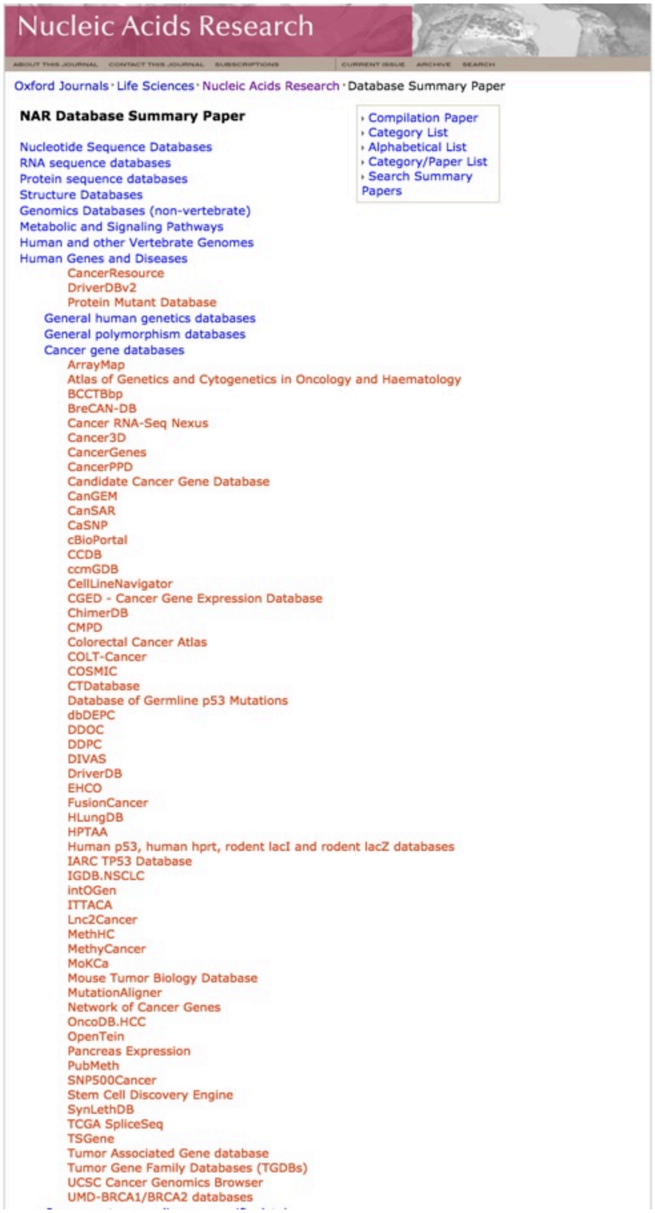




Jan. 1, 2015 – Dec. 31, 2015
63,932 visits from 27,409 visitors
>300 citations

Cancer-specific resources

- Pavlopoulou et al., **Human cancer databases (Review), Oncology Reports**, 33:3-18, 2015
- Yang et al., **Databases and web tools for cancer genomics study**, Genomics proteomics bioinformatics, 13: 46-50, 2015
- https://www.oxfordjournals.org/our_journals/nar/database/subcat/8/33



Oxford Journals · Life Sciences · Nucleic Acids Research · Database Summary Paper

NAR Database Summary Paper

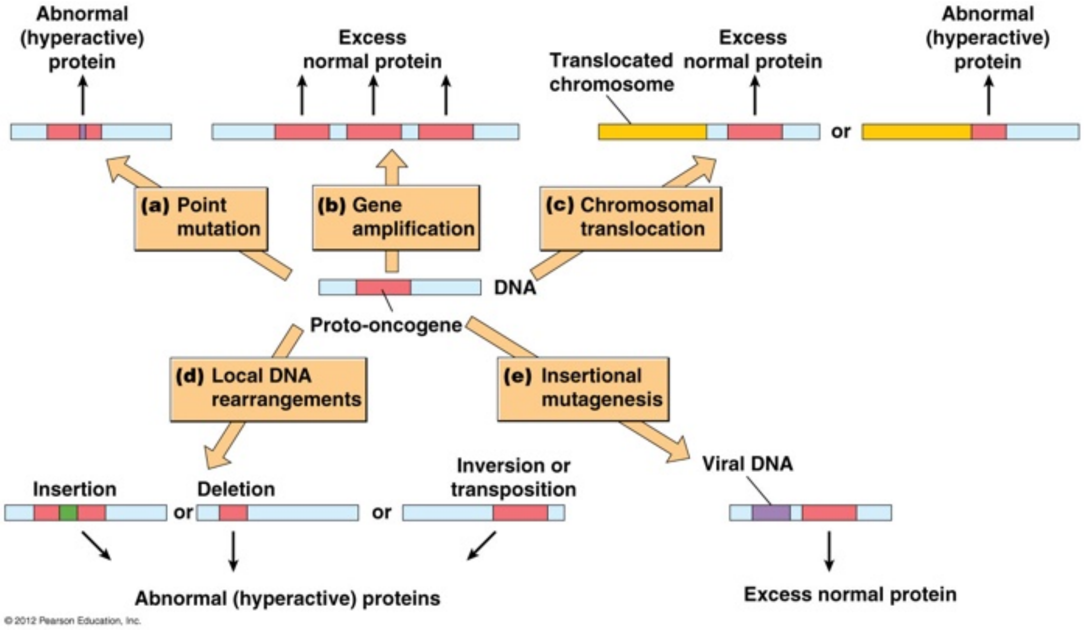
- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
- CancerResource
- DriverDBv2
- Protein Mutant Database

General human genetics databases
General polymorphism databases
Cancer gene databases
- ArrayMap
- Atlas of Genetics and Cytogenetics in Oncology and Haematology
- BCCTBbp
- BreCAN-DB
- Cancer RNA-Seq Nexus
- Cancer3D
- CancerGenes
- CancerPPD
- Candidate Cancer Gene Database
- CanGEM
- CanSAR
- CaSNP
- cBioPortal
- CCDB
- ccmGDB
- CellLineNavigator
- CGED - Cancer Gene Expression Database
- ChimerDB
- CMPD
- Colorectal Cancer Atlas
- COLT-Cancer
- COSMIC
- CTDatabase
- Database of Germline p53 Mutations
- dbDEPC
- DDOC
- DDPC
- DIVAS
- DriverDB
- EHCO
- FusionCancer
- HLungDB
- HPTAA
- Human p53, human hprt, rodent lacI and rodent lacZ databases
- IARC TP53 Database
- IGDB.NSCLC
- intOGen
- ITTACA
- Lnc2Cancer
- MethHC
- MethyCancer
- MoKCa
- Mouse Tumor Biology Database
- MutationAligner
- Network of Cancer Genes
- OncoDB.HCC
- OpenTein
- Pancreas Expression
- PubMeth
- SNP500Cancer
- Stem Cell Discovery Engine
- SynLethDB
- TCGA SpliceSeq
- TSGene
- Tumor Associated Gene database
- Tumor Gene Family Databases (TGDBs)
- UCSC Cancer Genomics Browser
- UMD-BRCA1/BRCA2 databases

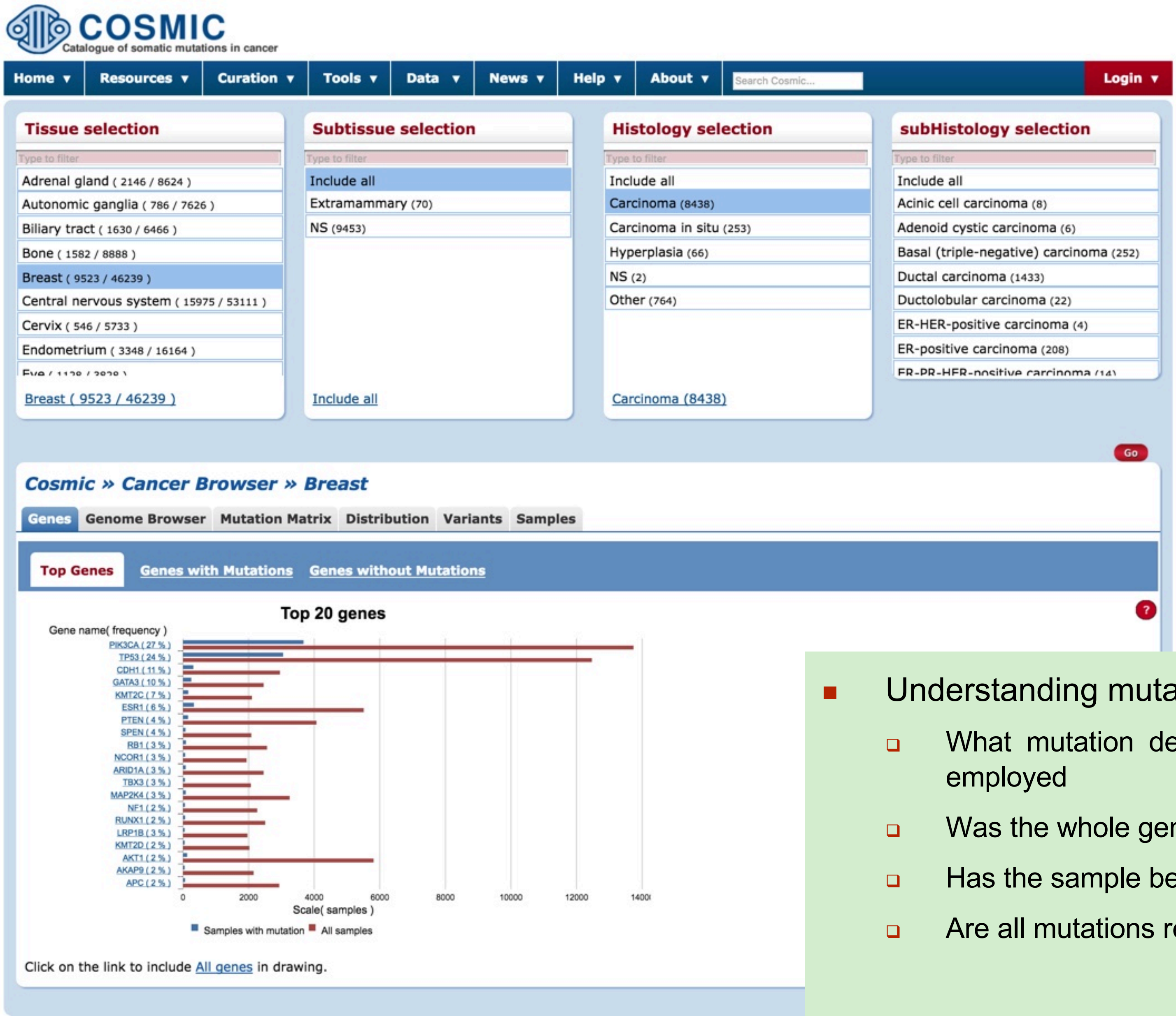
Catalogue of somatic mutations in cancer (COSMIC)

- COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers
- Wellcome Trust Sanger Institute
- http://cancer.sanger.ac.uk/cosmic
- Expert curation data and genome-wide screen data
- Search by gene, cancer type, mutation, or sample


Abnormal (hyperactive) protein
Excess normal protein
Translocated chromosome
Excess normal protein
or
Abnormal (hyperactive) protein
(a) Point mutation
(b) Gene amplification
(c) Chromosomal translocation
DNA
Proto-oncogene
(d) Local DNA rearrangements
(e) Insertional mutagenesis
Insertion
or
Deletion
or
Inversion or transposition
Viral DNA
Abnormal (hyperactive) proteins
Excess normal protein
© 2012 Pearson Education, Inc.

Translational Breast Cancer Research, 2016

COSMIC: breast cancer



Tissue selection

Type to filter

Adrenal gland (2146 / 8624)
Autonomic ganglia (786 / 7626)
Biliary tract (1630 / 6466)
Bone (1582 / 8888)
Breast (9523 / 46239)
Central nervous system (15975 / 53111)
Cervix (546 / 5733)
Endometrium (3348 / 16164)
Eye (1128 / 2828)

Breast (9523 / 46239)

Subtissue selection

Type to filter

Include all
Extramammary (70)
NS (9453)

Include all

Histology selection

Type to filter

Include all
Carcinoma (8438)
Carcinoma in situ (253)
Hyperplasia (66)
NS (2)
Other (764)

Carcinoma (8438)

subHistology selection

Type to filter

Include all
Acinic cell carcinoma (8)
Adenoid cystic carcinoma (6)
Basal (triple-negative) carcinoma (252)
Ductal carcinoma (1433)
Ductolobular carcinoma (22)
ER-HER-positive carcinoma (4)
ER-positive carcinoma (208)
ER-PR-HER-positive carcinoma (14)

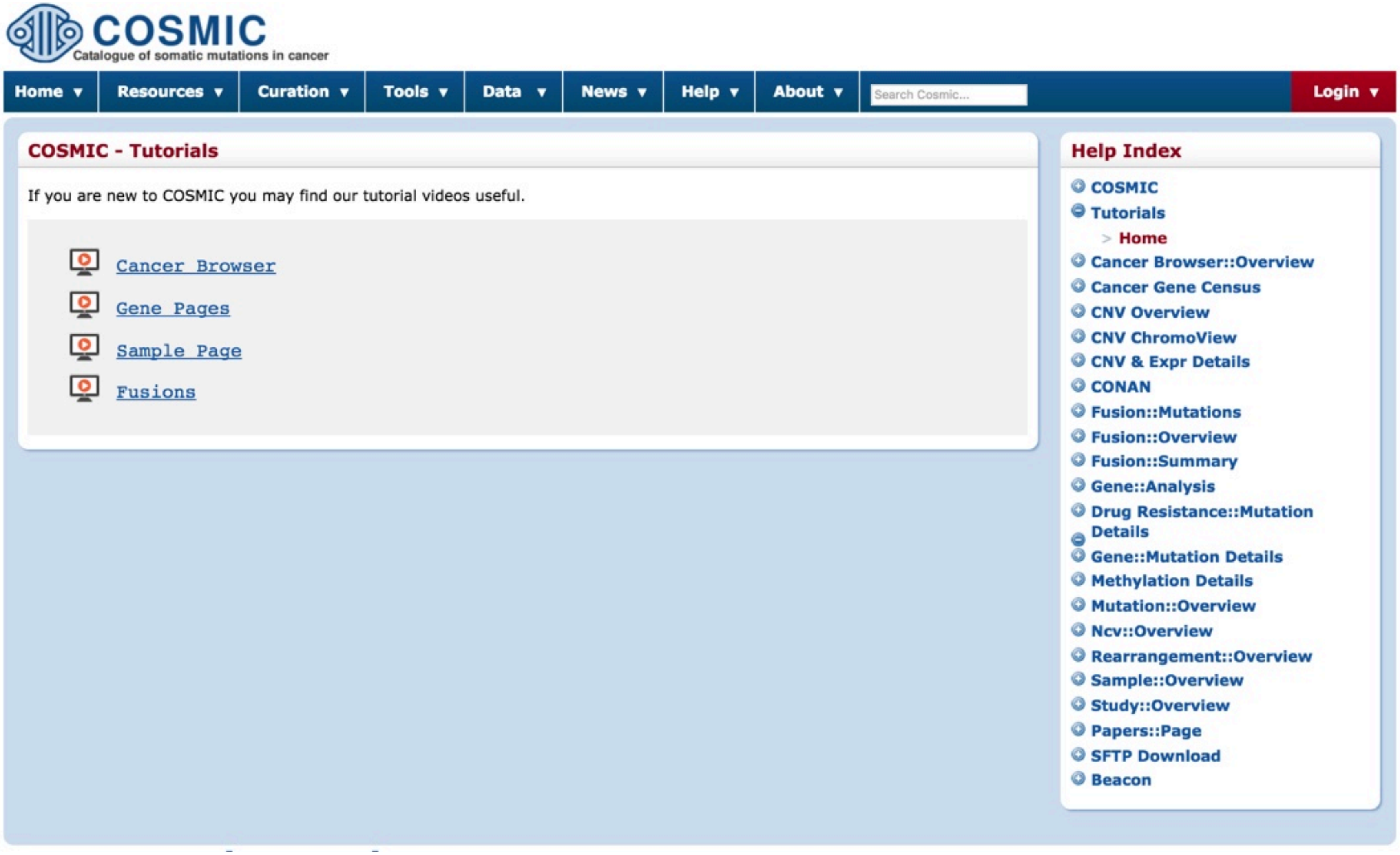
Cosmic » Cancer Browser » Breast

Click on the link to include All genes in drawing.

- Understanding mutation frequency
  - What mutation detection method was employed
  - Was the whole gene screened
  - Has the sample been screened before
  - Are all mutations real?

Translational Breast Cancer Research, 2016

COSMIC: getting help



COSMIC - Tutorials

If you are new to COSMIC you may find our tutorial videos useful.

- Cancer Browser
- Gene Pages
- Sample Page
- Fusions

Help Index

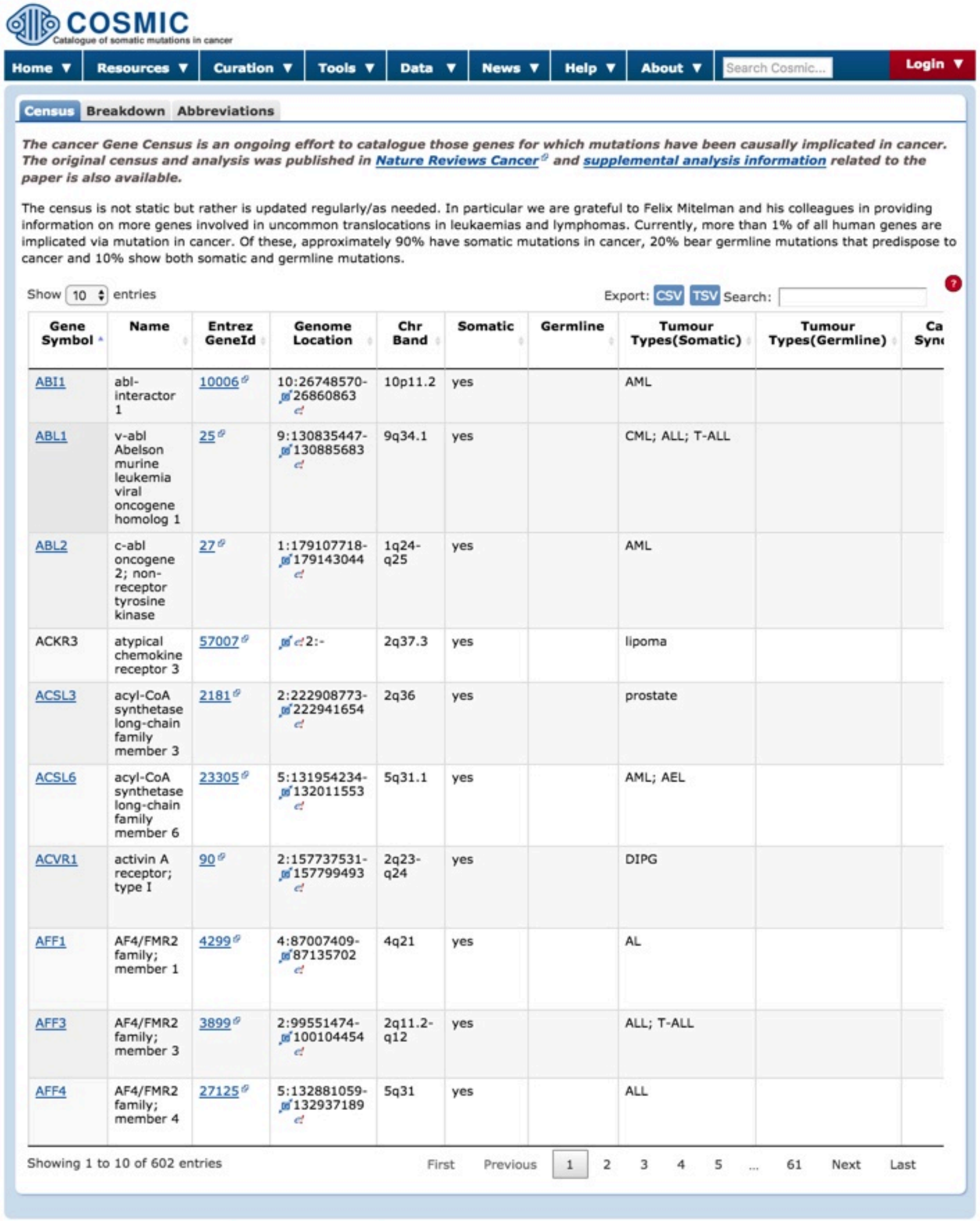
- COSMIC
- Tutorials
  - > Home
- Cancer Browser::Overview
- Cancer Gene Census
- CNV Overview
- CNV ChromoView
- CNV & Expr Details
- CONAN
- Fusion::Mutations
- Fusion::Overview
- Fusion::Summary
- Gene::Analysis
- Drug Resistance::Mutation Details
- Gene::Mutation Details
- Methylation Details
- Mutation::Overview
- Ncv::Overview
- Rearrangement::Overview
- Sample::Overview
- Study::Overview
- Papers::Page
- SFTP Download
- Beacon

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

Cancer Gene Census

- Futreal et al. **A census of human cancer genes**. Nature Reviews Cancer, 4:2004
- The Cancer Gene Census is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer.
- 602 genes as of October 2016.

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



The cancer Gene Census is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. The original census and analysis was published in Nature Reviews Cancer and supplemental analysis information related to the paper is also available.

The census is not static but rather is updated regularly/as needed. In particular we are grateful to Felix Mitelman and his colleagues in providing information on more genes involved in uncommon translocations in leukaemias and lymphomas. Currently, more than 1% of all human genes are implicated via mutation in cancer. Of these, approximately 90% have somatic mutations in cancer, 20% bear germline mutations that predispose to cancer and 10% show both somatic and germline mutations.

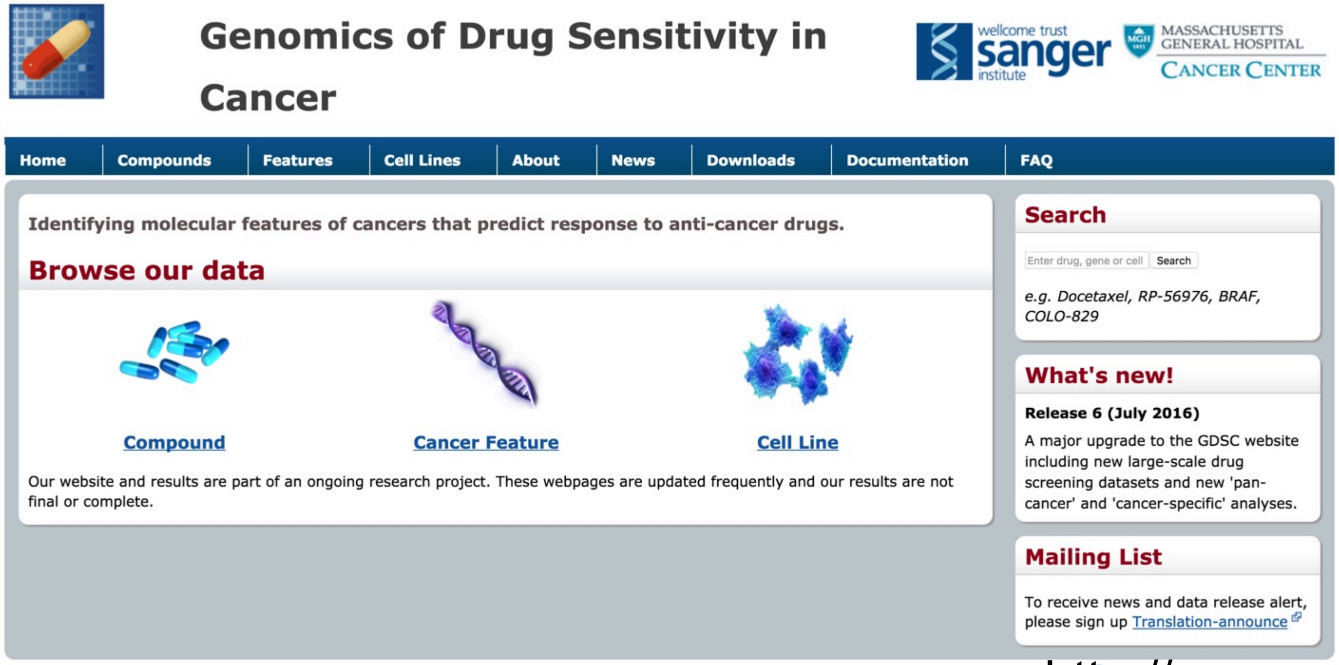
Show 10 entries   Export: CSV TSV Search:

Gene Symbol	Name	Entrez GeneId	Genome Location	Chr Band	Somatic	Germline	Tumour Types(Somatic)	Tumour Types(Germline)	Ca Sync
ABI1	abl-interactor 1	10006	10:26748570-26860863	10p11.2	yes		AML		
ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	25	9:130835447-130885683	9q34.1	yes		CML; ALL; T-ALL		
ABL2	c-abl oncogene 2; non-receptor tyrosine kinase	27	1:179107718-179143044	1q24-q25	yes		AML		
ACKR3	atypical chemokine receptor 3	57007	2:-	2q37.3	yes		lipoma		
ACSL3	acyl-CoA synthetase long-chain family member 3	2181	2:222908773-222941654	2q36	yes		prostate		
ACSL6	acyl-CoA synthetase long-chain family member 6	23305	5:131954234-132011553	5q31.1	yes		AML; AEL		
ACVR1	activin A receptor; type I	90	2:157737531-157799493	2q23-q24	yes		DIPG		
AFF1	AF4/FMR2 family; member 1	4299	4:87007409-87135702	4q21	yes		AL		
AFF3	AF4/FMR2 family; member 3	3899	2:99551474-100104454	2q11.2-q12	yes		ALL; T-ALL		
AFF4	AF4/FMR2 family; member 4	27125	5:132881059-132937189	5q31	yes		ALL		

Showing 1 to 10 of 602 entries

Genomics of Drug Sensitivity in Cancer (GDSC)

- A collaboration between the Cancer Genome Project at the Wellcome Trust Sanger Institute (UK) and the Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center (USA), funded by the Wellcome Trust.
- Goal: to identify molecular features of cancers that predict response to anti-cancer drugs.



Genomics of Drug Sensitivity in Cancer

Identifying molecular features of cancers that predict response to anti-cancer drugs.

Browse our data

Our website and results are part of an ongoing research project. These webpages are updated frequently and our results are not final or complete.

Search

e.g. Docetaxel, RP-56976, BRAF, COLO-829

What's new!

Release 6 (July 2016)

A major upgrade to the GDSC website including new large-scale drug screening datasets and new 'pan-cancer' and 'cancer-specific' analyses.

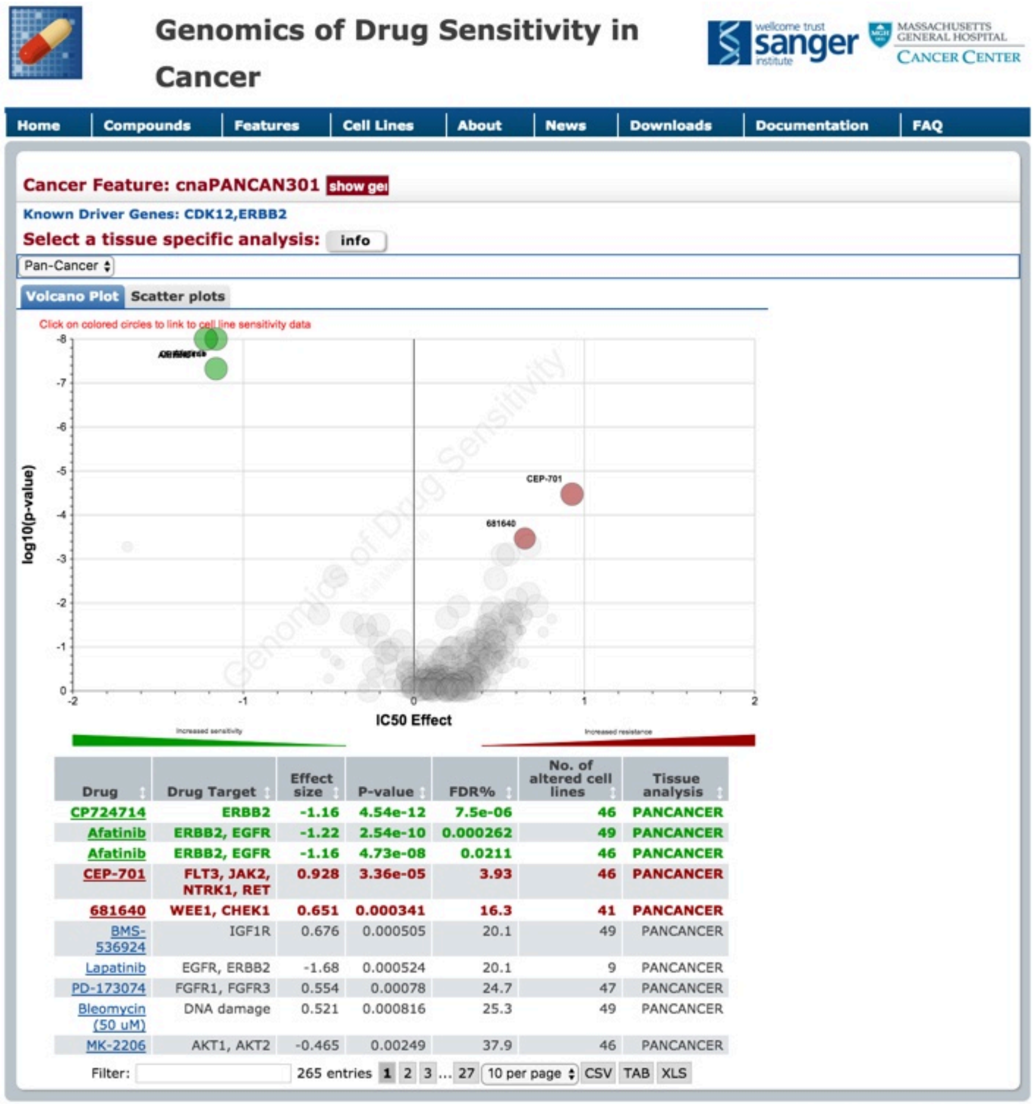
Mailing List

To receive news and data release alert, please sign up Translation-announce

http://www.cancerrxgene.org/

Translational Breast Cancer Research, 2016

GDSC: drug sensitivity vs Her2 amplification



Genomics of Drug Sensitivity in Cancer

wellcome trust sanger institute
MASSACHUSETTS GENERAL HOSPITAL CANCER CENTER
Home
Compounds
Features
Cell Lines
About
News
Downloads
Documentation
FAQ

Cancer Feature: cnaPANCAN301 show ge

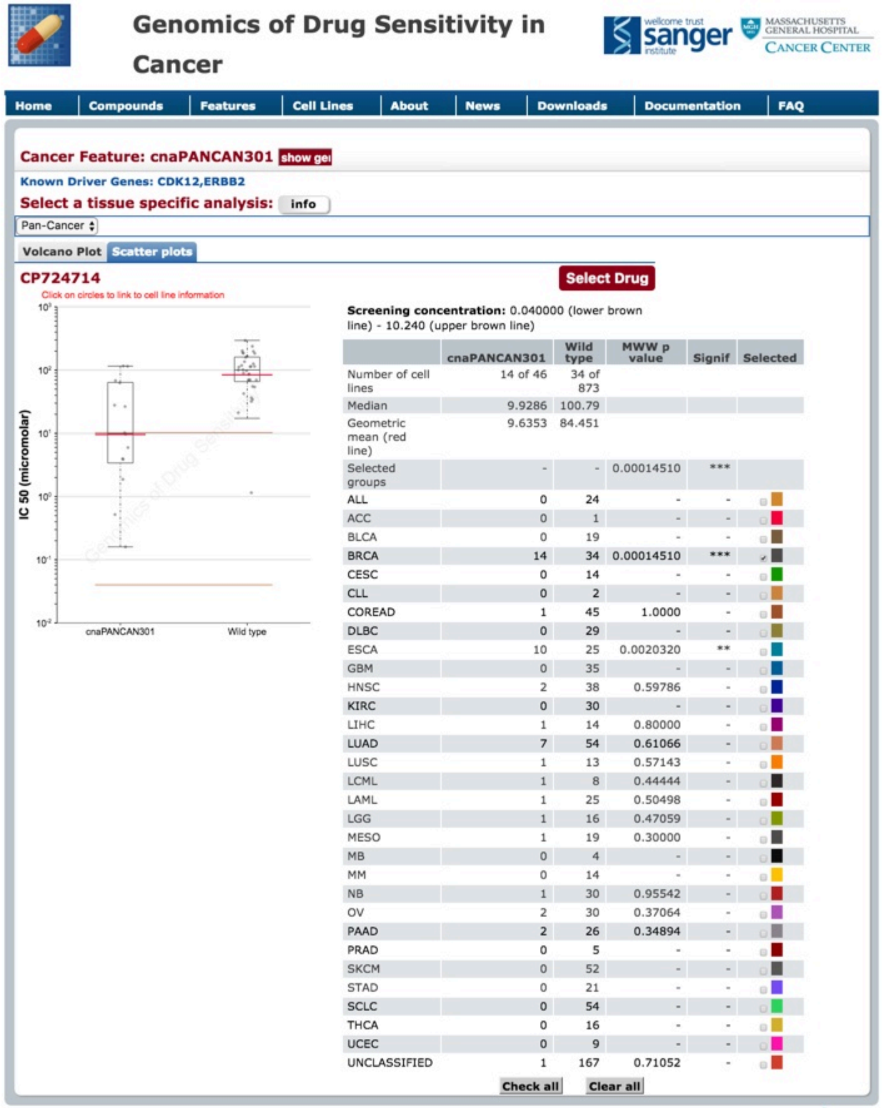
Known Driver Genes: CDK12,ERBB2

Select a tissue specific analysis: info

Pan-Cancer

Drug	Drug Target	Effect size	P-value	FDR%	No. of altered cell lines	Tissue analysis
CP724714	ERBB2	-1.16	4.54e-12	7.5e-06	46	PANCANCER
Afatinib	ERBB2, EGFR	-1.22	2.54e-10	0.000262	49	PANCANCER
Afatinib	ERBB2, EGFR	-1.16	4.73e-08	0.0211	46	PANCANCER
CEP-701	FLT3, JAK2, NTRK1, RET	0.928	3.36e-05	3.93	46	PANCANCER
681640	WEE1, CHEK1	0.651	0.000341	16.3	41	PANCANCER
BMS-536924	IGF1R	0.676	0.000505	20.1	49	PANCANCER
Lapatinib	EGFR, ERBB2	-1.68	0.000524	20.1	9	PANCANCER
PD-173074	FGFR1, FGFR3	0.554	0.00078	24.7	47	PANCANCER
Bleomycin (50 uM)	DNA damage	0.521	0.000816	25.3	49	PANCANCER
MK-2206	AKT1, AKT2	-0.465	0.00249	37.9	46	PANCANCER

Filter: 265 entries 1 2 3 ... 27 10 per page CSV TAB XLS



Genomics of Drug Sensitivity in Cancer

wellcome trust sanger institute
MASSACHUSETTS GENERAL HOSPITAL CANCER CENTER
Home
Compounds
Features
Cell Lines
About
News
Downloads
Documentation
FAQ

Cancer Feature: cnaPANCAN301 show ge

Known Driver Genes: CDK12,ERBB2

Select a tissue specific analysis: info

Pan-Cancer

Screening concentration: 0.040000 (lower brown line) - 10.240 (upper brown line)

	cnaPANCAN301	Wild type	MWW p value	Signif	Selected
Number of cell lines	14 of 46	34 of 873			
Median	9.9286	100.79			
Geometric mean (red line)	9.6353	84.451			
Selected groups	-	-	0.00014510	***	
ALL	0	24	-	-	□ 🟧
ACC	0	1	-	-	□ 🟥
BLCA	0	19	-	-	□ 🟫
BRCA	14	34	0.00014510	***	☑ ⬛
CESC	0	14	-	-	□ 🟩
CLL	0	2	-	-	□ 🟧
COREAD	1	45	1.0000	-	□ 🟫
DLBC	0	29	-	-	□ 🟫
ESCA	10	25	0.0020320	**	□ 🟦
GBM	0	35	-	-	□ 🟦
HNSC	2	38	0.59786	-	□ 🟦
KIRC	0	30	-	-	□ 🟪
LIHC	1	14	0.80000	-	□ 🟪
LUAD	7	54	0.61066	-	□ 🟧
LUSC	1	13	0.57143	-	□ 🟧
LCML	1	8	0.44444	-	□ ⬛
LAML	1	25	0.50498	-	□ 🟥
LGG	1	16	0.47059	-	□ 🟩
MESO	1	19	0.30000	-	□ ⬛
MB	0	4	-	-	□ ⬛
MM	0	14	-	-	□ 🟨
NB	1	30	0.95542	-	□ 🟥
OV	2	30	0.37064	-	□ 🟪
PAAD	2	26	0.34894	-	□ ⬜
PRAD	0	5	-	-	□ 🟥
SKCM	0	52	-	-	□ ⬛
STAD	0	21	-	-	□ 🟪
SCLC	0	54	-	-	□ 🟩
THCA	0	16	-	-	□ 🟨
UCEC	0	9	-	-	□ 🟪
UNCLASSIFIED	1	167	0.71052	-	□ 🟥

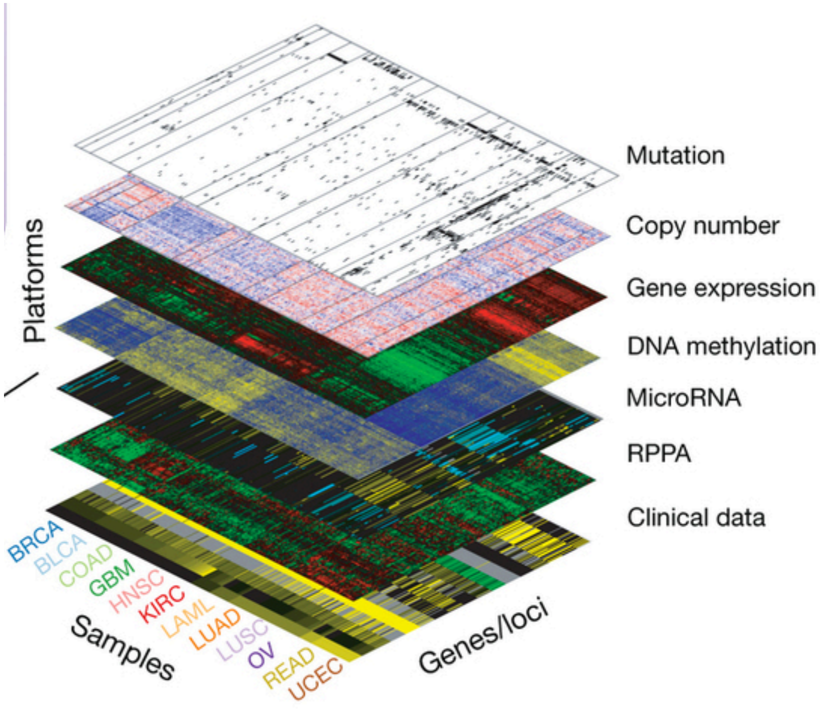
Check all Clear all

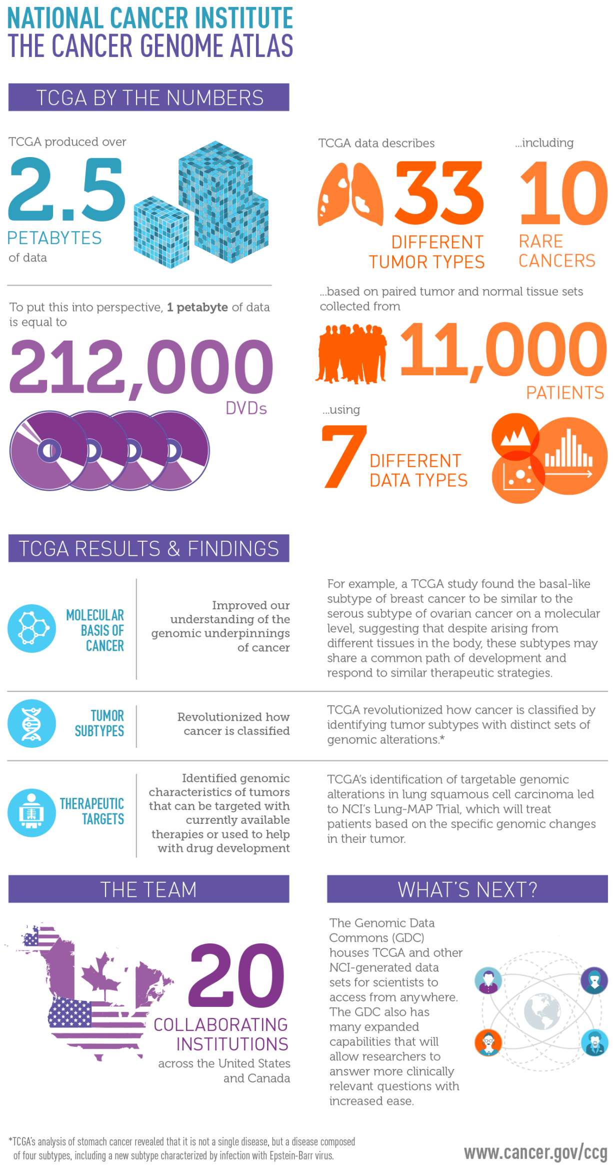
http://www.cancerrxgene.org/

Translational Breast Cancer Research, 2016

The Cancer Genome Atlas (TCGA)

- A collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI)
- To accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.





NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

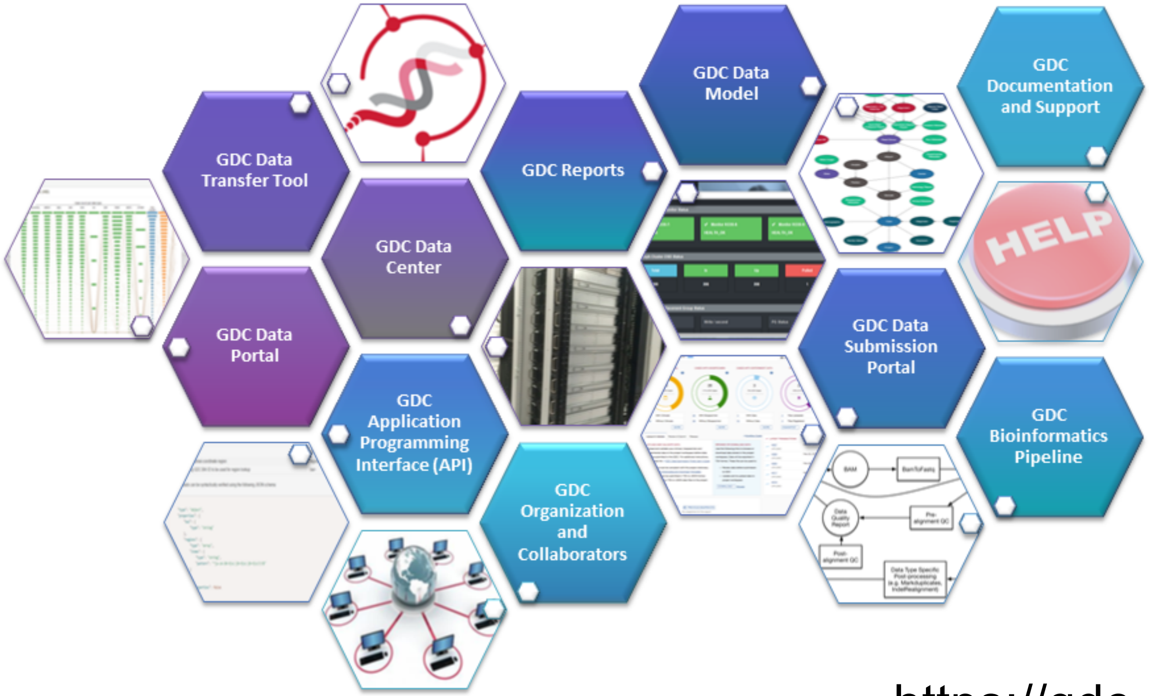
*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

Translational Breast Cancer Research, 2016

NCI Genomic Data Commons (GDC)

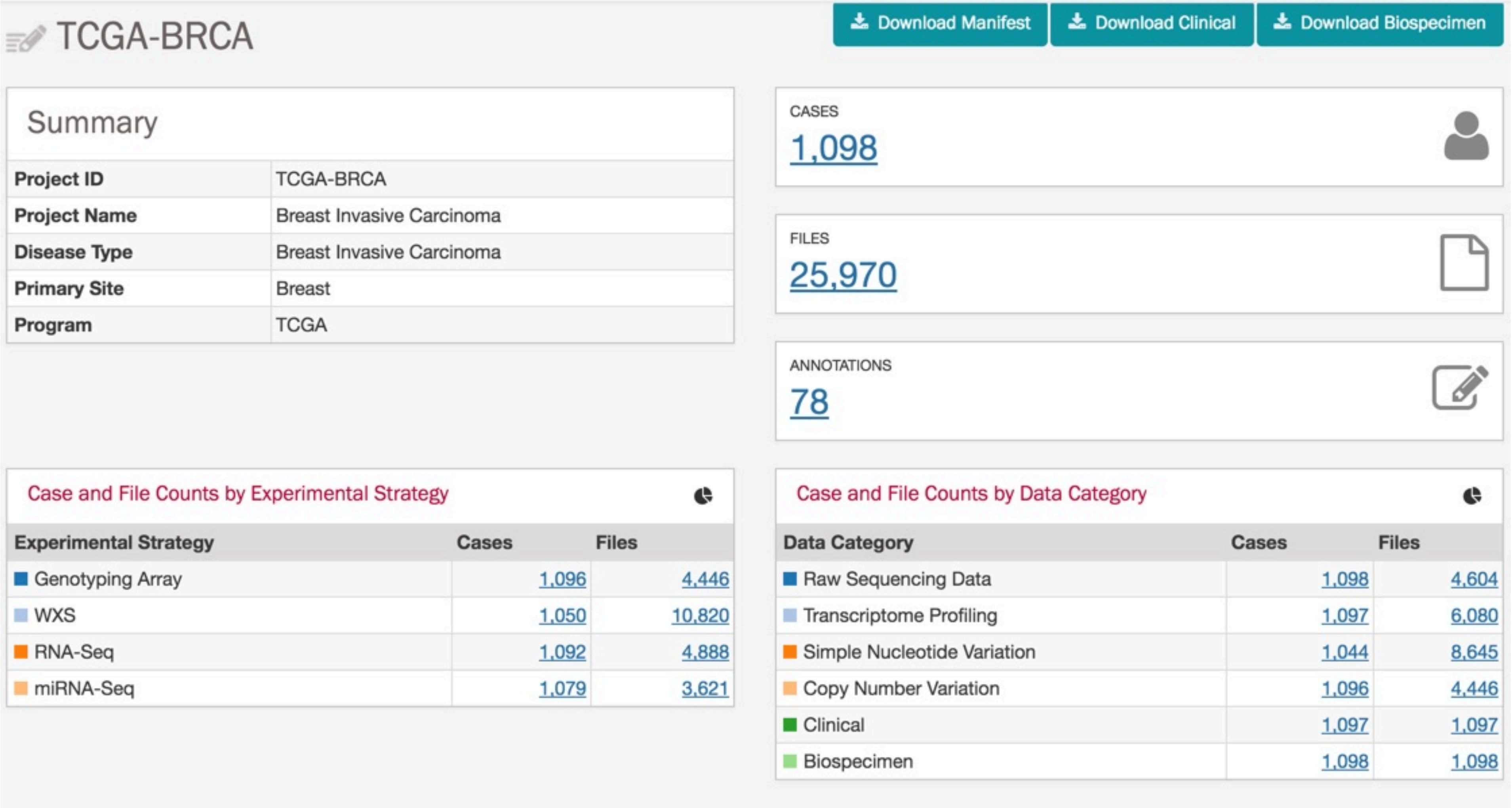
- A product of the NCI Center for Cancer Genomics (CCG)
- Mission: to provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine
- Associated projects: TCGA, Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative, and Cancer Genome Characterization Initiative (CGCI)


GDC Data Transfer Tool
GDC Data Center
GDC Reports
GDC Data Model
GDC Documentation and Support
HELP
GDC Data Portal
GDC Application Programming Interface (API)
GDC Organization and Collaborators
GDC Data Submission Portal
GDC Bioinformatics Pipeline

https://gdc.cancer.gov/

Translational Breast Cancer Research, 2016

GDC: TCGA breast cancer



TCGA-BRCA

Summary

Project ID	TCGA-BRCA
Project Name	Breast Invasive Carcinoma
Disease Type	Breast Invasive Carcinoma
Primary Site	Breast
Program	TCGA

CASES
1,098

FILES
25,970

ANNOTATIONS
78

Case and File Counts by Experimental Strategy

Experimental Strategy	Cases	Files
🟦 Genotyping Array	1,096	4,446
🟦 WXS	1,050	10,820
🟧 RNA-Seq	1,092	4,888
🟧 miRNA-Seq	1,079	3,621

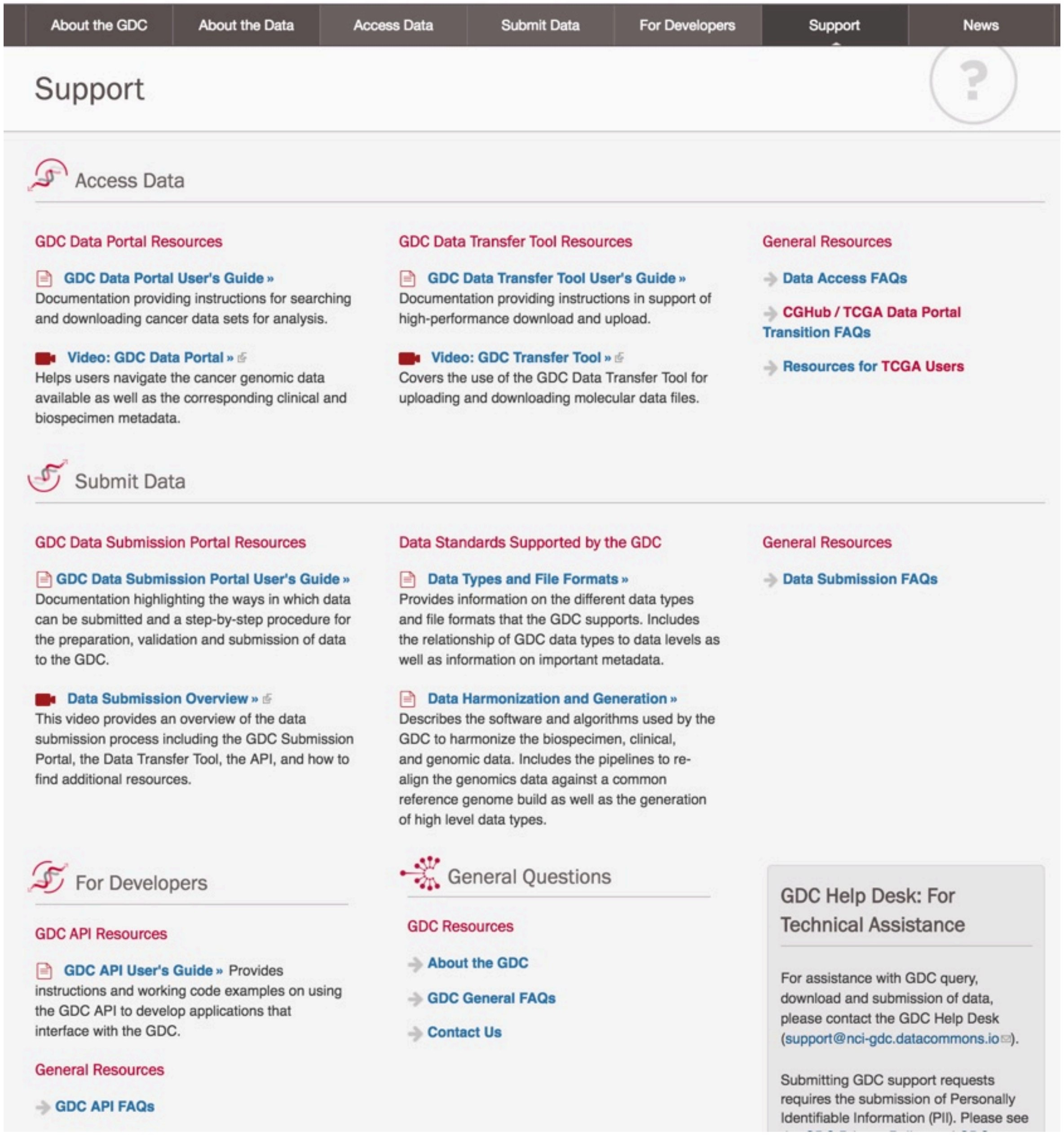
Case and File Counts by Data Category

Data Category	Cases	Files
🟦 Raw Sequencing Data	1,098	4,604
🟦 Transcriptome Profiling	1,097	6,080
🟧 Simple Nucleotide Variation	1,044	8,645
🟧 Copy Number Variation	1,096	4,446
🟩 Clinical	1,097	1,097
🟩 Biospecimen	1,098	1,098

https://gdc-portal.nci.nih.gov/projects/TCGA-BRCA

Translational Breast Cancer Research, 2016

NCI Genomic Data Commons



Support

Access Data

GDC Data Portal Resources

GDC Data Portal User's Guide »
Documentation providing instructions for searching and downloading cancer data sets for analysis.

Video: GDC Data Portal »
Helps users navigate the cancer genomic data available as well as the corresponding clinical and biospecimen metadata.

GDC Data Transfer Tool Resources

GDC Data Transfer Tool User's Guide »
Documentation providing instructions in support of high-performance download and upload.

Video: GDC Transfer Tool »
Covers the use of the GDC Data Transfer Tool for uploading and downloading molecular data files.

General Resources

- **Data Access FAQs**
- **CGHub / TCGA Data Portal Transition FAQs**
- **Resources for TCGA Users**

Submit Data

GDC Data Submission Portal Resources

GDC Data Submission Portal User's Guide »
Documentation highlighting the ways in which data can be submitted and a step-by-step procedure for the preparation, validation and submission of data to the GDC.

Data Submission Overview »
This video provides an overview of the data submission process including the GDC Submission Portal, the Data Transfer Tool, the API, and how to find additional resources.

Data Standards Supported by the GDC

Data Types and File Formats »
Provides information on the different data types and file formats that the GDC supports. Includes the relationship of GDC data types to data levels as well as information on important metadata.

Data Harmonization and Generation »
Describes the software and algorithms used by the GDC to harmonize the biospecimen, clinical, and genomic data. Includes the pipelines to re-align the genomics data against a common reference genome build as well as the generation of high level data types.

General Resources

- **Data Submission FAQs**

For Developers

GDC API Resources

GDC API User's Guide » Provides instructions and working code examples on using the GDC API to develop applications that interface with the GDC.

General Resources

- **GDC API FAQs**

General Questions

GDC Resources

- **About the GDC**
- **GDC General FAQs**
- **Contact Us**

GDC Help Desk: For Technical Assistance

For assistance with GDC query, download and submission of data, please contact the GDC Help Desk (support@nci-gdc.datacommons.io).

Submitting GDC support requests requires the submission of Personally Identifiable Information (PII). Please see

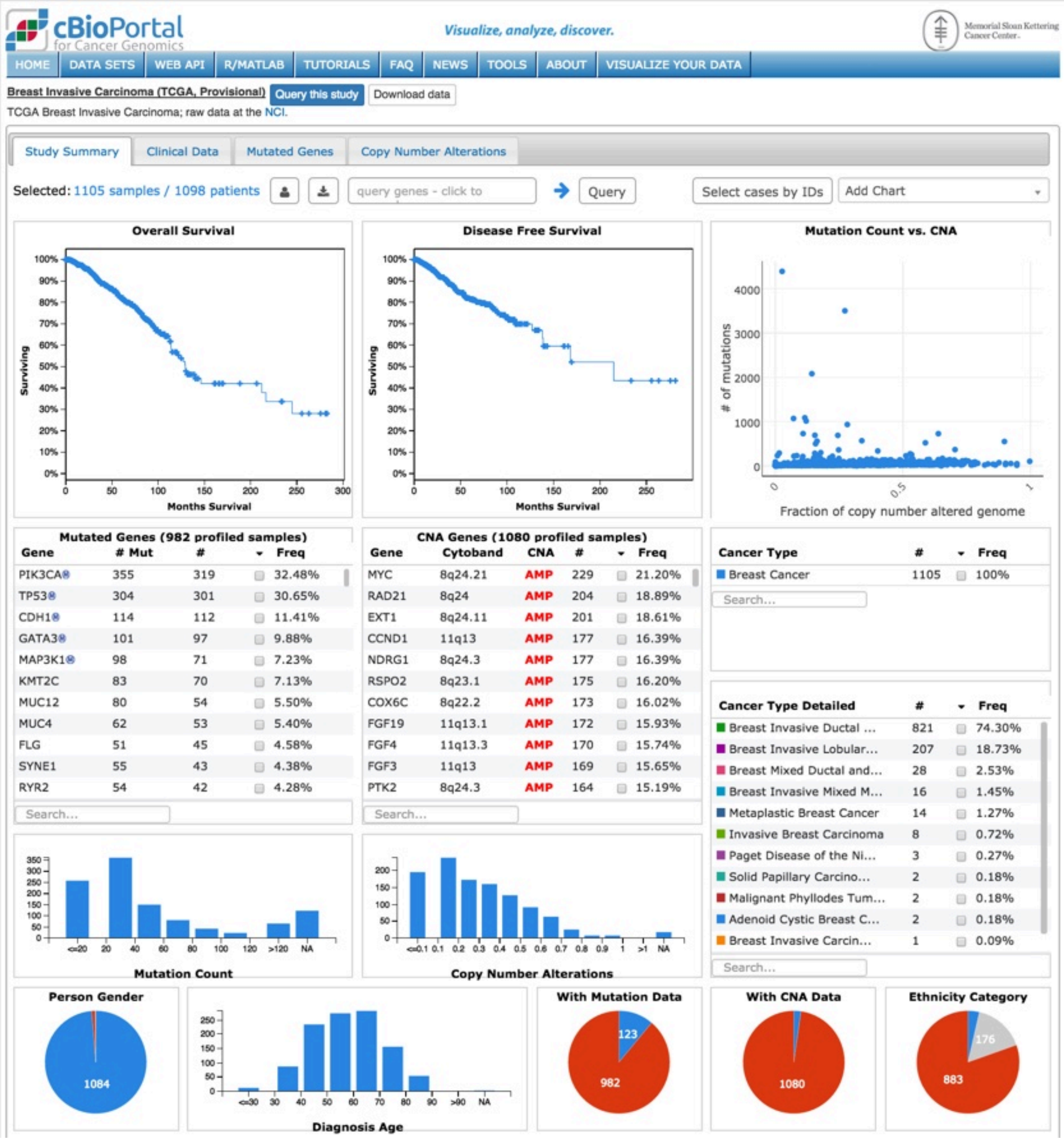
https://gdc.cancer.gov/

Translational Breast Cancer Research, 2016

cBioPortal for Cancer Genomics

- http://www.cbioportal.org/
- Visualization, analysis and download of large-scale cancer genomics data sets
- 147 cancer genomics studies as of October 2016
- References
  - Gao et al., Sci Signal 2013
  - Cerami et al. Cancer Discov, 2012

cBioPortal: TCGA breast cancer overview



Breast Invasive Carcinoma (TCGA, Provisional)

TCGA Breast Invasive Carcinoma; raw data at the NCI.

Selected: 1105 samples / 1098 patients

Mutated Genes (982 profiled samples)

Gene	# Mut	#	Freq
PIK3CA	355	319	□ 32.48%
TP53	304	301	□ 30.65%
CDH1	114	112	□ 11.41%
GATA3	101	97	□ 9.88%
MAP3K1	98	71	□ 7.23%
KMT2C	83	70	□ 7.13%
MUC12	80	54	□ 5.50%
MUC4	62	53	□ 5.40%
FLG	51	45	□ 4.58%
SYNE1	55	43	□ 4.38%
RYR2	54	42	□ 4.28%

CNA Genes (1080 profiled samples)

Gene	Cytoband	CNA	#	Freq
MYC	8q24.21	AMP	229	□ 21.20%
RAD21	8q24	AMP	204	□ 18.89%
EXT1	8q24.11	AMP	201	□ 18.61%
CCND1	11q13	AMP	177	□ 16.39%
NDRG1	8q24.3	AMP	177	□ 16.39%
RSPO2	8q23.1	AMP	175	□ 16.20%
COX6C	8q22.2	AMP	173	□ 16.02%
FGF19	11q13.1	AMP	172	□ 15.93%
FGF4	11q13.3	AMP	170	□ 15.74%
FGF3	11q13	AMP	169	□ 15.65%
PTK2	8q24.3	AMP	164	□ 15.19%

Cancer Type	#	Freq
Breast Cancer	1105	□ 100%

Cancer Type Detailed	#	Freq
Breast Invasive Ductal ...	821	□ 74.30%
Breast Invasive Lobular...	207	□ 18.73%
Breast Mixed Ductal and...	28	□ 2.53%
Breast Invasive Mixed M...	16	□ 1.45%
Metaplastic Breast Cancer	14	□ 1.27%
Invasive Breast Carcinoma	8	□ 0.72%
Paget Disease of the Ni...	3	□ 0.27%
Solid Papillary Carcino...	2	□ 0.18%
Malignant Phyllodes Tum...	2	□ 0.18%
Adenoid Cystic Breast C...	2	□ 0.18%
Breast Invasive Carcin...	1	□ 0.09%

Translational Breast Cancer Research, 2016

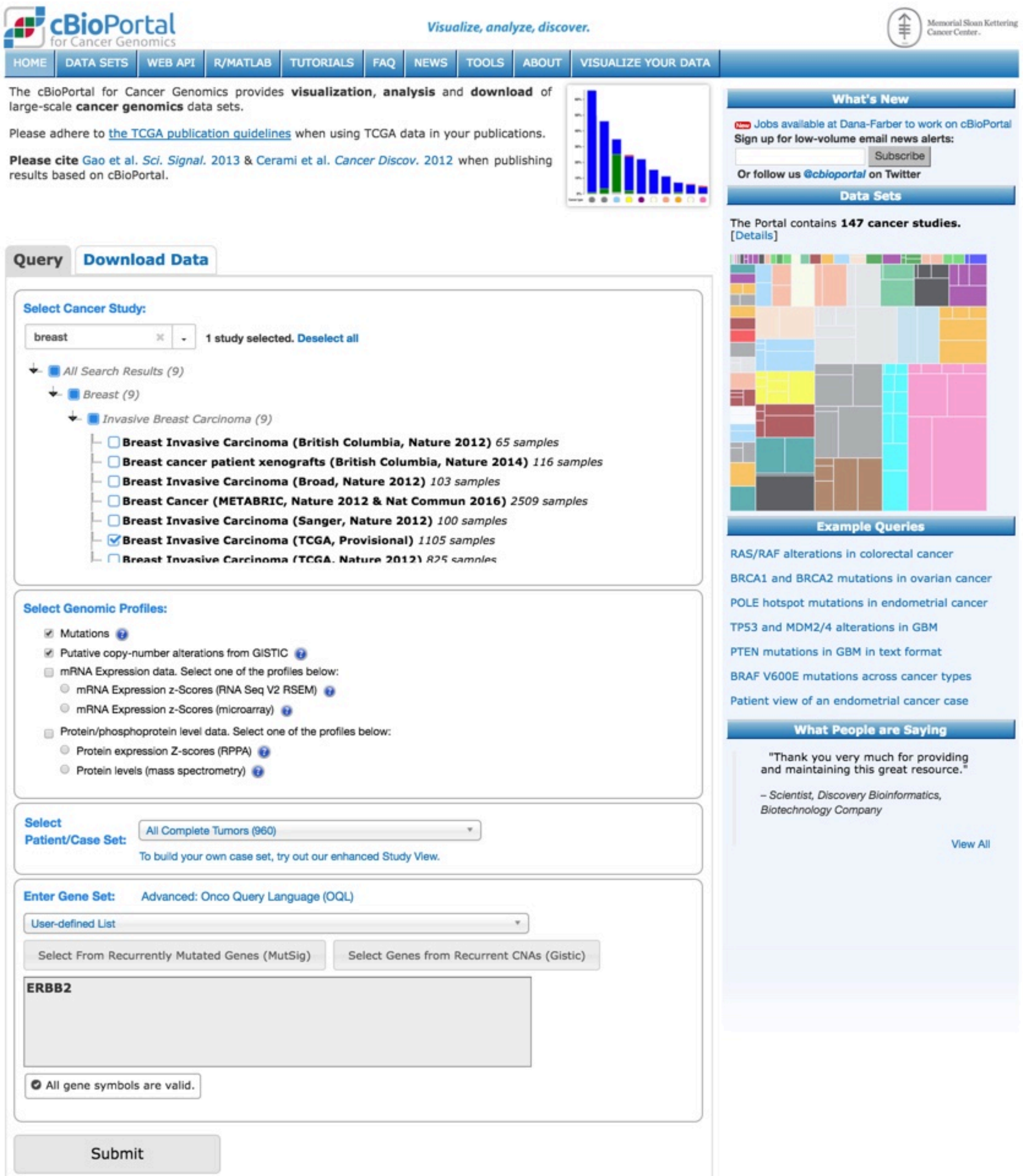
cBioPortal: query interface

Select cancer study

Select genomic profiles

Select patient/case set

Enter gene set



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.

Select Cancer Study:

breast ✕ ⌄ 1 study selected. Deselect all

- ■ *All Search Results (9)*
  - ■ *Breast (9)*
    - ■ *Invasive Breast Carcinoma (9)*
      - □ **Breast Invasive Carcinoma (British Columbia, Nature 2012)** *65 samples*
      - □ **Breast cancer patient xenografts (British Columbia, Nature 2014)** *116 samples*
      - □ **Breast Invasive Carcinoma (Broad, Nature 2012)** *103 samples*
      - □ **Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)** *2509 samples*
      - □ **Breast Invasive Carcinoma (Sanger, Nature 2012)** *100 samples*
      - ☑ **Breast Invasive Carcinoma (TCGA, Provisional)** *1105 samples*
      - □ **Breast Invasive Carcinoma (TCGA, Nature 2012)** *825 samples*

Select Genomic Profiles:

- ☑ Mutations
- ☑ Putative copy-number alterations from GISTIC
- □ mRNA Expression data. Select one of the profiles below:
  - ○ mRNA Expression z-Scores (RNA Seq V2 RSEM)
  - ○ mRNA Expression z-Scores (microarray)
- □ Protein/phosphoprotein level data. Select one of the profiles below:
  - ○ Protein expression Z-scores (RPPA)
  - ○ Protein levels (mass spectrometry)

Select Patient/Case Set: All Complete Tumors (960)

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

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

User-defined List

Select From Recurrently Mutated Genes (MutSig)
Select Genes from Recurrent CNAs (Gistic)

ERBB2

✓ All gene symbols are valid.

What's New

New Jobs available at Dana-Farber to work on cBioPortal

Sign up for low-volume email news alerts:

Or follow us *@cbioportal* on Twitter

Data Sets

The Portal contains **147 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

"Thank you very much for providing and maintaining this great resource."

– Scientist, Discovery Bioinformatics, Biotechnology Company

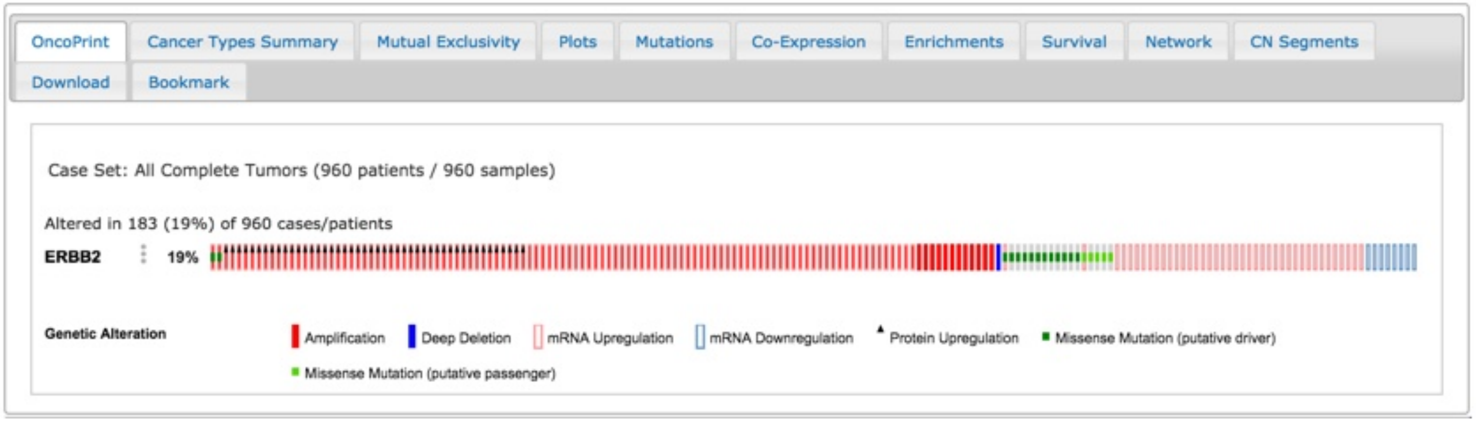
View All

Translational Breast Cancer Research, 2016

cBioPortal: oncoprint

- Compact visualization of distinct genomic alterations, including somatic mutations, copy number alterations, and gene expression changes across a set of cases.

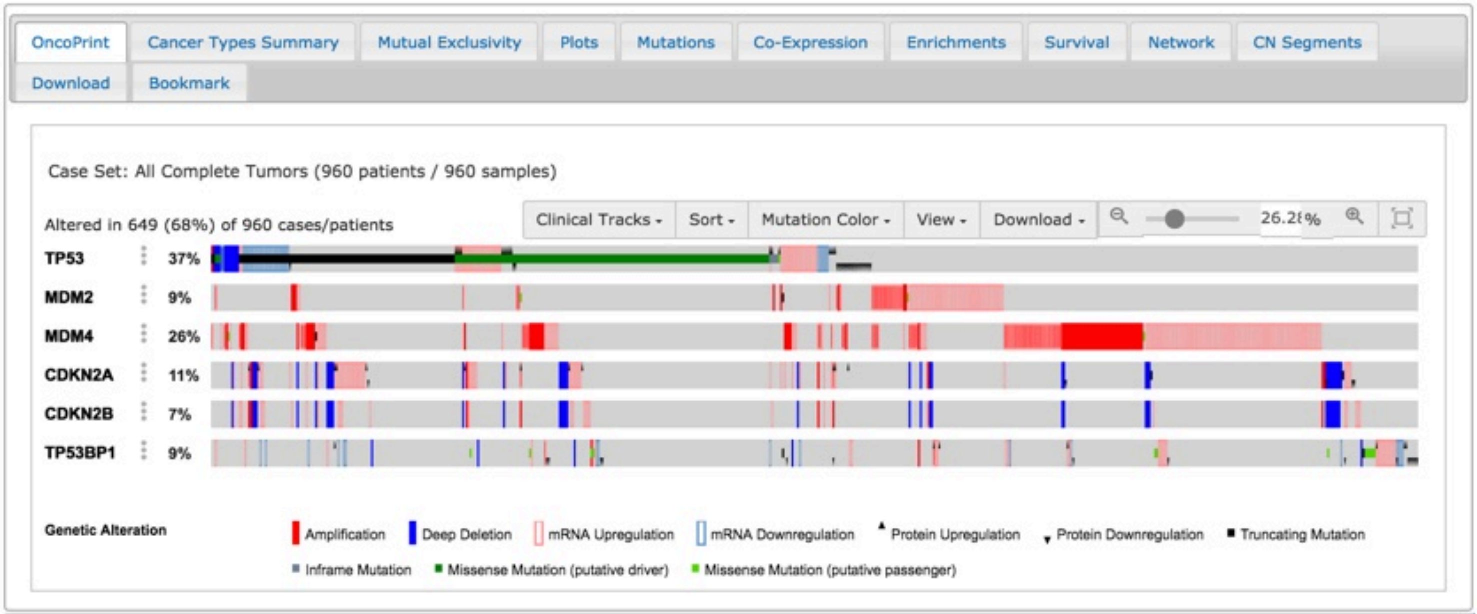
ERBB2



Case Set: All Complete Tumors (960 patients / 960 samples)

Altered in 183 (19%) of 960 cases/patients

p53 signaling

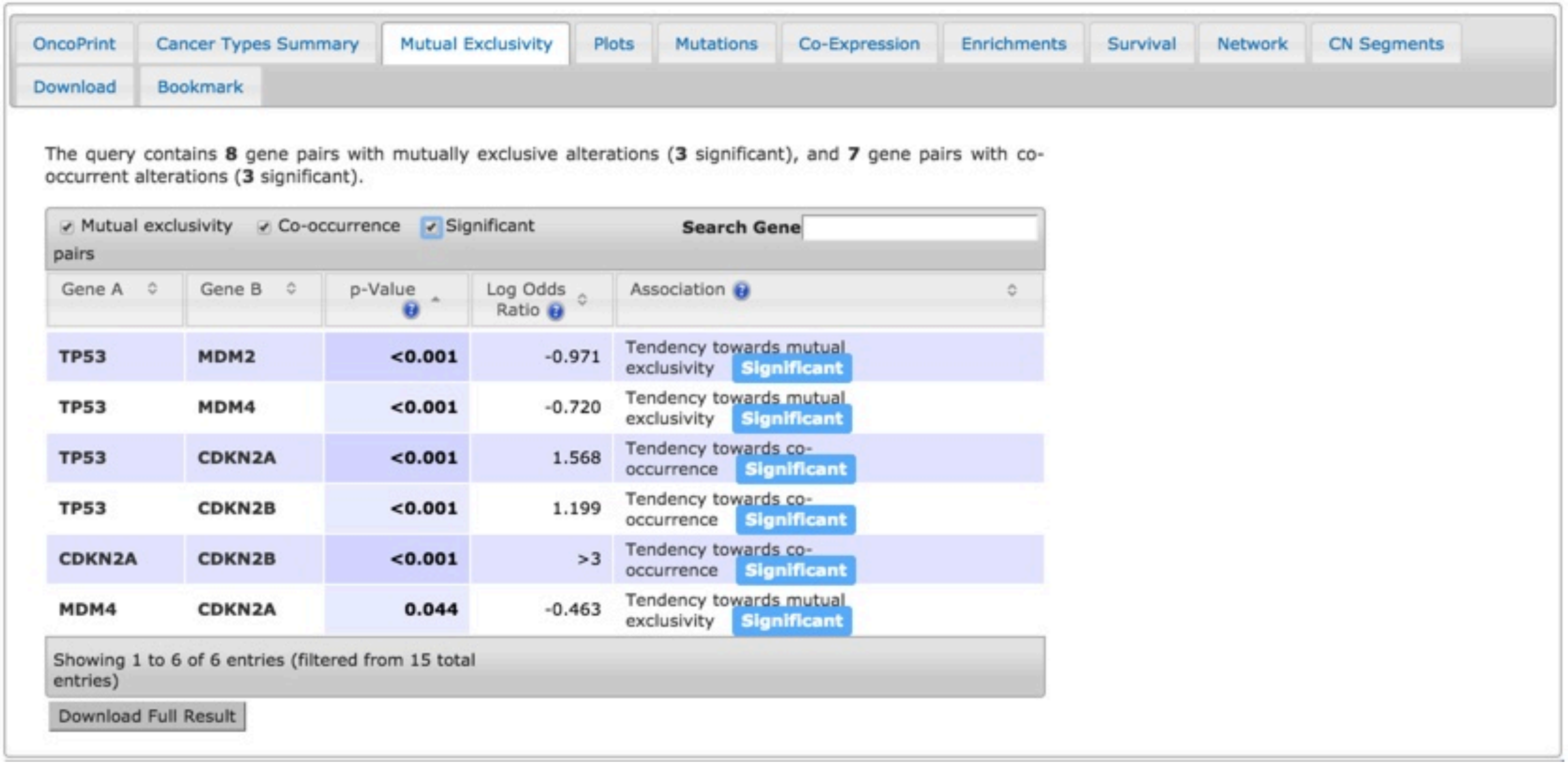


Case Set: All Complete Tumors (960 patients / 960 samples)

Altered in 649 (68%) of 960 cases/patients

Translational Breast Cancer Research, 2016

cBioPortal: mutual exclusivity



The query contains **8** gene pairs with mutually exclusive alterations (**3** significant), and **7** gene pairs with co-occurrent alterations (**3** significant).

☑ Mutual exclusivity ☑ Co-occurrence ☑ Significant pairs **Search Gene**

Gene A	Gene B	p-Value	Log Odds Ratio	Association
TP53	**MDM2**	**<0.001**	-0.971	Tendency towards mutual exclusivity **Significant**
TP53	**MDM4**	**<0.001**	-0.720	Tendency towards mutual exclusivity **Significant**
TP53	**CDKN2A**	**<0.001**	1.568	Tendency towards co-occurrence **Significant**
TP53	**CDKN2B**	**<0.001**	1.199	Tendency towards co-occurrence **Significant**
CDKN2A	**CDKN2B**	**<0.001**	>3	Tendency towards co-occurrence **Significant**
MDM4	**CDKN2A**	**0.044**	-0.463	Tendency towards mutual exclusivity **Significant**

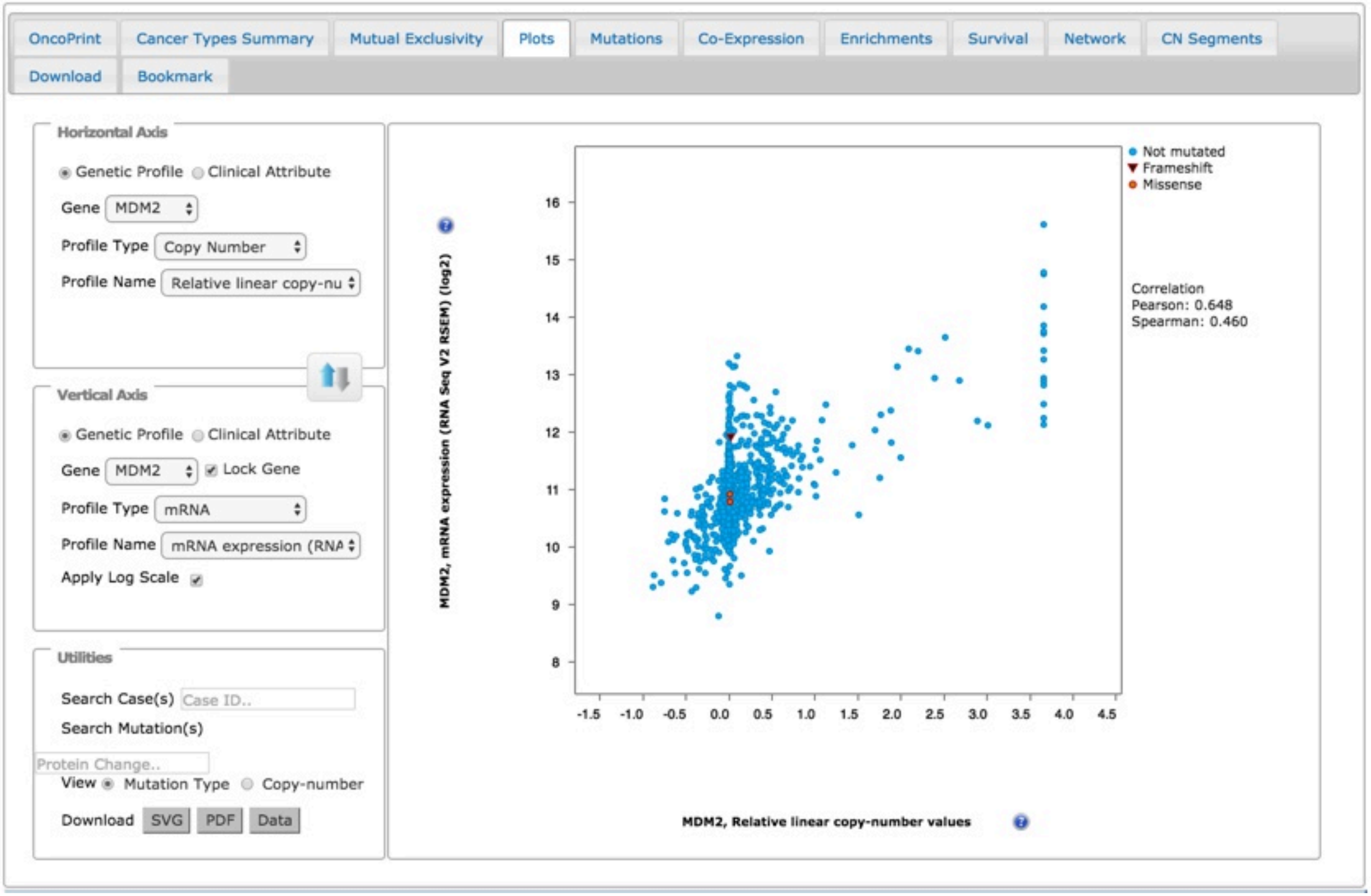
Showing 1 to 6 of 6 entries (filtered from 15 total entries)

Mutual exclusivity => functional link

- Alteration to the second gene within the same pathway offers no further selective advantage
- Alteration to the second gene within the same pathway leads to a disadvantage for the cell, *i.e.*, synthetic lethality.

Ciriello et al., Genome Res, 2012

cBioPortal: copy number vs mRNA expression



Horizontal Axis

● Genetic Profile ○ Clinical Attribute

Gene: MDM2

Profile Type: Copy Number

Profile Name: Relative linear copy-nu

Vertical Axis

● Genetic Profile ○ Clinical Attribute

Gene: MDM2 ☑ Lock Gene

Profile Type: mRNA

Profile Name: mRNA expression (RNA

Apply Log Scale ☑

Utilities

Search Case(s): Case ID..

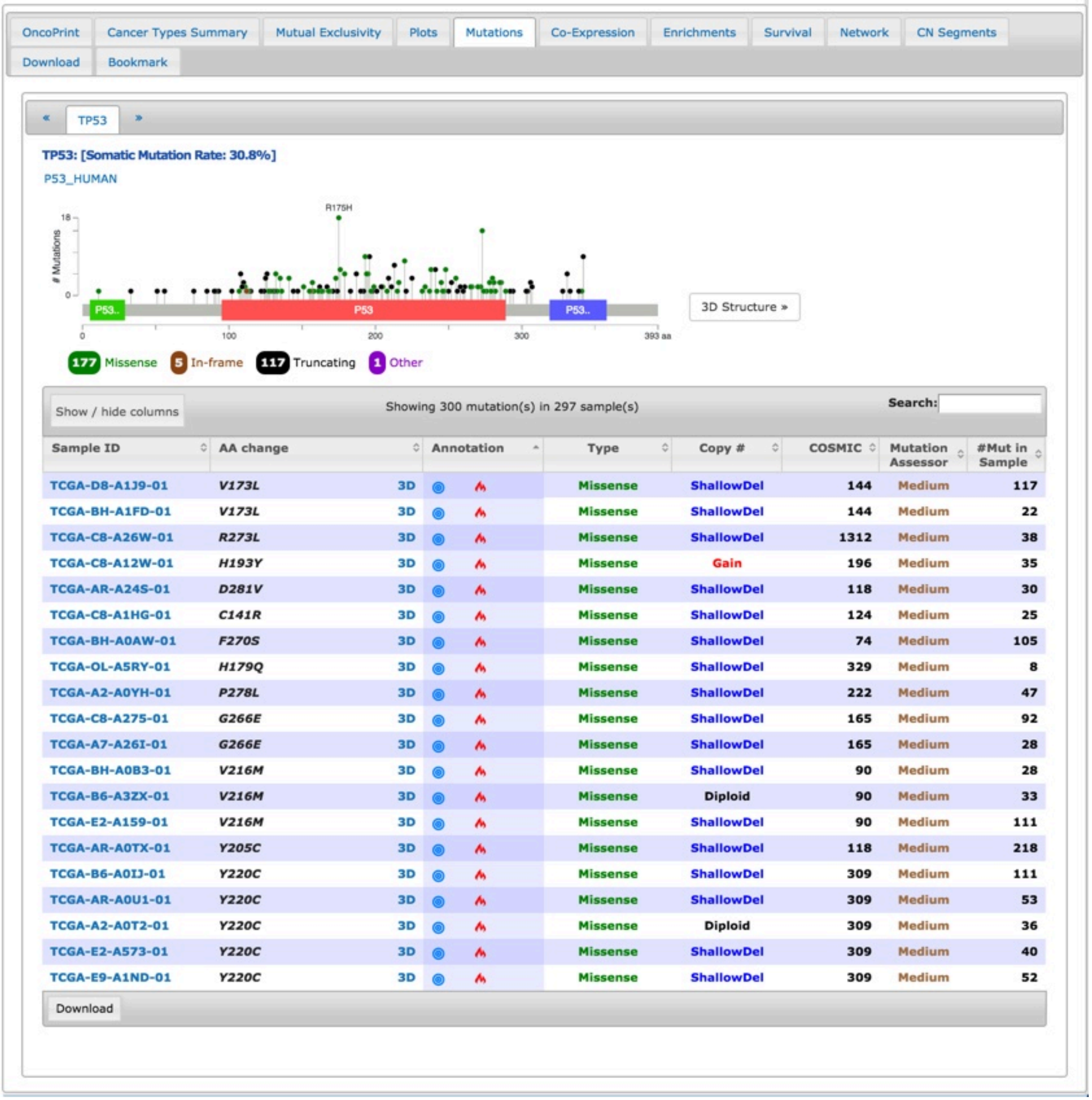
Search Mutation(s): Protein Change..

View ● Mutation Type ○ Copy-number

Correlation
Pearson: 0.648
Spearman: 0.460

Translational Breast Cancer Research, 2016

cBioPortal: mutations



TP53: [Somatic Mutation Rate: 30.8%]

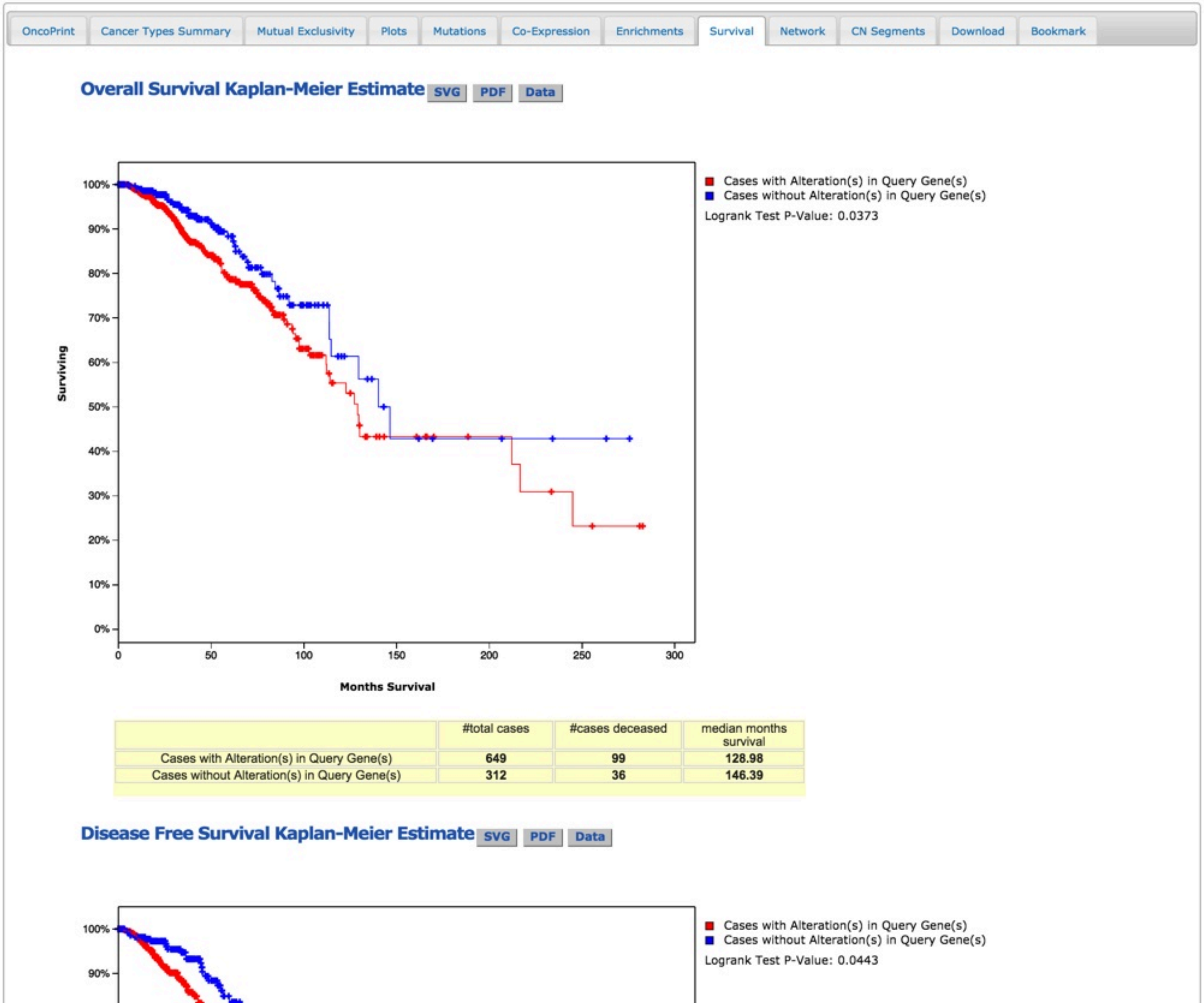
P53_HUMAN

177 Missense 5 In-frame 117 Truncating 1 Other

Showing 300 mutation(s) in 297 sample(s)

Sample ID	AA change	Annotation	Type	Copy #	COSMIC	Mutation Assessor	#Mut in Sample
TCGA-D8-A1J9-01	V173L 3D	◉ 🔥	Missense	ShallowDel	144	Medium	117
TCGA-BH-A1FD-01	V173L 3D	◉ 🔥	Missense	ShallowDel	144	Medium	22
TCGA-C8-A26W-01	R273L 3D	◉ 🔥	Missense	ShallowDel	1312	Medium	38
TCGA-C8-A12W-01	H193Y 3D	◉ 🔥	Missense	Gain	196	Medium	35
TCGA-AR-A24S-01	D281V 3D	◉ 🔥	Missense	ShallowDel	118	Medium	30
TCGA-C8-A1HG-01	C141R 3D	◉ 🔥	Missense	ShallowDel	124	Medium	25
TCGA-BH-A0AW-01	F270S 3D	◉ 🔥	Missense	ShallowDel	74	Medium	105
TCGA-OL-A5RY-01	H179Q 3D	◉ 🔥	Missense	ShallowDel	329	Medium	8
TCGA-A2-A0YH-01	P278L 3D	◉ 🔥	Missense	ShallowDel	222	Medium	47
TCGA-C8-A275-01	G266E 3D	◉ 🔥	Missense	ShallowDel	165	Medium	92
TCGA-A7-A26I-01	G266E 3D	◉ 🔥	Missense	ShallowDel	165	Medium	28
TCGA-BH-A0B3-01	V216M 3D	◉ 🔥	Missense	ShallowDel	90	Medium	28
TCGA-B6-A3ZX-01	V216M 3D	◉ 🔥	Missense	Diploid	90	Medium	33
TCGA-E2-A159-01	V216M 3D	◉ 🔥	Missense	ShallowDel	90	Medium	111
TCGA-AR-A0TX-01	Y205C 3D	◉ 🔥	Missense	ShallowDel	118	Medium	218
TCGA-B6-A0IJ-01	Y220C 3D	◉ 🔥	Missense	ShallowDel	309	Medium	111
TCGA-AR-A0U1-01	Y220C 3D	◉ 🔥	Missense	ShallowDel	309	Medium	53
TCGA-A2-A0T2-01	Y220C 3D	◉ 🔥	Missense	Diploid	309	Medium	36
TCGA-E2-A573-01	Y220C 3D	◉ 🔥	Missense	ShallowDel	309	Medium	40
TCGA-E9-A1ND-01	Y220C 3D	◉ 🔥	Missense	ShallowDel	309	Medium	52

cBioPortal: survival



Overall Survival Kaplan-Meier Estimate

🟥 Cases with Alteration(s) in Query Gene(s)
🟦 Cases without Alteration(s) in Query Gene(s)
Logrank Test P-Value: 0.0373

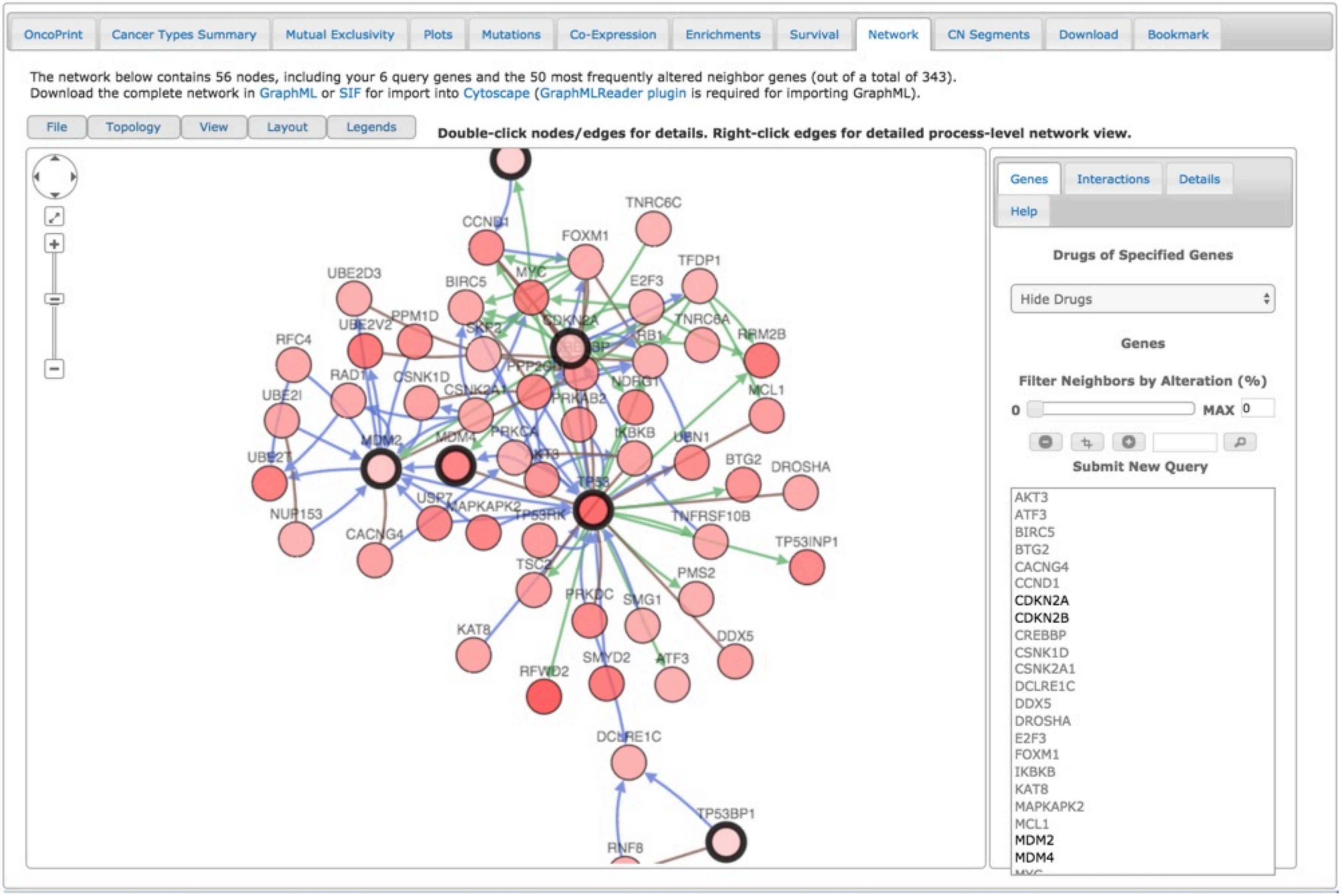
	#total cases	#cases deceased	median months survival
Cases with Alteration(s) in Query Gene(s)	649	99	128.98
Cases without Alteration(s) in Query Gene(s)	312	36	146.39

Disease Free Survival Kaplan-Meier Estimate

🟥 Cases with Alteration(s) in Query Gene(s)
🟦 Cases without Alteration(s) in Query Gene(s)
Logrank Test P-Value: 0.0443

Translational Breast Cancer Research, 2016

cBioPortal: exploring the interactome



The network below contains 56 nodes, including your 6 query genes and the 50 most frequently altered neighbor genes (out of a total of 343).
Download the complete network in GraphML or SIF for import into Cytoscape (GraphMLReader plugin is required for importing GraphML).

Double-click nodes/edges for details. Right-click edges for detailed process-level network view.

Drugs of Specified Genes

Hide Drugs

Genes

Filter Neighbors by Alteration (%)

0 MAX 0

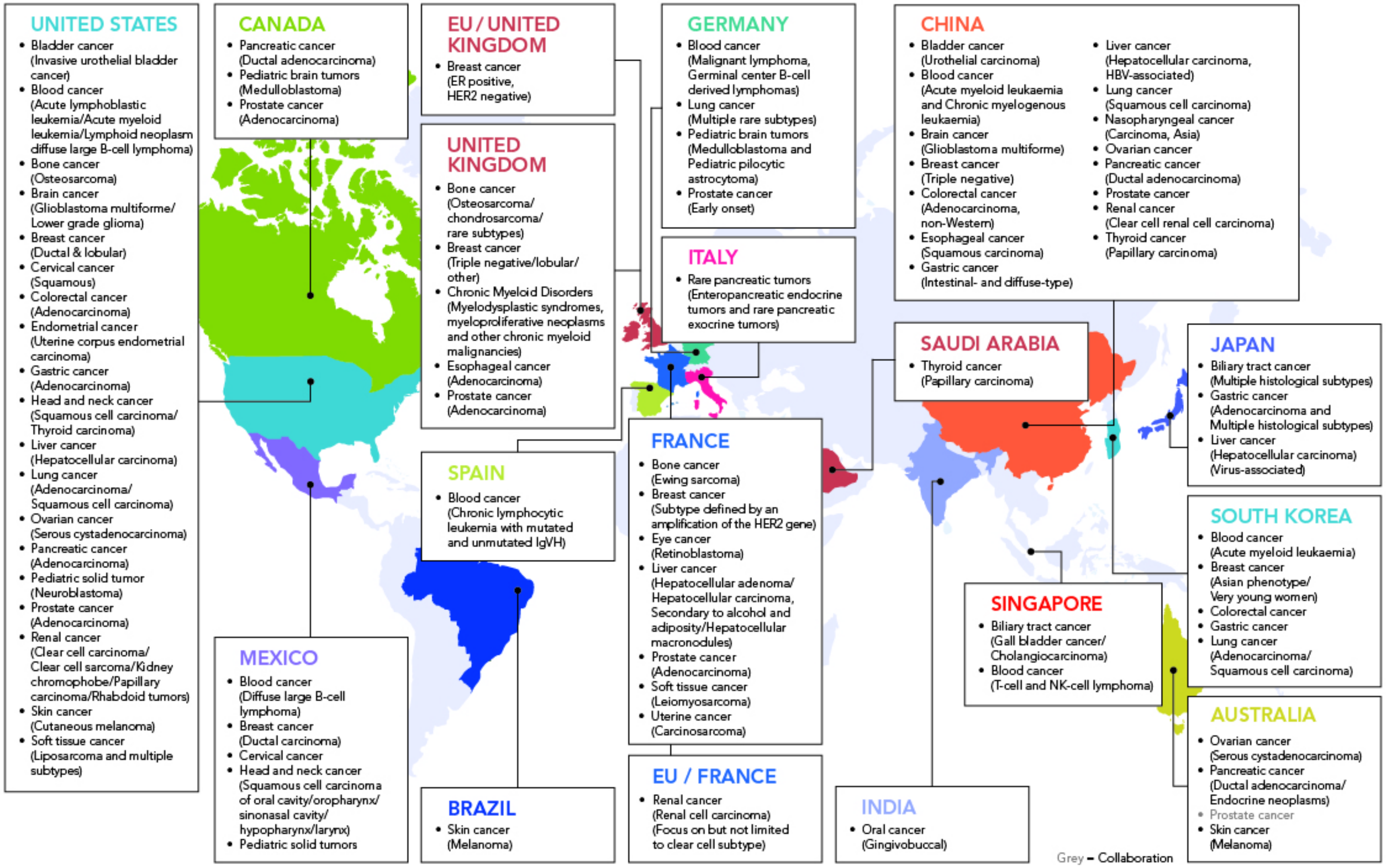
Submit New Query

AKT3
ATF3
BIRC5
BTG2
CACNG4
CCND1
CDKN2A
CDKN2B
CREBBP
CSNK1D
CSNK2A1
DCLRE1C
DDX5
DROSHA
E2F3
FOXM1
IKBKB
KAT8
MAPKAPK2
MCL1
MDM2
MDM4

Translational Breast Cancer Research, 2016

International Cancer Genome Consortium (ICGC)

- To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.



UNITED STATES

- Bladder cancer (Invasive urothelial bladder cancer)
- Blood cancer (Acute lymphoblastic leukemia/Acute myeloid leukemia/Lymphoid neoplasm diffuse large B-cell lymphoma)
- Bone cancer (Osteosarcoma)
- Brain cancer (Glioblastoma multiforme/ Lower grade glioma)
- Breast cancer (Ductal & lobular)
- Cervical cancer (Squamous)
- Colorectal cancer (Adenocarcinoma)
- Endometrial cancer (Uterine corpus endometrial carcinoma)
- Gastric cancer (Adenocarcinoma)
- Head and neck cancer (Squamous cell carcinoma/ Thyroid carcinoma)
- Liver cancer (Hepatocellular carcinoma)
- Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Adenocarcinoma)
- Pediatric solid tumor (Neuroblastoma)
- Prostate cancer (Adenocarcinoma)
- Renal cancer (Clear cell carcinoma/ Clear cell sarcoma/Kidney chromophobe/Papillary carcinoma/Rhabdoid tumors)
- Skin cancer (Cutaneous melanoma)
- Soft tissue cancer (Liposarcoma and multiple subtypes)

CANADA

- Pancreatic cancer (Ductal adenocarcinoma)
- Pediatric brain tumors (Medulloblastoma)
- Prostate cancer (Adenocarcinoma)

MEXICO

- Blood cancer (Diffuse large B-cell lymphoma)
- Breast cancer (Ductal carcinoma)
- Cervical cancer
- Head and neck cancer (Squamous cell carcinoma of oral cavity/oropharynx/ sinonasal cavity/ hypopharynx/larynx)
- Pediatric solid tumors

EU / UNITED KINGDOM

- Breast cancer (ER positive, HER2 negative)

UNITED KINGDOM

- Bone cancer (Osteosarcoma/ chondrosarcoma/ rare subtypes)
- Breast cancer (Triple negative/lobular/ other)
- Chronic Myeloid Disorders (Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies)
- Esophageal cancer (Adenocarcinoma)
- Prostate cancer (Adenocarcinoma)

SPAIN

- Blood cancer (Chronic lymphocytic leukemia with mutated and unmutated IgVH)

BRAZIL

- Skin cancer (Melanoma)

GERMANY

- Blood cancer (Malignant lymphoma, Germinal center B-cell derived lymphomas)
- Lung cancer (Multiple rare subtypes)
- Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
- Prostate cancer (Early onset)

ITALY

- Rare pancreatic tumors (Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors)

FRANCE

- Bone cancer (Ewing sarcoma)
- Breast cancer (Subtype defined by an amplification of the HER2 gene)
- Eye cancer (Retinoblastoma)
- Liver cancer (Hepatocellular adenoma/ Hepatocellular carcinoma, Secondary to alcohol and adiposity/Hepatocellular macronodules)
- Prostate cancer (Adenocarcinoma)
- Soft tissue cancer (Leiomyosarcoma)
- Uterine cancer (Carcinosarcoma)

EU / FRANCE

- Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)

CHINA

- Bladder cancer (Urothelial carcinoma)
- Blood cancer (Acute myeloid leukaemia and Chronic myelogenous leukaemia)
- Brain cancer (Glioblastoma multiforme)
- Breast cancer (Triple negative)
- Colorectal cancer (Adenocarcinoma, non-Western)
- Esophageal cancer (Squamous carcinoma)
- Gastric cancer (Intestinal- and diffuse-type)
- Liver cancer (Hepatocellular carcinoma, HBV-associated)
- Lung cancer (Squamous cell carcinoma)
- Nasopharyngeal cancer (Carcinoma, Asia)
- Ovarian cancer
- Pancreatic cancer (Ductal adenocarcinoma)
- Prostate cancer
- Renal cancer (Clear cell renal cell carcinoma)
- Thyroid cancer (Papillary carcinoma)

SAUDI ARABIA

- Thyroid cancer (Papillary carcinoma)

INDIA

- Oral cancer (Gingivobuccal)

SINGAPORE

- Biliary tract cancer (Gall bladder cancer/ Cholangiocarcinoma)
- Blood cancer (T-cell and NK-cell lymphoma)

JAPAN

- Biliary tract cancer (Multiple histological subtypes)
- Gastric cancer (Adenocarcinoma and Multiple histological subtypes)
- Liver cancer (Hepatocellular carcinoma) (Virus-associated)

SOUTH KOREA

- Blood cancer (Acute myeloid leukaemia)
- Breast cancer (Asian phenotype/ Very young women)
- Colorectal cancer
- Gastric cancer
- Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)

AUSTRALIA

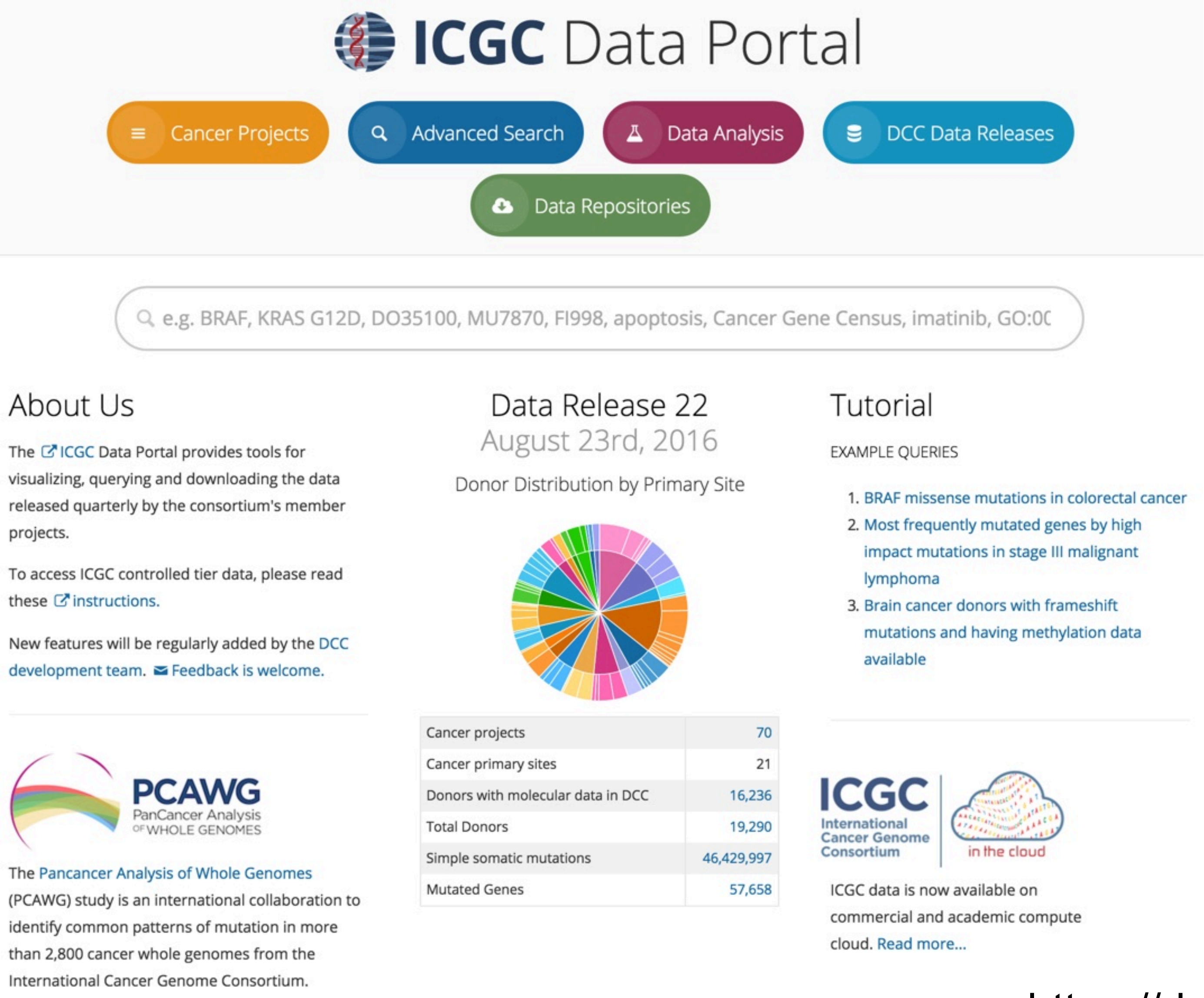
- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Ductal adenocarcinoma/ Endocrine neoplasms)
- Prostate cancer
- Skin cancer (Melanoma)

Grey – Collaboration

Translational Breast Cancer Research, 2016

http://www.icgc.org/

ICGC Data Portal



About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these instructions.

New features will be regularly added by the DCC development team. ✉ Feedback is welcome.

PCAWG PanCancer Analysis OF WHOLE GENOMES

The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium.

Data Release 22

August 23rd, 2016

Donor Distribution by Primary Site

Cancer projects	70
Cancer primary sites	21
Donors with molecular data in DCC	16,236
Total Donors	19,290
Simple somatic mutations	46,429,997
Mutated Genes	57,658

Tutorial

EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

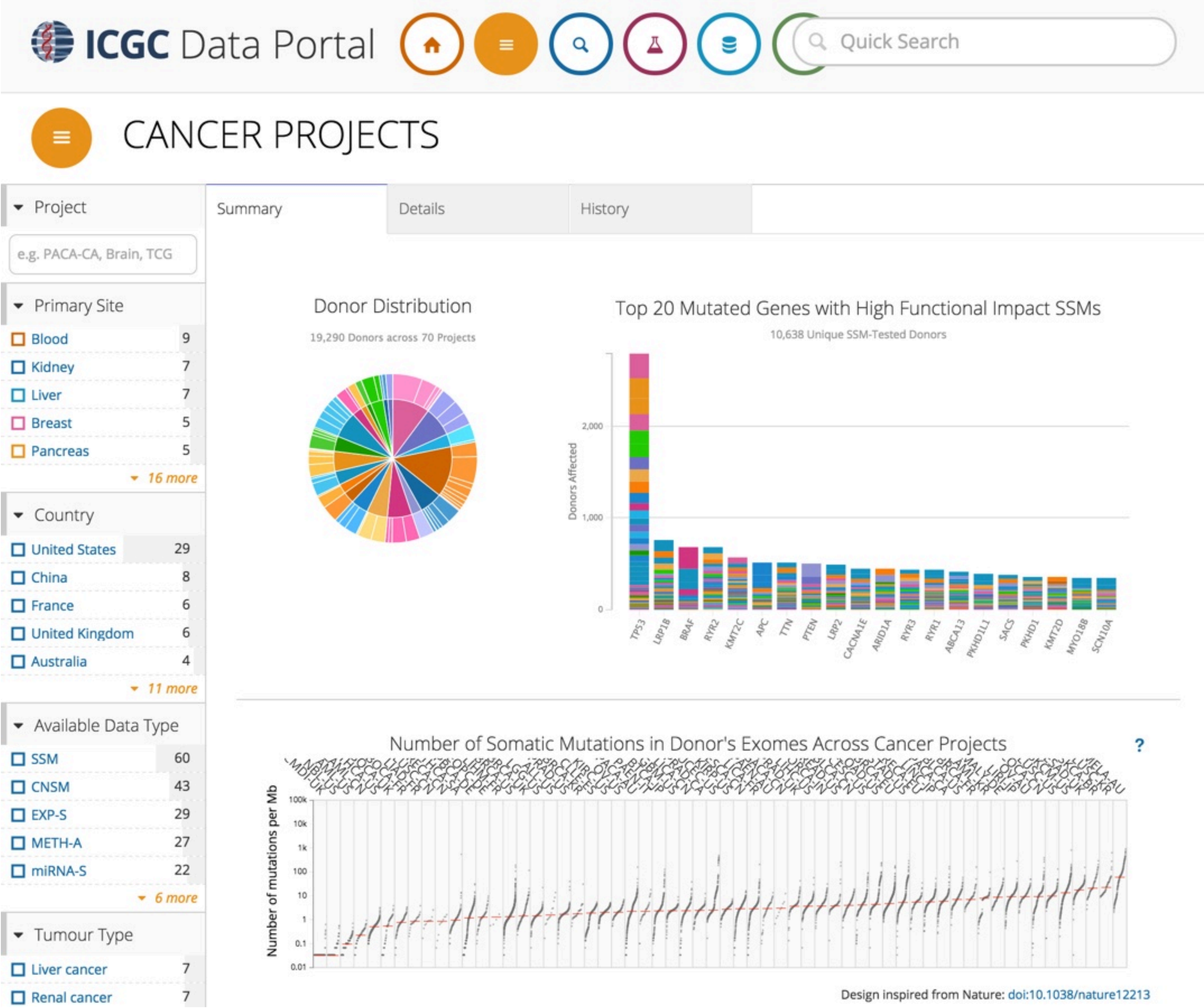
ICGC International Cancer Genome Consortium | in the cloud

ICGC data is now available on commercial and academic compute cloud. Read more...

https://dcc.icgc.org/

Translational Breast Cancer Research, 2016

ICGC Data Portal: cancer projects



ICGC Data Portal

Quick Search

CANCER PROJECTS

Summary | Details | History

Project

e.g. PACA-CA, Brain, TCG

Primary Site

□ Blood	9
□ Kidney	7
□ Liver	7
□ Breast	5
□ Pancreas	5

16 more

Country

□ United States	29
□ China	8
□ France	6
□ United Kingdom	6
□ Australia	4

11 more

Available Data Type

□ SSM	60
□ CNSM	43
□ EXP-S	29
□ METH-A	27
□ miRNA-S	22

6 more

Tumour Type

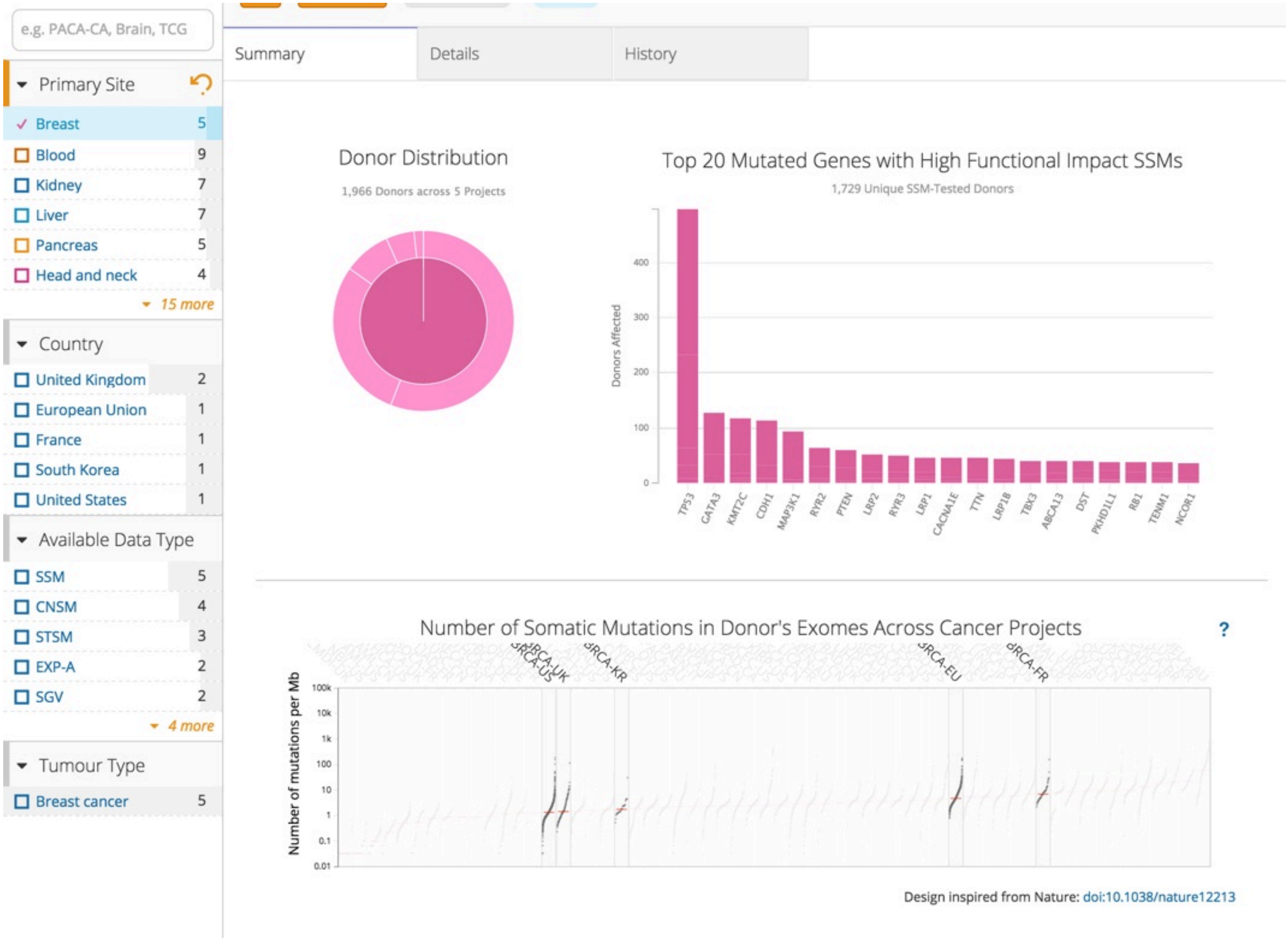
□ Liver cancer	7
□ Renal cancer	7

Donor Distribution
19,290 Donors across 70 Projects
Top 20 Mutated Genes with High Functional Impact SSMs
10,638 Unique SSM-Tested Donors
Donors Affected
2,000
1,000
0
TP53
LRP1B
BRAF
RYR2
KMT2C
APC
TTN
PTEN
LRP2
CACNA1E
ARID1A
RYR3
RYR1
ABCA13
PKHD1L1
SACS
PKHD1
KMT2D
MYO18B
SCN10A
Number of Somatic Mutations in Donor's Exomes Across Cancer Projects
?
Number of mutations per Mb
100k
10k
1k
100
10
1
0.1
0.01

Design inspired from Nature: doi:10.1038/nature12213

Translational Breast Cancer Research, 2016

ICGC Data Portal: breast cancer



e.g. PACA-CA, Brain, TCG

Summary | Details | History

Primary Site	
✓ Breast	5
□ Blood	9
□ Kidney	7
□ Liver	7
□ Pancreas	5
□ Head and neck	4

15 more

Country	
□ United Kingdom	2
□ European Union	1
□ France	1
□ South Korea	1
□ United States	1

Available Data Type	
□ SSM	5
□ CNSM	4
□ STSM	3
□ EXP-A	2
□ SGV	2

4 more

Tumour Type	
□ Breast cancer	5

Donor Distribution

1,966 Donors across 5 Projects

Top 20 Mutated Genes with High Functional Impact SSMs

1,729 Unique SSM-Tested Donors

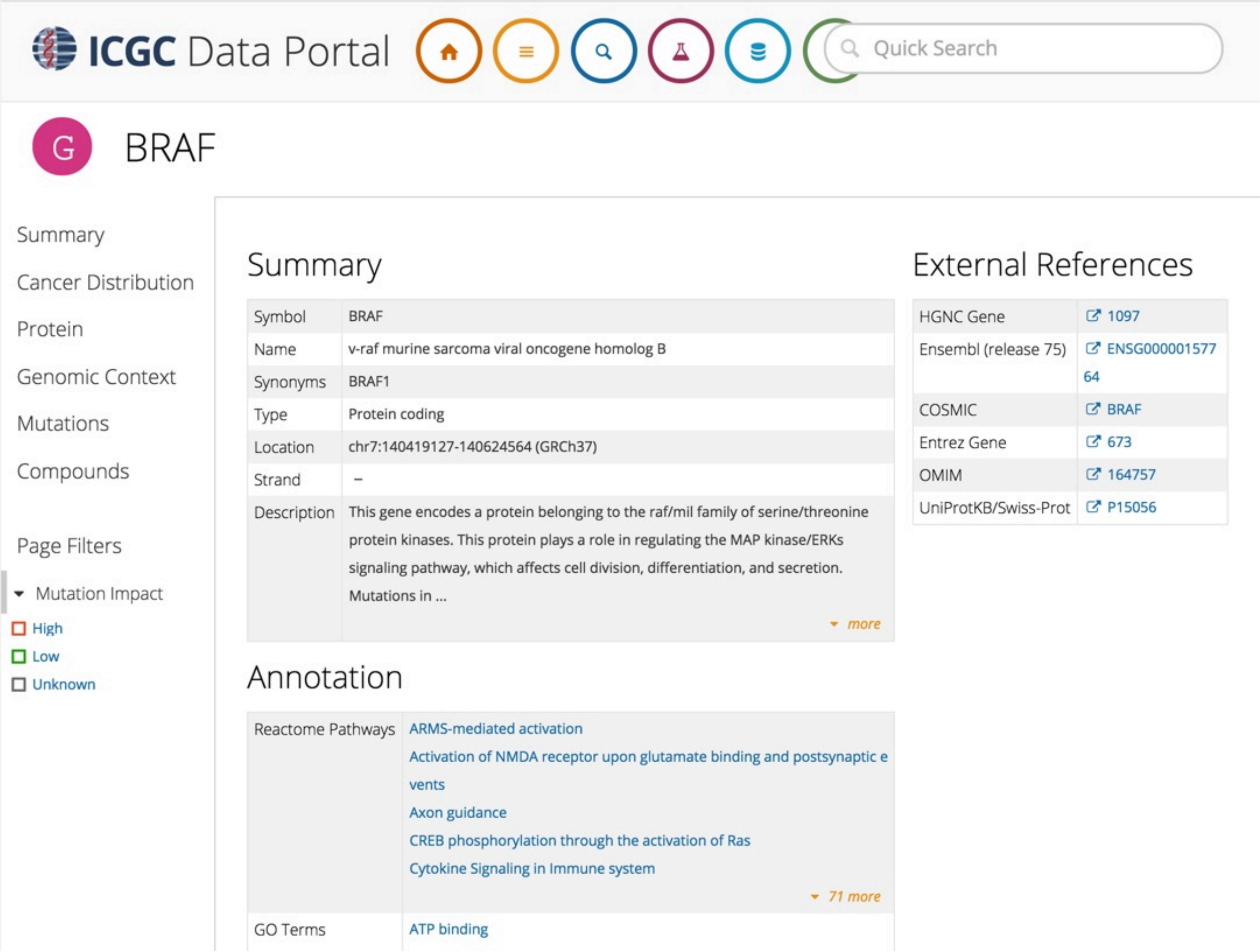
Number of Somatic Mutations in Donor's Exomes Across Cancer Projects

?

Design inspired from Nature: doi:10.1038/nature12213

Translational Breast Cancer Research, 2016

ICGC Data Portal: BRAF



BRAF

Summary

Symbol	BRAF
Name	v-raf murine sarcoma viral oncogene homolog B
Synonyms	BRAF1
Type	Protein coding
Location	chr7:140419127-140624564 (GRCh37)
Strand	–
Description	This gene encodes a protein belonging to the raf/mil family of serine/threonine protein kinases. This protein plays a role in regulating the MAP kinase/ERKs signaling pathway, which affects cell division, differentiation, and secretion. Mutations in ... more

External References

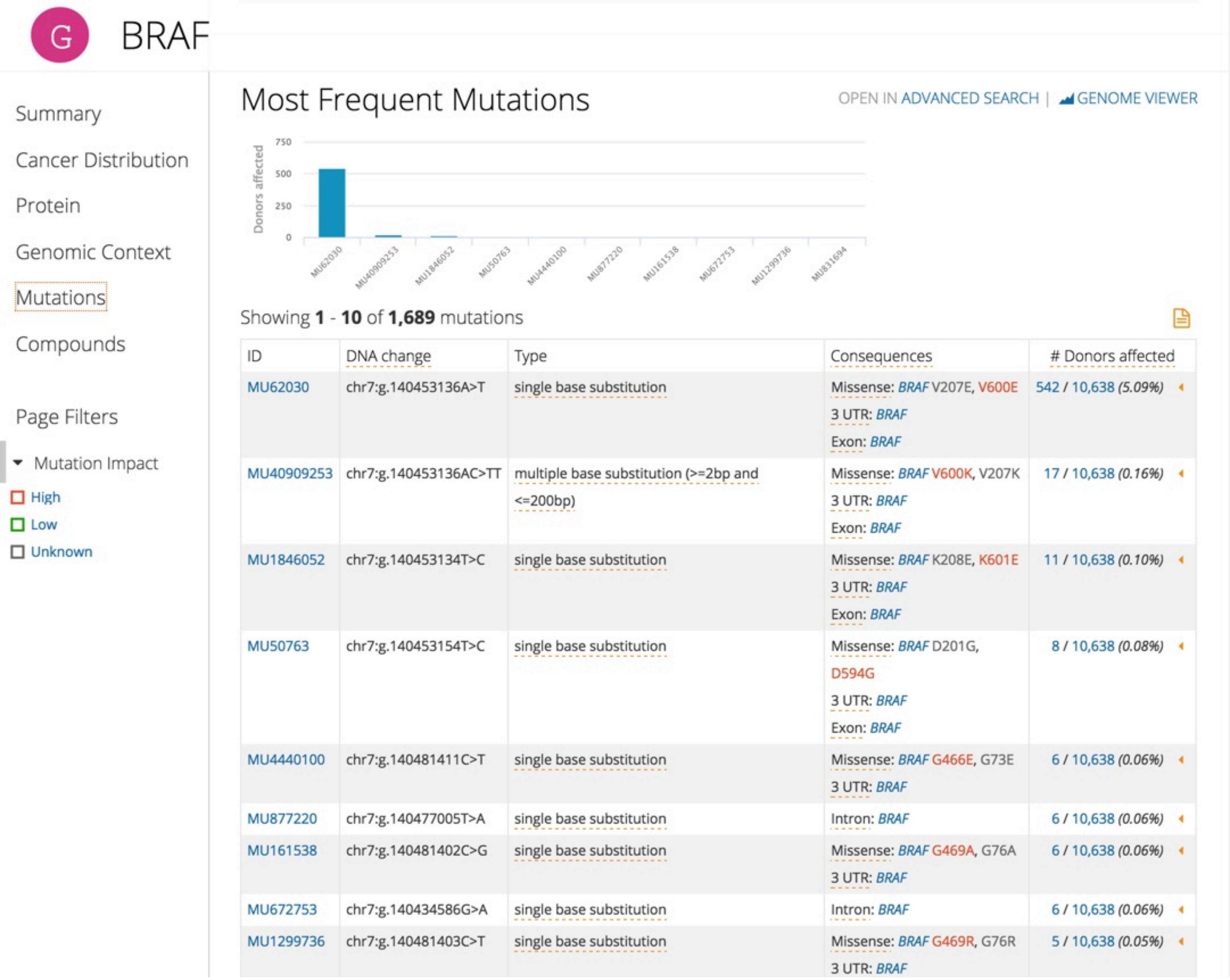
HGNC Gene	1097
Ensembl (release 75)	ENSG00000157764
COSMIC	BRAF
Entrez Gene	673
OMIM	164757
UniProtKB/Swiss-Prot	P15056

Annotation

Reactome Pathways	ARMS-mediated activation Activation of NMDA receptor upon glutamate binding and postsynaptic events Axon guidance CREB phosphorylation through the activation of Ras Cytokine Signaling in Immune system 71 more
GO Terms	ATP binding

Translational Breast Cancer Research, 2016

ICGC Data Portal: BRAF mutations



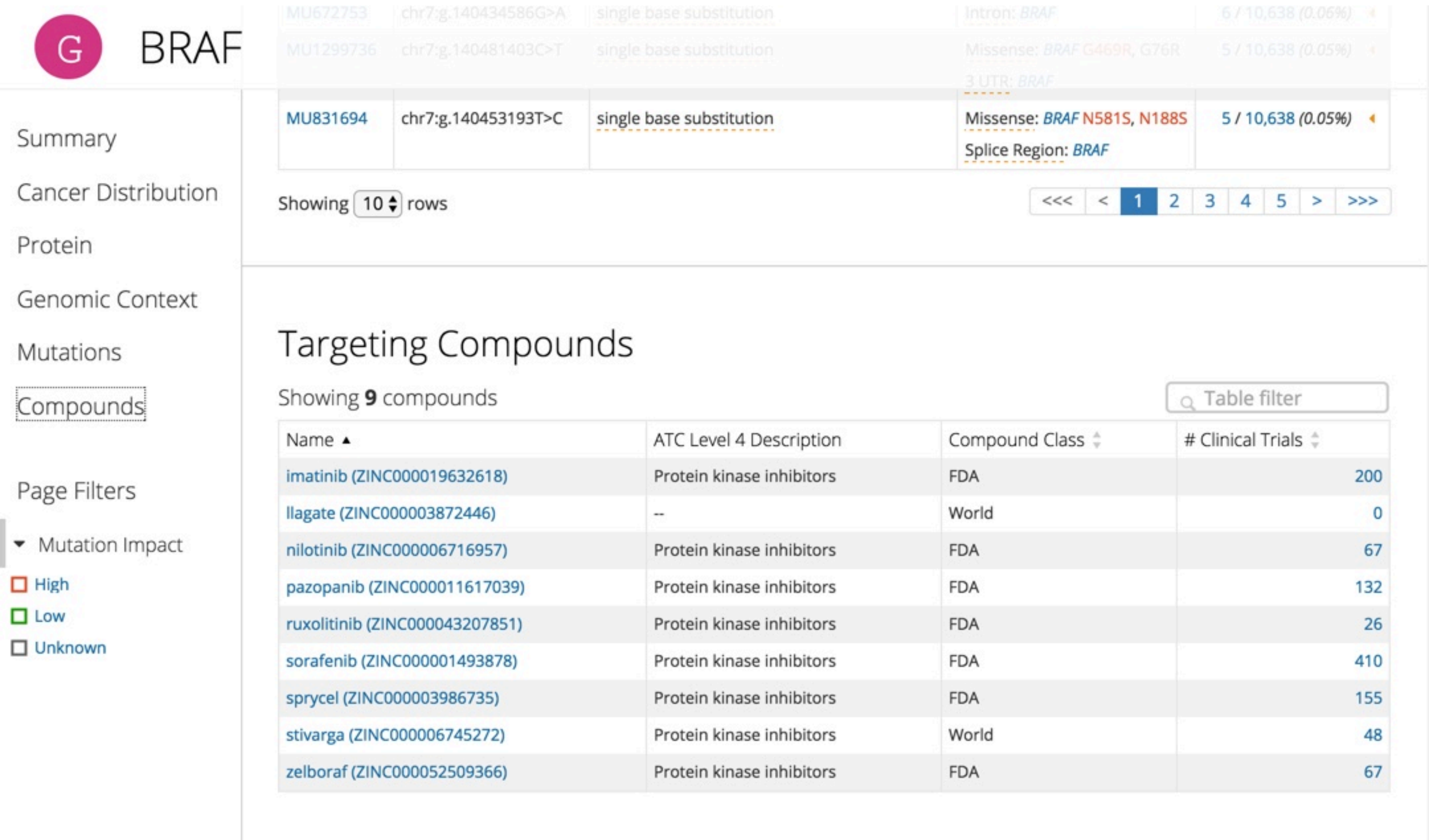
Most Frequent Mutations

Showing **1** - **10** of **1,689** mutations

ID	DNA change	Type	Consequences	# Donors affected
MU62030	chr7:g.140453136A>T	single base substitution	Missense: *BRAF* V207E, V600E 3 UTR: *BRAF* Exon: *BRAF*	542 / 10,638 *(5.09%)*
MU40909253	chr7:g.140453136AC>TT	multiple base substitution (>=2bp and <=200bp)	Missense: *BRAF* V600K, V207K 3 UTR: *BRAF* Exon: *BRAF*	17 / 10,638 *(0.16%)*
MU1846052	chr7:g.140453134T>C	single base substitution	Missense: *BRAF* K208E, K601E 3 UTR: *BRAF* Exon: *BRAF*	11 / 10,638 *(0.10%)*
MU50763	chr7:g.140453154T>C	single base substitution	Missense: *BRAF* D201G, D594G 3 UTR: *BRAF* Exon: *BRAF*	8 / 10,638 *(0.08%)*
MU4440100	chr7:g.140481411C>T	single base substitution	Missense: *BRAF* G466E, G73E 3 UTR: *BRAF*	6 / 10,638 *(0.06%)*
MU877220	chr7:g.140477005T>A	single base substitution	Intron: *BRAF*	6 / 10,638 *(0.06%)*
MU161538	chr7:g.140481402C>G	single base substitution	Missense: *BRAF* G469A, G76A 3 UTR: *BRAF*	6 / 10,638 *(0.06%)*
MU672753	chr7:g.140434586G>A	single base substitution	Intron: *BRAF*	6 / 10,638 *(0.06%)*
MU1299736	chr7:g.140481403C>T	single base substitution	Missense: *BRAF* G469R, G76R 3 UTR: *BRAF*	5 / 10,638 *(0.05%)*

Translational Breast Cancer Research, 2016

ICGC Data Portal: BRAF targeting compounds



MU672753	chr7:g.140434586G>A	single base substitution	Intron: BRAF	6 / 10,638 (0.06%)
MU1299736	chr7:g.140481403C>T	single base substitution	Missense: BRAF G469R, G76R 3 UTR: BRAF	5 / 10,638 (0.05%)
MU831694	chr7:g.140453193T>C	single base substitution	Missense: BRAF N581S, N188S Splice Region: BRAF	5 / 10,638 (0.05%)

Showing 10 rows

<<< < 1 2 3 4 5 > >>>

Targeting Compounds

Showing **9** compounds

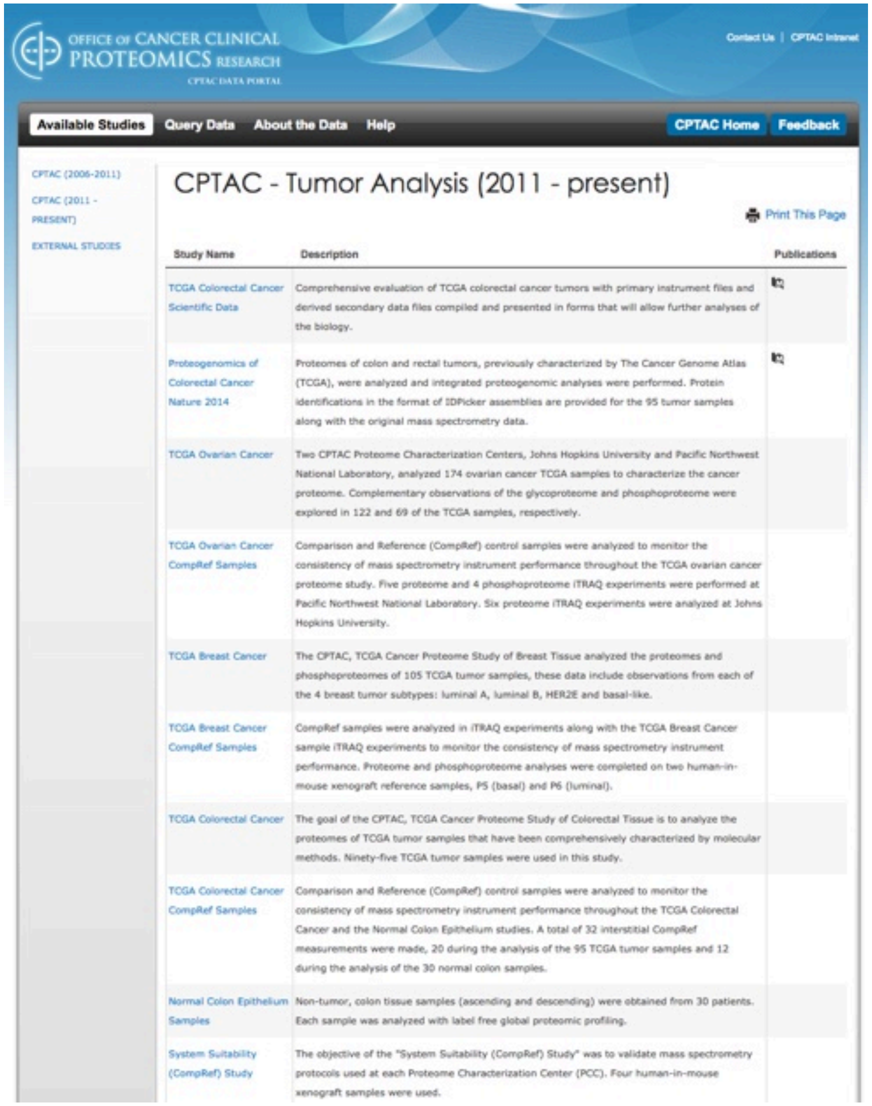
Table filter

Name ▲	ATC Level 4 Description	Compound Class	# Clinical Trials
imatinib (ZINC000019632618)	Protein kinase inhibitors	FDA	200
llagate (ZINC000003872446)	--	World	0
nilotinib (ZINC000006716957)	Protein kinase inhibitors	FDA	67
pazopanib (ZINC000011617039)	Protein kinase inhibitors	FDA	132
ruxolitinib (ZINC000043207851)	Protein kinase inhibitors	FDA	26
sorafenib (ZINC000001493878)	Protein kinase inhibitors	FDA	410
sprycel (ZINC000003986735)	Protein kinase inhibitors	FDA	155
stivarga (ZINC000006745272)	Protein kinase inhibitors	World	48
zelboraf (ZINC000052509366)	Protein kinase inhibitors	FDA	67

Translational Breast Cancer Research, 2016

Clinical Proteome Tumor Analysis Consortium (CPTAC)

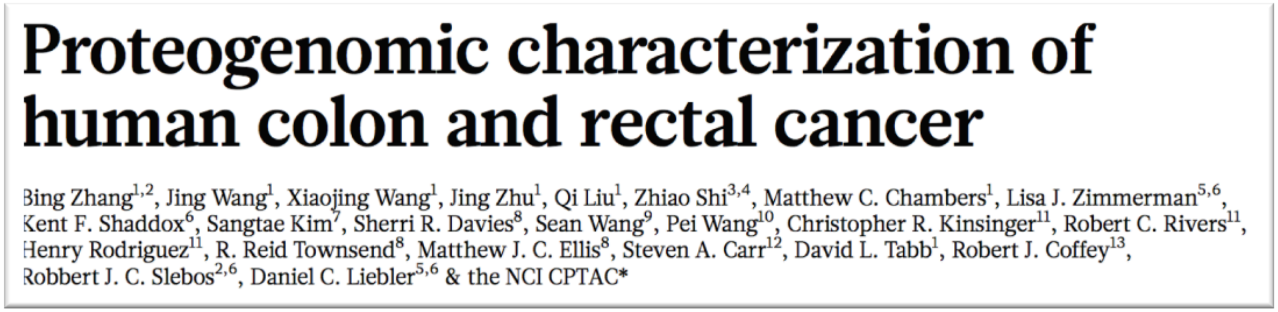
- Goals
  - Global proteomic characterization of TCGA tumors
  - Proteogenomic data integration
- Five centers established in 2011
  - Broad Institute
  - John Hopkins University
  - Pacific Northwest National Laboratory
  - Washington University
  - Vanderbilt University
- Tumor samples
  - Breast (Broad and Wash U)
  - Colon and Rectal (Vanderbilt)
  - Ovarian (JHU and PNNL)
- CPTAC data portal
  - https://cptac-data-portal.georgetown.edu



CPTAC - Tumor Analysis (2011 - present)

Study Name	Description	Publications
TCGA Colorectal Cancer Scientific Data	Comprehensive evaluation of TCGA colorectal cancer tumors with primary instrument files and derived secondary data files compiled and presented in forms that will allow further analyses of the biology.	🔍
Proteogenomics of Colorectal Cancer Nature 2014	Proteomes of colon and rectal tumors, previously characterized by The Cancer Genome Atlas (TCGA), were analyzed and integrated proteogenomic analyses were performed. Protein identifications in the format of IDPicker assemblies are provided for the 95 tumor samples along with the original mass spectrometry data.	🔍
TCGA Ovarian Cancer	Two CPTAC Proteome Characterization Centers, Johns Hopkins University and Pacific Northwest National Laboratory, analyzed 174 ovarian cancer TCGA samples to characterize the cancer proteome. Complementary observations of the glycoproteome and phosphoproteome were explored in 122 and 69 of the TCGA samples, respectively.	
TCGA Ovarian Cancer CompRef Samples	Comparison and Reference (CompRef) control samples were analyzed to monitor the consistency of mass spectrometry instrument performance throughout the TCGA ovarian cancer proteome study. Five proteome and 4 phosphoproteome iTRAQ experiments were performed at Pacific Northwest National Laboratory. Six proteome iTRAQ experiments were analyzed at Johns Hopkins University.	
TCGA Breast Cancer	The CPTAC, TCGA Cancer Proteome Study of Breast Tissue analyzed the proteomes and phosphoproteomes of 105 TCGA tumor samples, these data include observations from each of the 4 breast tumor subtypes: luminal A, luminal B, HER2E and basal-like.	
TCGA Breast Cancer CompRef Samples	CompRef samples were analyzed in iTRAQ experiments along with the TCGA Breast Cancer sample iTRAQ experiments to monitor the consistency of mass spectrometry instrument performance. Proteome and phosphoproteome analyses were completed on two human-in-mouse xenograft reference samples, P5 (basal) and P6 (luminal).	
TCGA Colorectal Cancer	The goal of the CPTAC, TCGA Cancer Proteome Study of Colorectal Tissue is to analyze the proteomes of TCGA tumor samples that have been comprehensively characterized by molecular methods. Ninety-five TCGA tumor samples were used in this study.	
TCGA Colorectal Cancer CompRef Samples	Comparison and Reference (CompRef) control samples were analyzed to monitor the consistency of mass spectrometry instrument performance throughout the TCGA Colorectal Cancer and the Normal Colon Epithelium studies. A total of 32 interstitial CompRef measurements were made, 20 during the analysis of the 95 TCGA tumor samples and 12 during the analysis of the 30 normal colon samples.	
Normal Colon Epithelium Samples	Non-tumor, colon tissue samples (ascending and descending) were obtained from 30 patients. Each sample was analyzed with label free global proteomic profiling.	
System Suitability (CompRef) Study	The objective of the "System Suitability (CompRef) Study" was to validate mass spectrometry protocols used at each Proteome Characterization Center (PCC). Four human-in-mouse xenograft samples were used.	

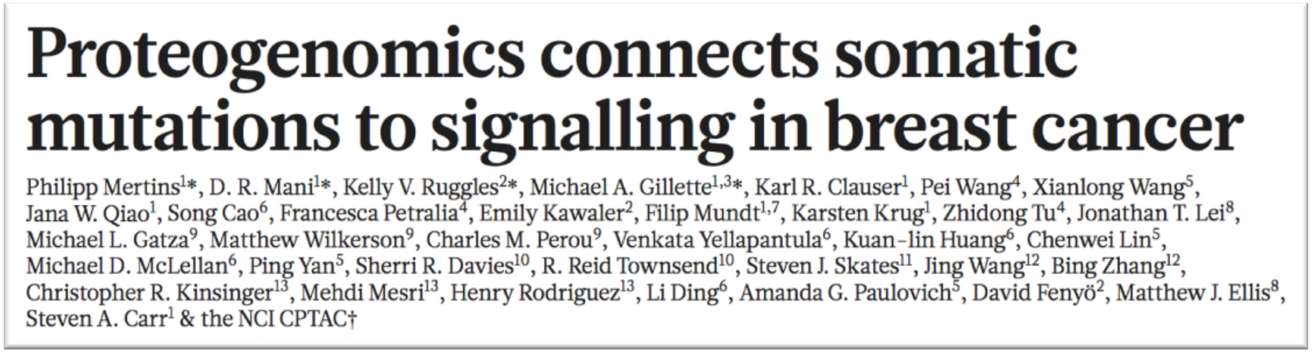
Clinical Proteome Tumor Analysis Consortium (CPTAC)



Proteogenomic characterization of human colon and rectal cancer

Bing Zhang[1,2], Jing Wang[1], Xiaojing Wang[1], Jing Zhu[1], Qi Liu[1], Zhiao Shi[3,4], Matthew C. Chambers[1], Lisa J. Zimmerman[5,6], Kent F. Shaddox[6], Sangtae Kim[7], Sherri R. Davies[8], Sean Wang[9], Pei Wang[10], Christopher R. Kinsinger[11], Robert C. Rivers[11], Henry Rodriguez[11], R. Reid Townsend[8], Matthew J. C. Ellis[8], Steven A. Carr[12], David L. Tabb[1], Robert J. Coffey[13], Robbert J. C. Slebos[2,6], Daniel C. Liebler[5,6] & the NCI CPTAC*

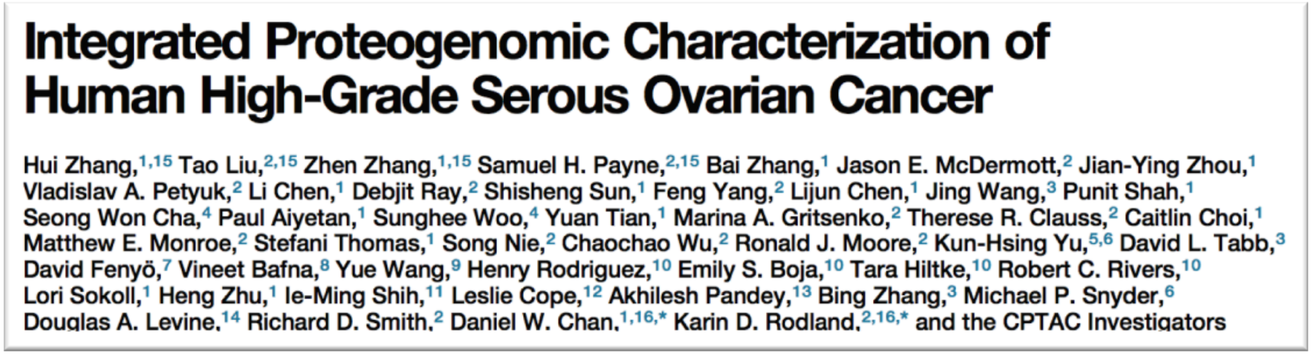
Nature, 2014



Proteogenomics connects somatic mutations to signalling in breast cancer

Philipp Mertins[1]*, D. R. Mani[1]*, Kelly V. Ruggles[2]*, Michael A. Gillette[1,3]*, Karl R. Clauser[1], Pei Wang[4], Xianlong Wang[5], Jana W. Qiao[1], Song Cao[6], Francesca Petralia[4], Emily Kawaler[2], Filip Mundt[1,7], Karsten Krug[1], Zhidong Tu[4], Jonathan T. Lei[8], Michael L. Gatza[9], Matthew Wilkerson[9], Charles M. Perou[9], Venkata Yellapantula[6], Kuan-lin Huang[6], Chenwei Lin[5], Michael D. McLellan[6], Ping Yan[5], Sherri R. Davies[10], R. Reid Townsend[10], Steven J. Skates[11], Jing Wang[12], Bing Zhang[12], Christopher R. Kinsinger[13], Mehdi Mesri[13], Henry Rodriguez[13], Li Ding[6], Amanda G. Paulovich[5], David Fenyö[2], Matthew J. Ellis[8], Steven A. Carr[1] & the NCI CPTAC†

Nature, 2016

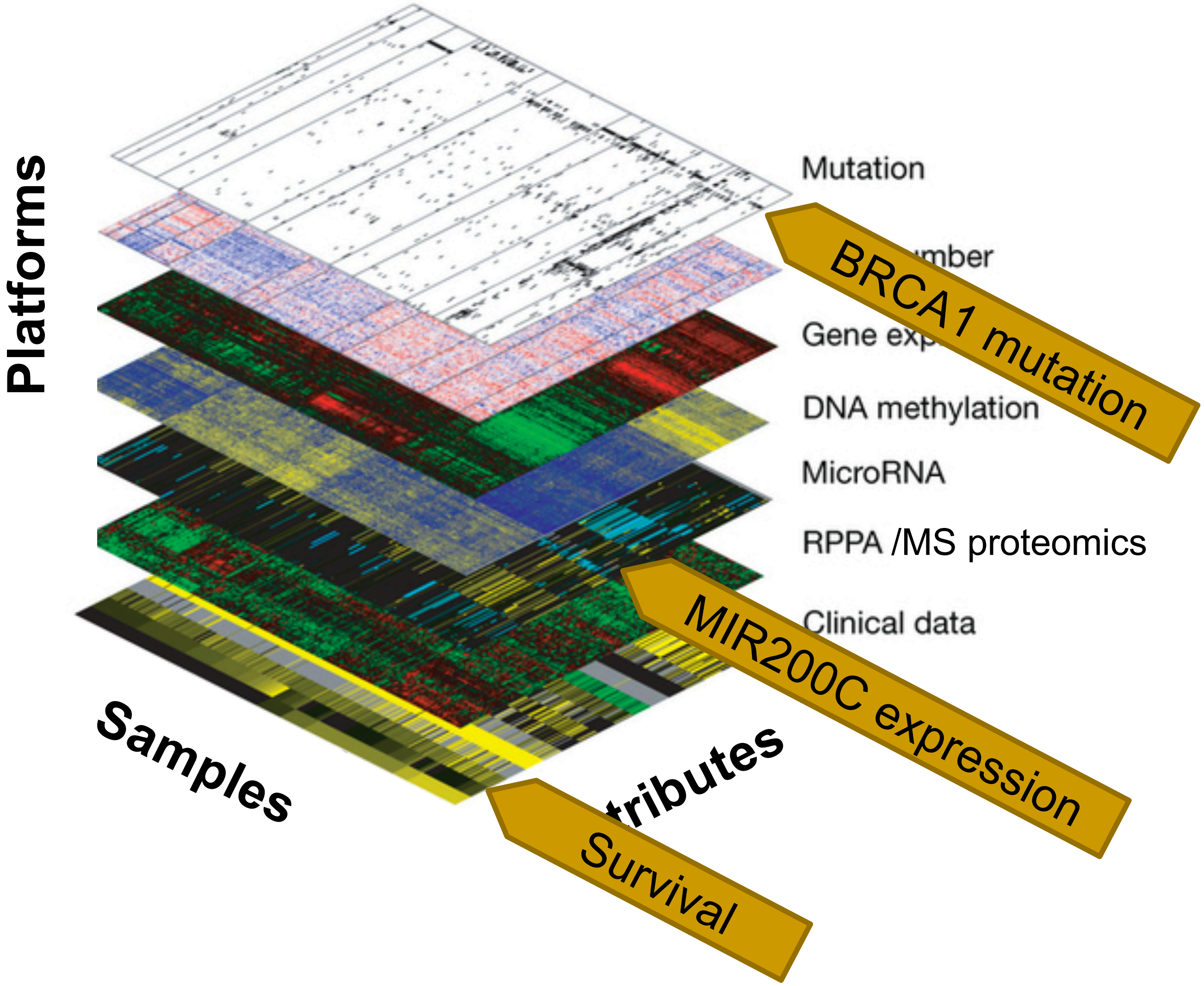


Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer

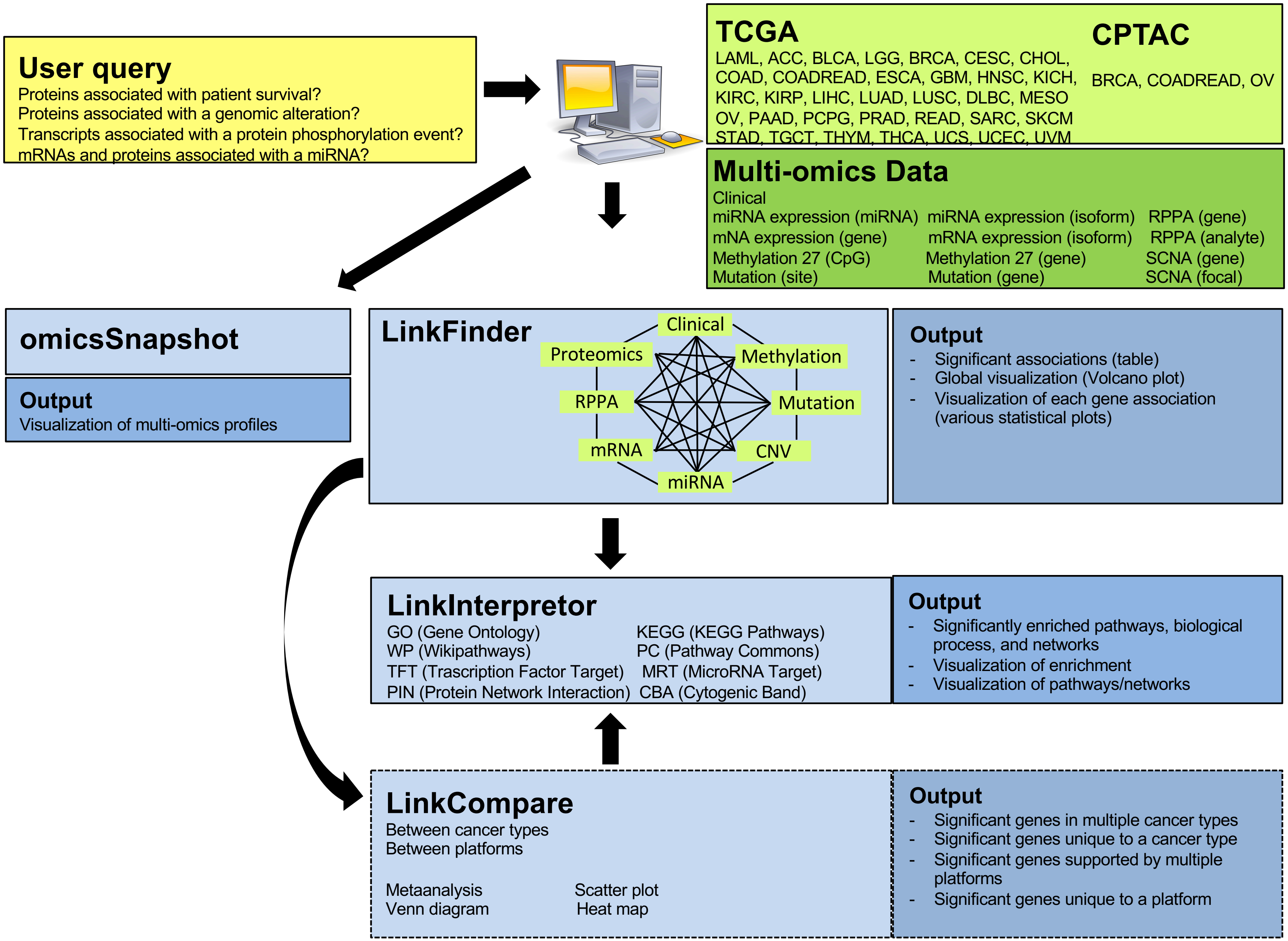
Hui Zhang,[1,15] Tao Liu,[2,15] Zhen Zhang,[1,15] Samuel H. Payne,[2,15] Bai Zhang,[1] Jason E. McDermott,[2] Jian-Ying Zhou,[1] Vladislav A. Petyuk,[2] Li Chen,[1] Debjit Ray,[2] Shisheng Sun,[1] Feng Yang,[2] Lijun Chen,[1] Jing Wang,[3] Punit Shah,[1] Seong Won Cha,[4] Paul Aiyetan,[1] Sunghee Woo,[4] Yuan Tian,[1] Marina A. Gritsenko,[2] Therese R. Clauss,[2] Caitlin Choi,[1] Matthew E. Monroe,[2] Stefani Thomas,[1] Song Nie,[2] Chaochao Wu,[2] Ronald J. Moore,[2] Kun-Hsing Yu,[5,6] David L. Tabb,[3] David Fenyö,[7] Vineet Bafna,[8] Yue Wang,[9] Henry Rodriguez,[10] Emily S. Boja,[10] Tara Hiltke,[10] Robert C. Rivers,[10] Lori Sokoll,[1] Heng Zhu,[1] Ie-Ming Shih,[11] Leslie Cope,[12] Akhilesh Pandey,[13] Bing Zhang,[3] Michael P. Snyder,[6] Douglas A. Levine,[14] Richard D. Smith,[2] Daniel W. Chan,[1,16,*] Karin D. Rodland,[2,16,*] and the CPTAC Investigators

Cell, 2016

LinkedOmics: cross-omics association analysis



LinkedOmics: cross-omics association analysis



User query

Proteins associated with patient survival?
Proteins associated with a genomic alteration?
Transcripts associated with a protein phosphorylation event?
mRNAs and proteins associated with a miRNA?

TCGA

LAML, ACC, BLCA, LGG, BRCA, CESC, CHOL,
COAD, COADREAD, ESCA, GBM, HNSC, KICH,
KIRC, KIRP, LIHC, LUAD, LUSC, DLBC, MESO
OV, PAAD, PCPG, PRAD, READ, SARC, SKCM
STAD, TGCT, THYM, THCA, UCS, UCEC, UVM

CPTAC

BRCA, COADREAD, OV

Multi-omics Data

Clinical
miRNA expression (miRNA)
mNA expression (gene)
Methylation 27 (CpG)
Mutation (site)

miRNA expression (isoform)
mRNA expression (isoform)
Methylation 27 (gene)
Mutation (gene)

RPPA (gene)
RPPA (analyte)
SCNA (gene)
SCNA (focal)

omicsSnapshot

Output

Visualization of multi-omics profiles

LinkFinder

Output

- Significant associations (table)
- Global visualization (Volcano plot)
- Visualization of each gene association (various statistical plots)

LinkInterpretor

GO (Gene Ontology)
WP (Wikipathways)
TFT (Trascription Factor Target)
PIN (Protein Network Interaction)

KEGG (KEGG Pathways)
PC (Pathway Commons)
MRT (MicroRNA Target)
CBA (Cytogenic Band)

Output

- Significantly enriched pathways, biological process, and networks
- Visualization of enrichment
- Visualization of pathways/networks

LinkCompare

Between cancer types
Between platforms

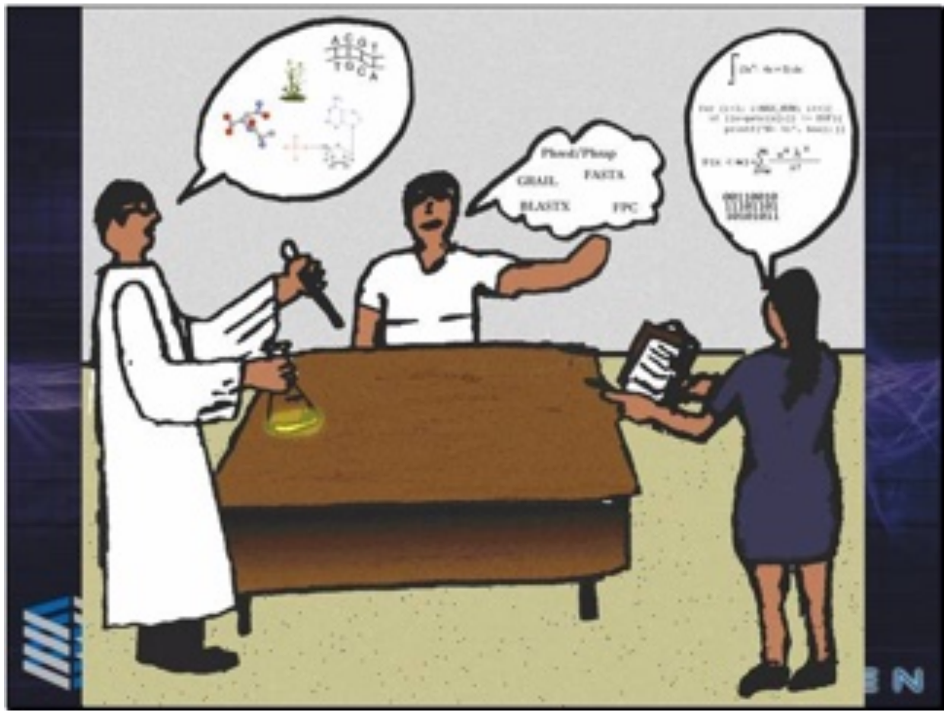
Metaanalysis
Venn diagram

Scatter plot
Heat map

Output

- Significant genes in multiple cancer types
- Significant genes unique to a cancer type
- Significant genes supported by multiple platforms
- Significant genes unique to a platform

Become a superuser



Graph courtesy of http://www.incogen.com/

- Algorithm developer
  - Statisticians
  - Mathematicians
  - Computer scientists
- Tool developer
  - Bioinformaticians
- Data provider/consumer
  - Biologists

+

R
Linux
Python