IMIM Parc de Salut MAR Barcelona

Institut Hospital del Mar d'Investigacions Mèdiques

Functional analysis of omics data

february 2020

MARGenomics

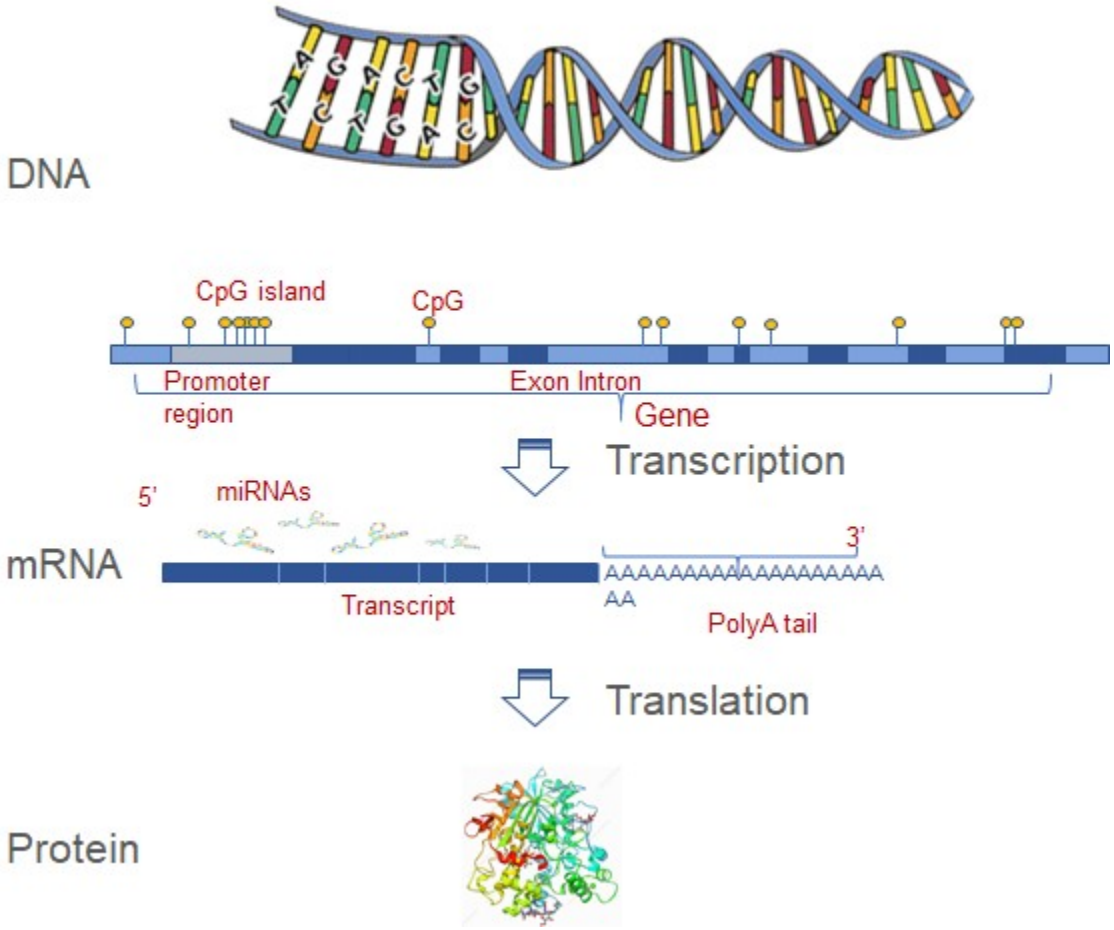
Summary

1. Omics
2. Public sources
3. Functional analysis

Summary

1. Omics
2. Public sources
3. Functional analysis

1. Omics



Genome

- SNPs and other variants
- Copy number alterations
- Rearrangements

Transcriptome

- Gene expression
- Alternative splicing

Epigenome

- CpGs methylation
- Histone modification
- miRNAs

Summary

1. Omics
2. Public sources
3. Functional analysis

2. Public sources (open data)

Projects

ENCODE
TCGA
GTEx
Cancer Cell Line Encyclopedia
The Human Protein Atlas
FANTOM5
Researchers
Publications
...

Web sources

GDAC iCGC
cBioPortal
GEPIA XENA
GEO SRA
dbGaP
...

Databases

GO
REACTOME
KEGG
BioCarta
mSigDB
...

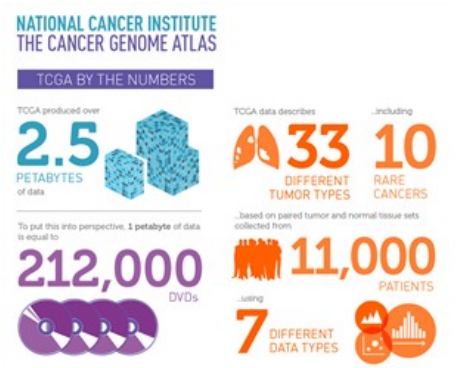
2. Public sources

- Advantages
  - Thousands of sample data available for several pathologies, projects, omic and technologies
  - Analyses available in many of them
  - Usable in your publications
- Limitations
  - Discordances
  - Constantly updating, reproducibility problems
  - Cancer biased

2. TCGA

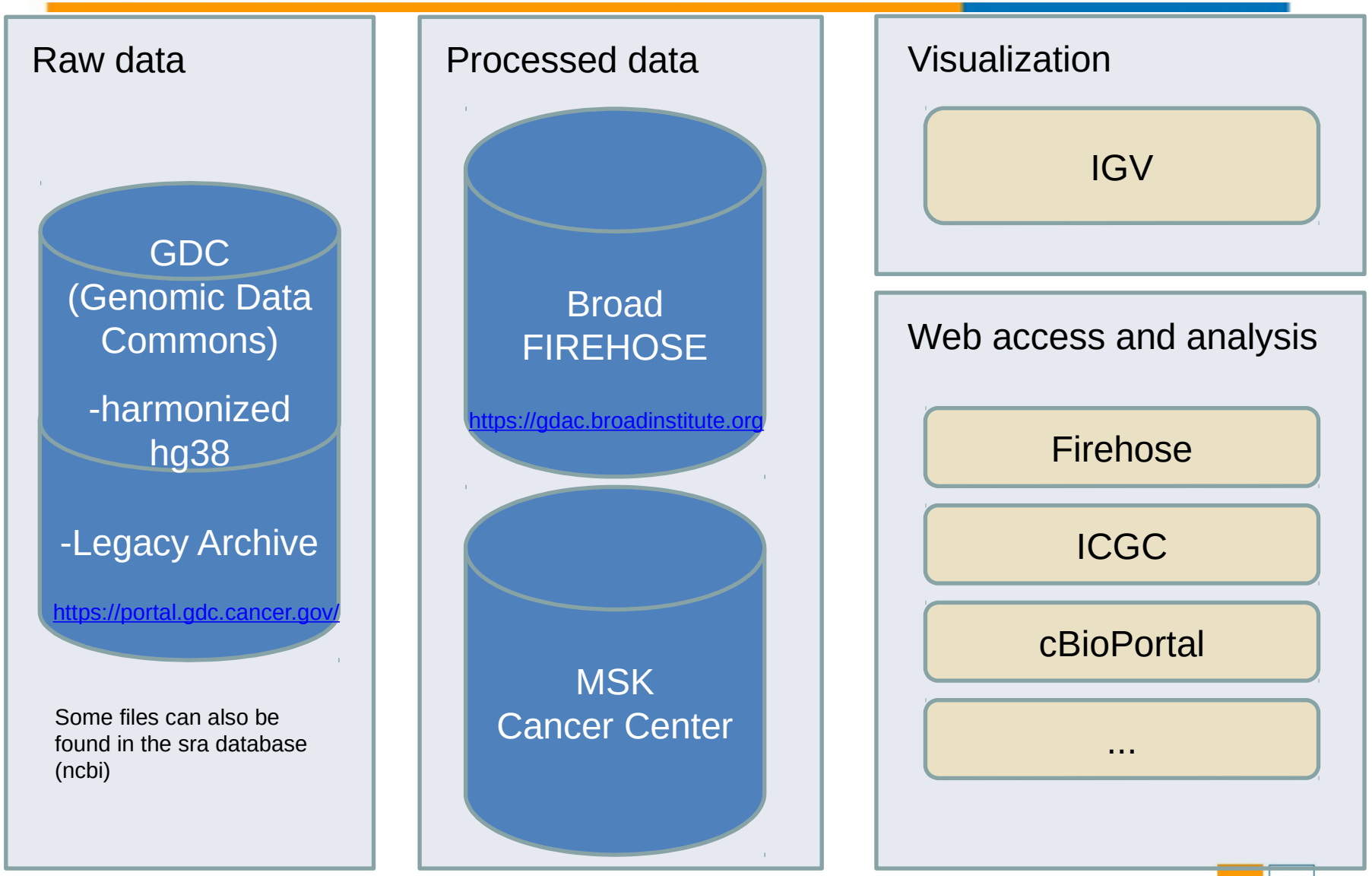
GOAL: To improve the ability to diagnose, treat and prevent cancer

- Search for genetic changes
- 33 different cancer types
- Started in 2006 from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI).
- Network of research
- Data publicly available
- Cancer samples and some normal ones
- The Pan-Cancer Atlas compares TCGA tumor types
- Many publications produced



http://cancergenome.nih.gov/

2. TCGA


Raw data
GDC
(Genomic Data Commons)
-harmonized hg38
-Legacy Archive
https://portal.gdc.cancer.gov/
Some files can also be found in the sra database (ncbi)
Processed data
Broad FIREHOSE
https://gdac.broadinstitute.org
MSK Cancer Center
Visualization
IGV
Web access and analysis
Firehose
ICGC
cBioPortal
...

2. TCGA Firehose

Dashboards Data Analyses Software Documentation FAQ Download Contact Us

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Browse
Bladder urothelial carcinoma	BLCA	412	Browse	Browse
Breast invasive carcinoma	BRCA	1098	Browse	Browse
Cervical and endocervical cancers	CESC	307	Browse	Browse
Cholangiocarcinoma	CHOL	51	Browse	Browse
Colon adenocarcinoma	COAD	460	Browse	Browse
Colorectal adenocarcinoma	COADREAD	631	Browse	Browse
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browse	Browse
Esophageal carcinoma	ESCA	185	Browse	Browse
FFPE Pilot Phase II	FPPP	38	None	Browse
Glioblastoma multiforme	GBM	613	Browse	Browse
Glioma	GBMLGG	1129	Browse	Browse
Head and Neck squamous cell carcinoma	HNSC	528	Browse	Browse
Kidney Chromophobe	KICH	113	Browse	Browse
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	Browse	Browse
Kidney renal clear cell carcinoma	KIRC	537	Browse	Browse
Kidney renal papillary cell carcinoma	KIRP	323	Browse	Browse
Acute Myeloid Leukemia	LAML	200	Browse	Browse
Brain Lower Grade Glioma	LGG	516	Browse	Browse
Liver hepatocellular carcinoma	LIHC	377	Browse	Browse
Lung adenocarcinoma	LUAD	585	Browse	Browse
Lung squamous cell carcinoma	LUSC	504	Browse	Browse
Mesothelioma	MESO	87	Browse	Browse
Ovarian serous cystadenocarcinoma	OV	602	Browse	Browse
Pancreatic adenocarcinoma	PAAD	185	Browse	Browse

Data Summary for Lung adenocarcinoma

Version: stddata__2016_01_28

Cohort	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF	rawMAF
LUAD	585	522	516	120	578	32	515	0	513	365	230	542

Browse Samples Browse Workflow Graph Browse Analyses

Thyroid carcinoma	THCA	503	Browse	Browse
Thymoma	THYM	124	Browse	Browse
Uterine Corpus Endometrial Carcinoma	UCEC	560	Browse	Browse
Uterine Carcinosarcoma	UCS	57	Browse	Browse
Uveal Melanoma	UVM	80	Browse	Browse

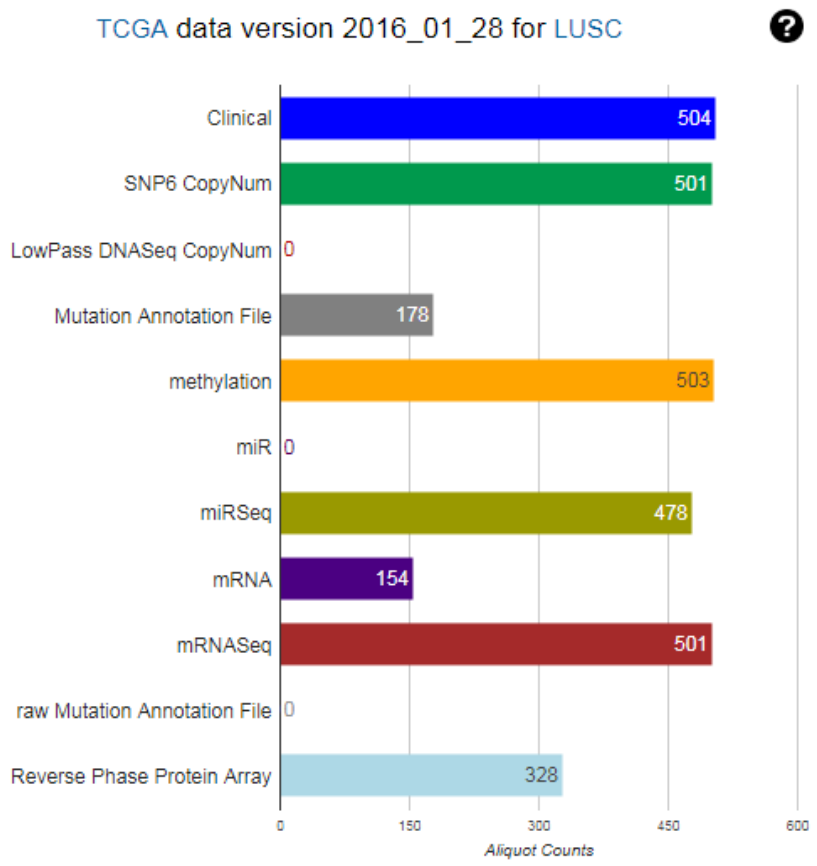
https://gdac.broadinstitute.org/

2. TCGA Firehose. LUSC

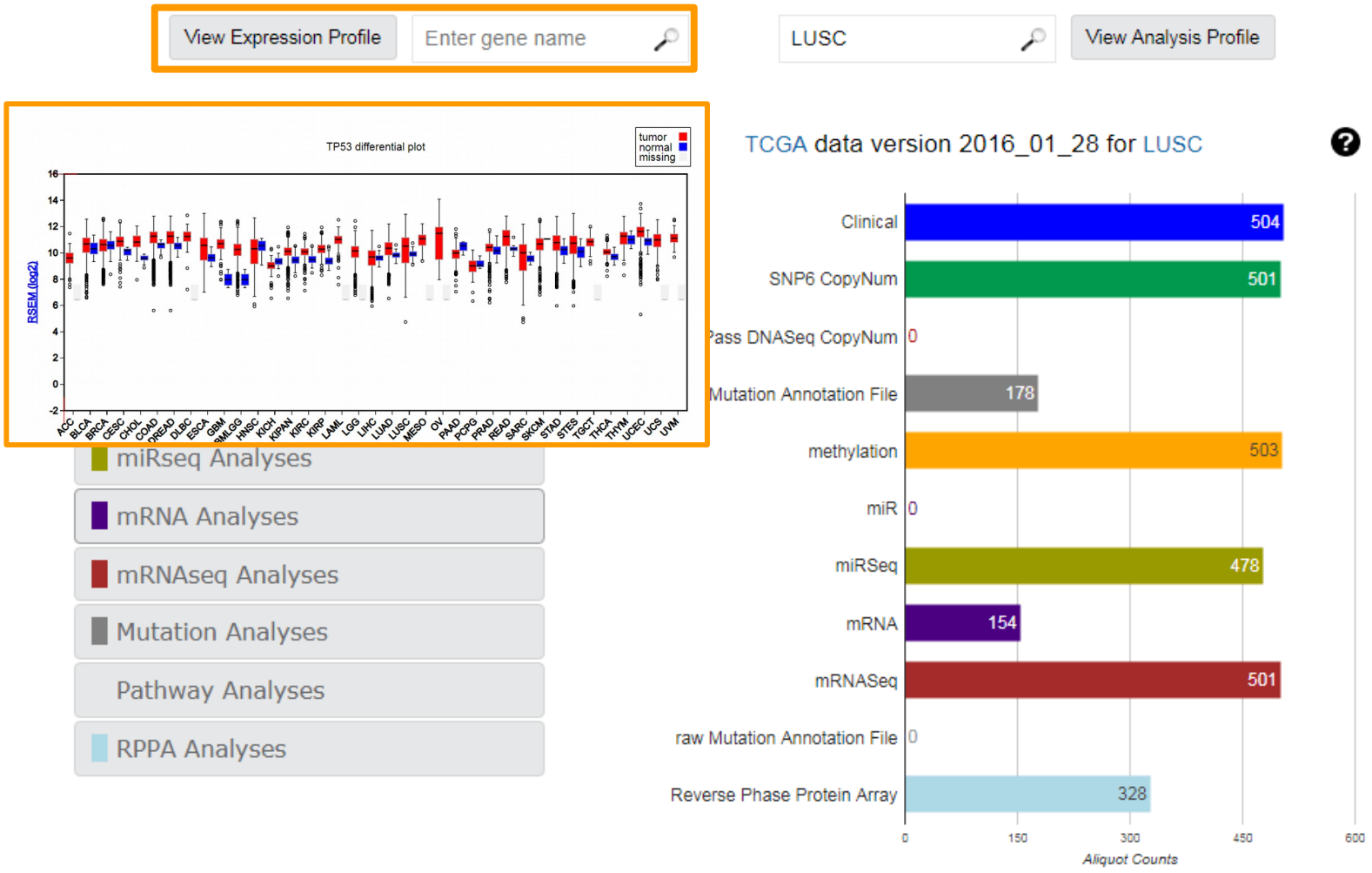
View Expression Profile | Enter gene name | LUSC | View Analysis Profile

Lung squamous cell carcinoma (LUSC)

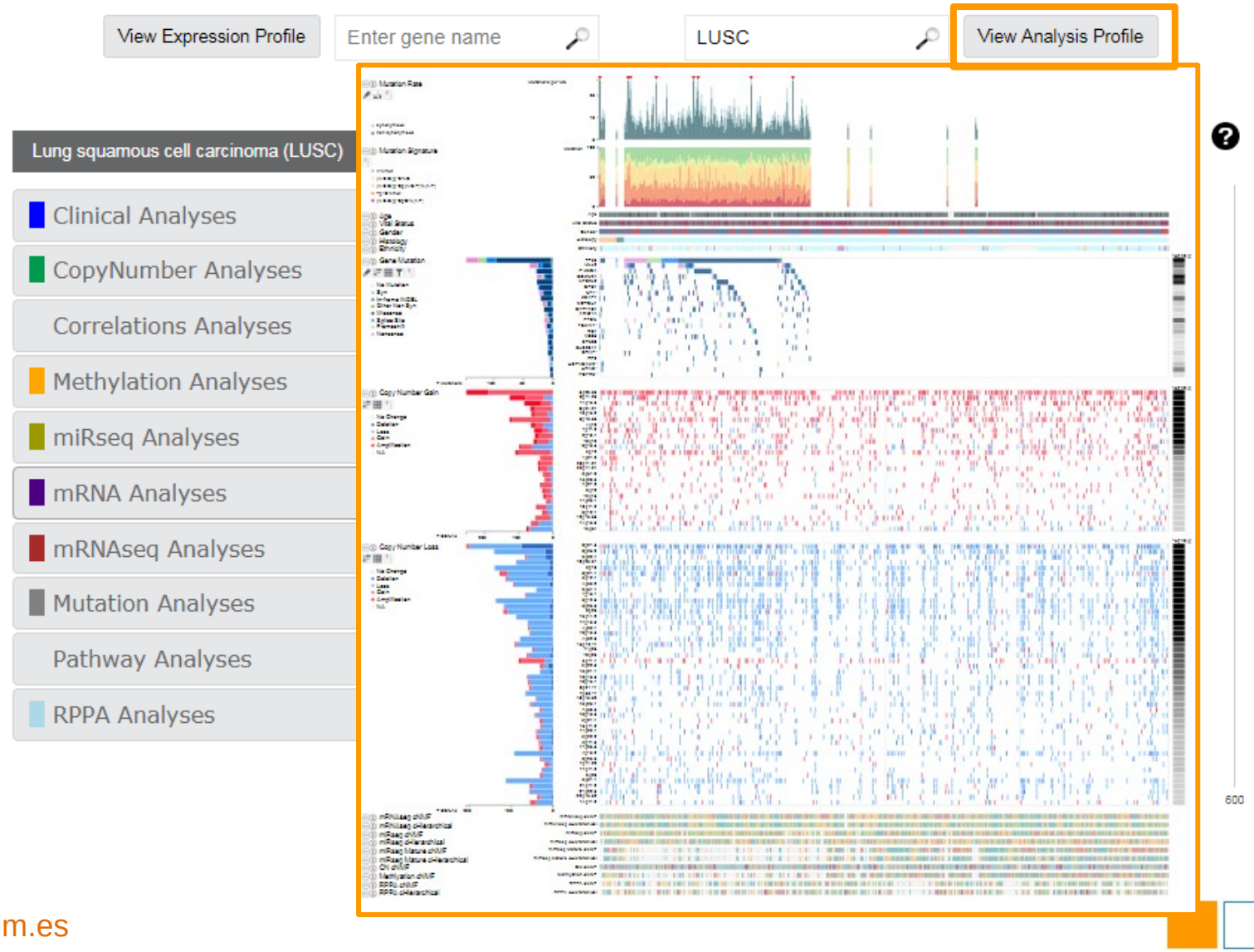
- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- Methylation Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses



2. TCGA Firehose. LUSC



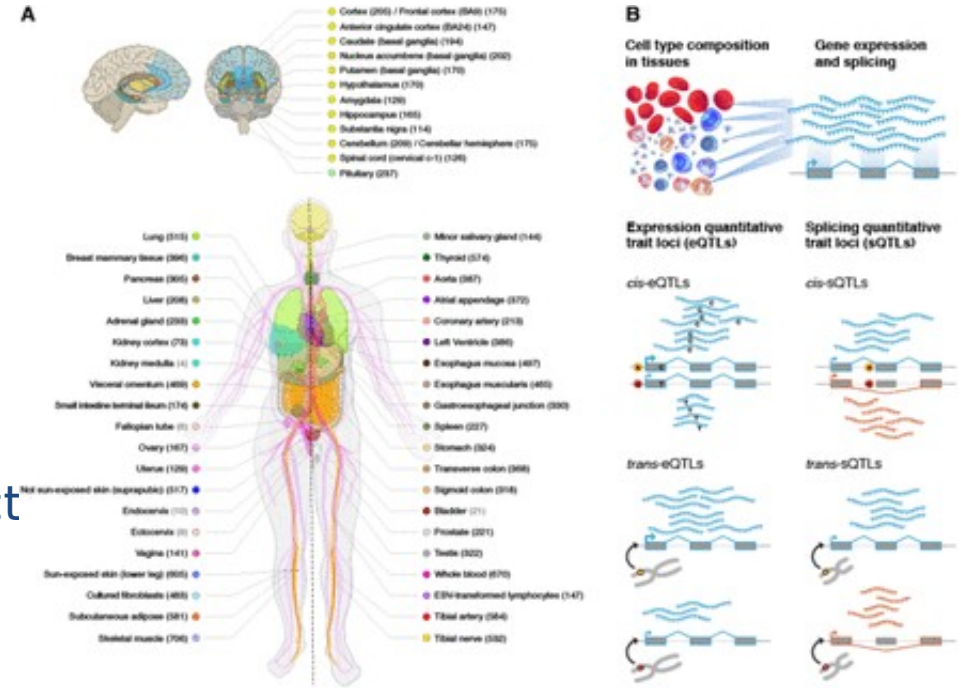
2. TCGA Firehose. LUSC



2. GTEx

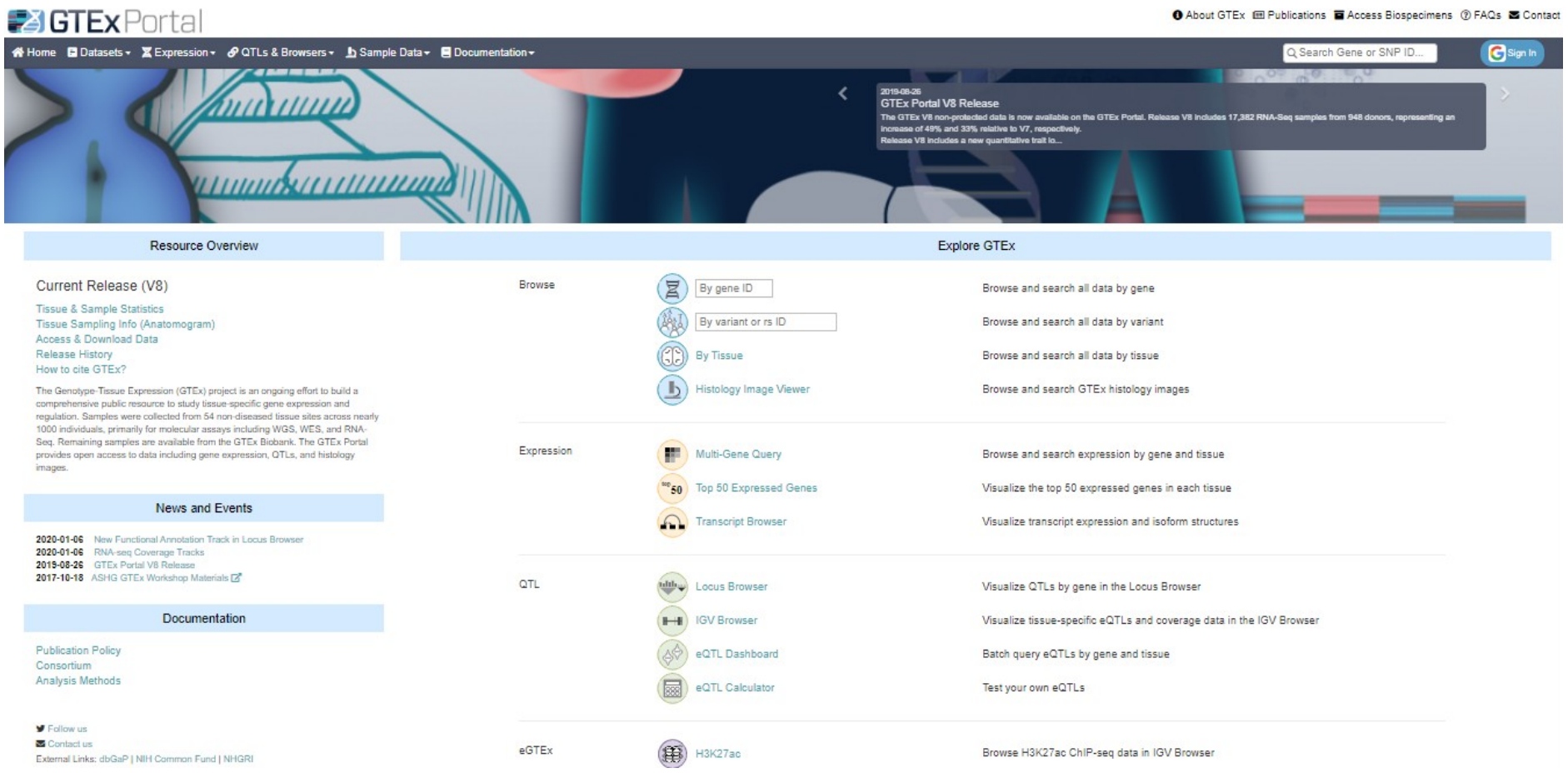
GOAL: To build a catalog of genetic effects on gene expression across a large number of human tissues

- Launched in 2010
- ~50 tissues
- ~1000 postmortem donors
- ~17000 RNA-Seq samples
- Optimize protocols for:
  - postmortem tissue collect
  - donor recruitment
  - biospecimen processing
  - data sharing
- https://www.cbioportal.org/tutorials



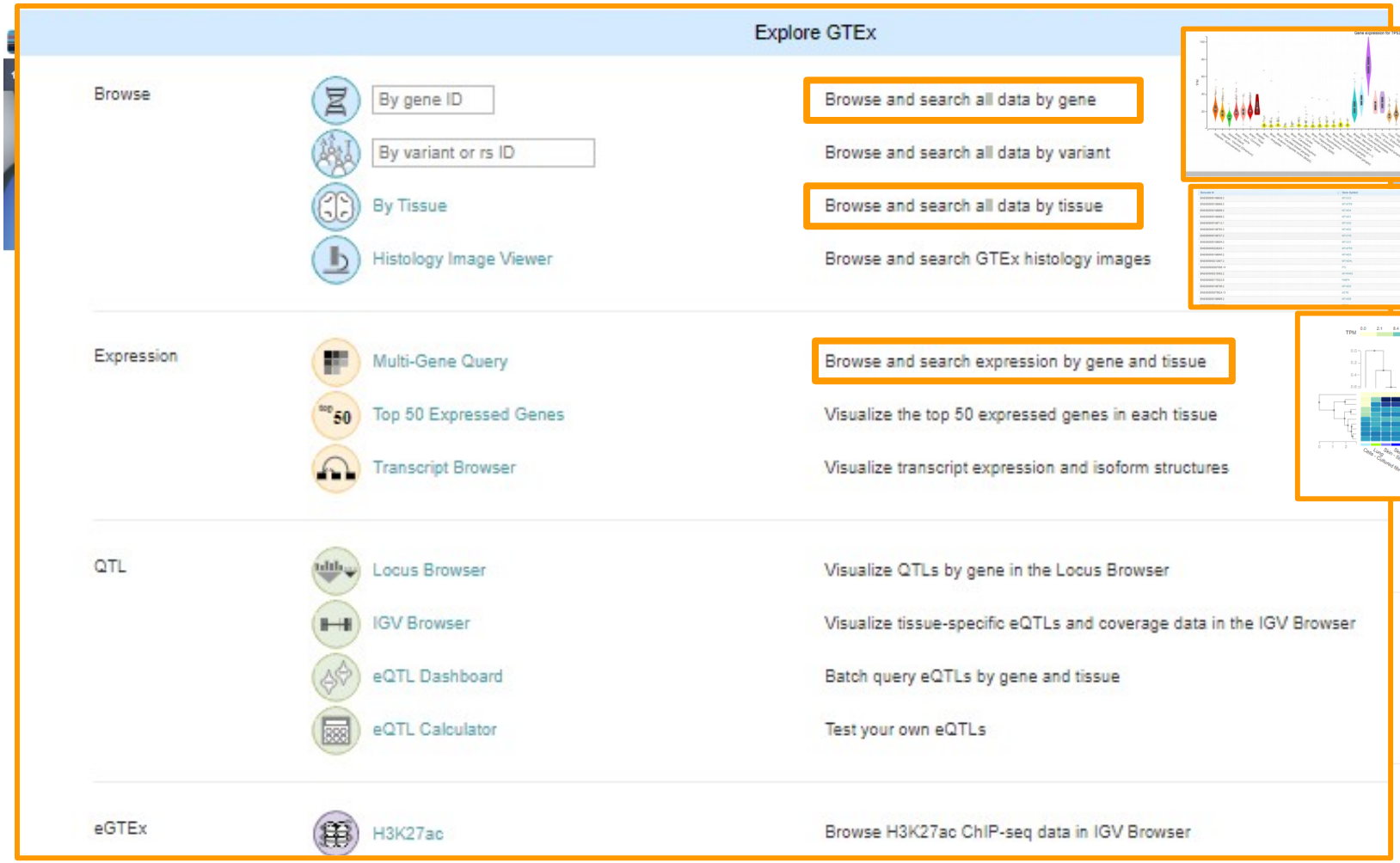
https://www.biorxiv.org/content/10.1101/787903v1.full

2. GTEx



https://gtexportal.org/

2. GTEx



https://gtexportal.org/

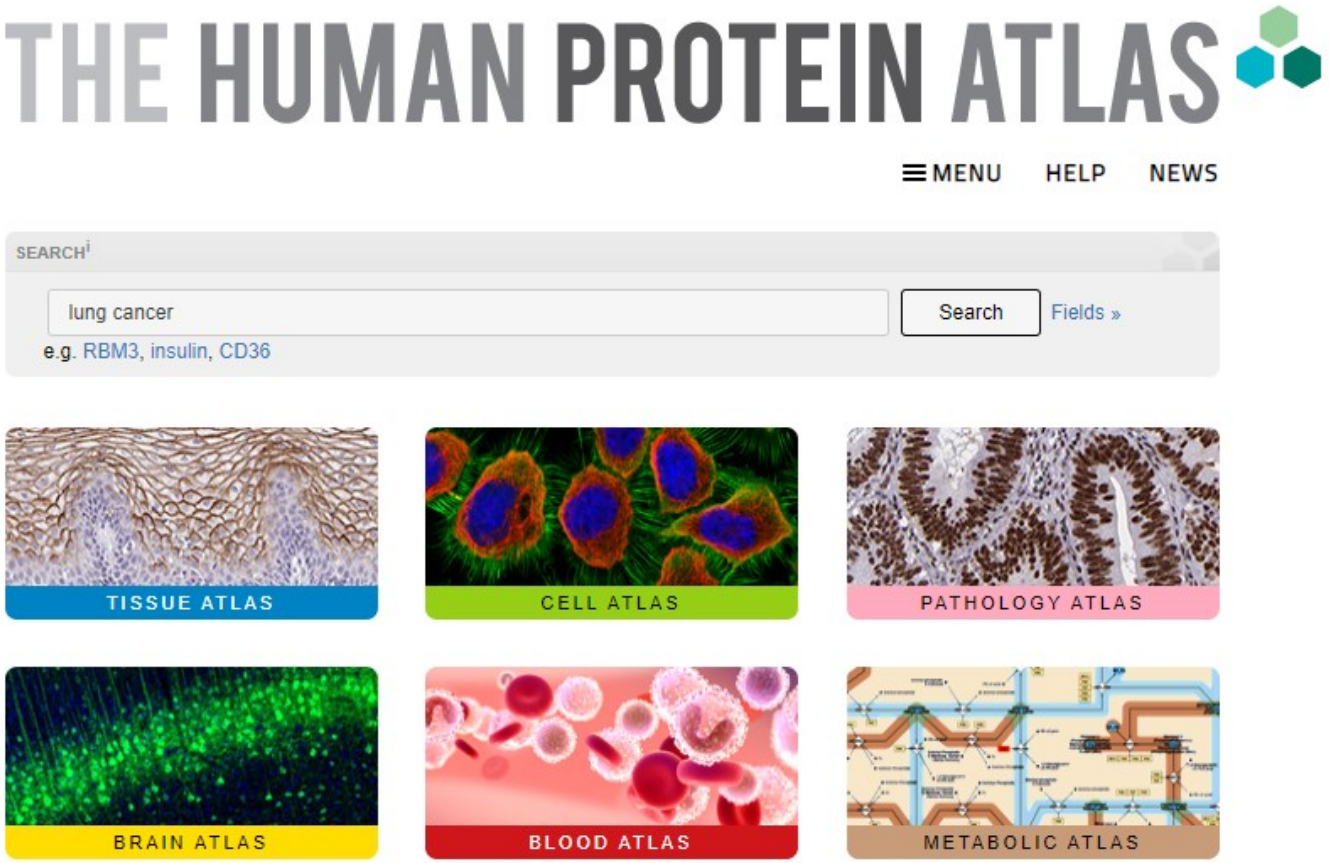
2. The Human Protein Atlas

GOAL: To map all the human proteins in cells, tissues and organs using integration of various omics technologies

- Initiated in 2003
- Swedish-based program
- Technologies including:
  - antibody-based imaging
  - mass spectrometry-based proteomics
  - transcriptomics
- Commercial antibodies information
- Contains information from other projects such as GTEx or TCGA

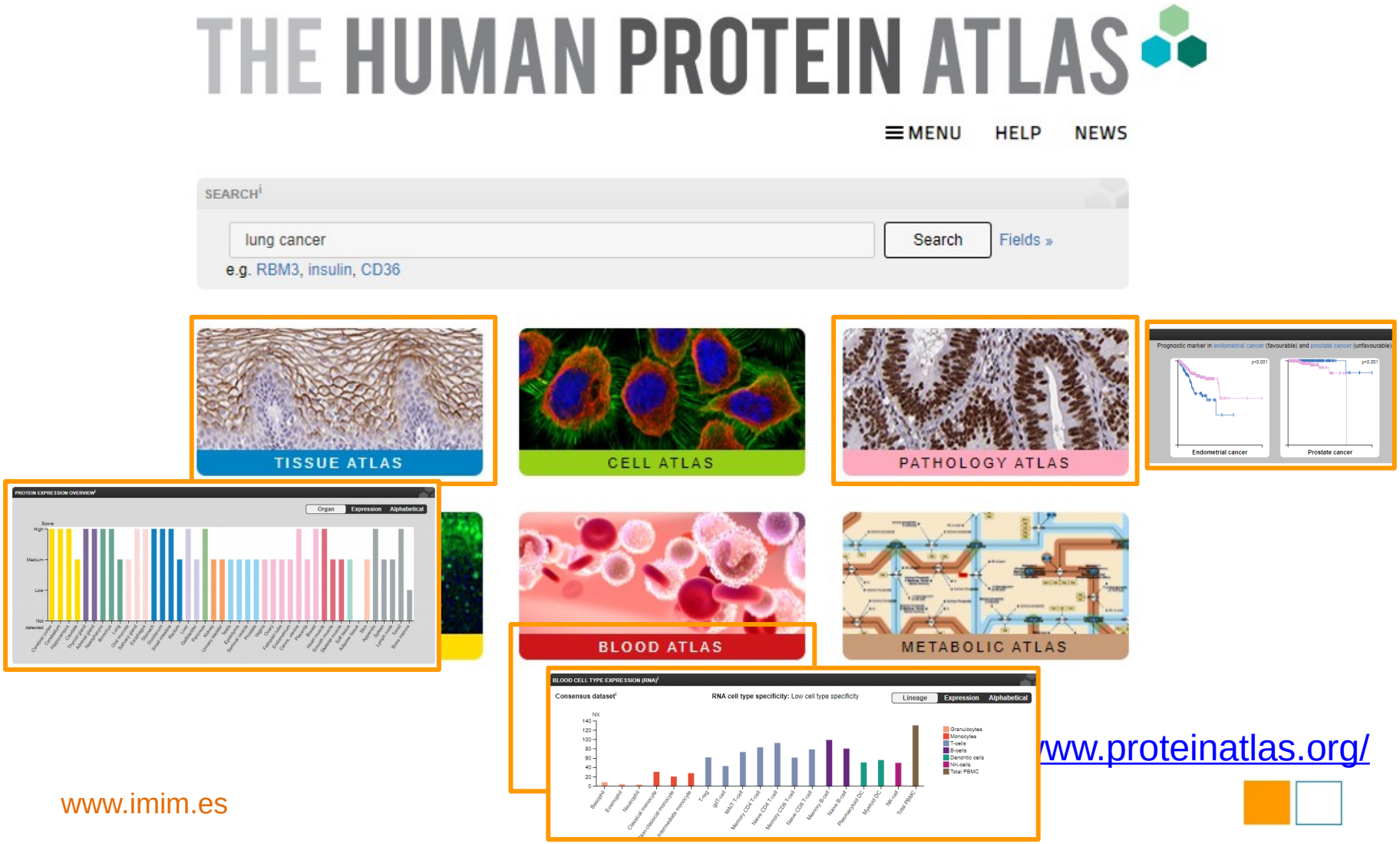
https://www.proteinatlas.org/

2. The Human Protein Atlas



https://www.proteinatlas.org/

2. The Human Protein Atlas



www.proteinatlas.org/

2. cBioPortal

Resource for interactive exploration of multidimensional cancer genomics data sets

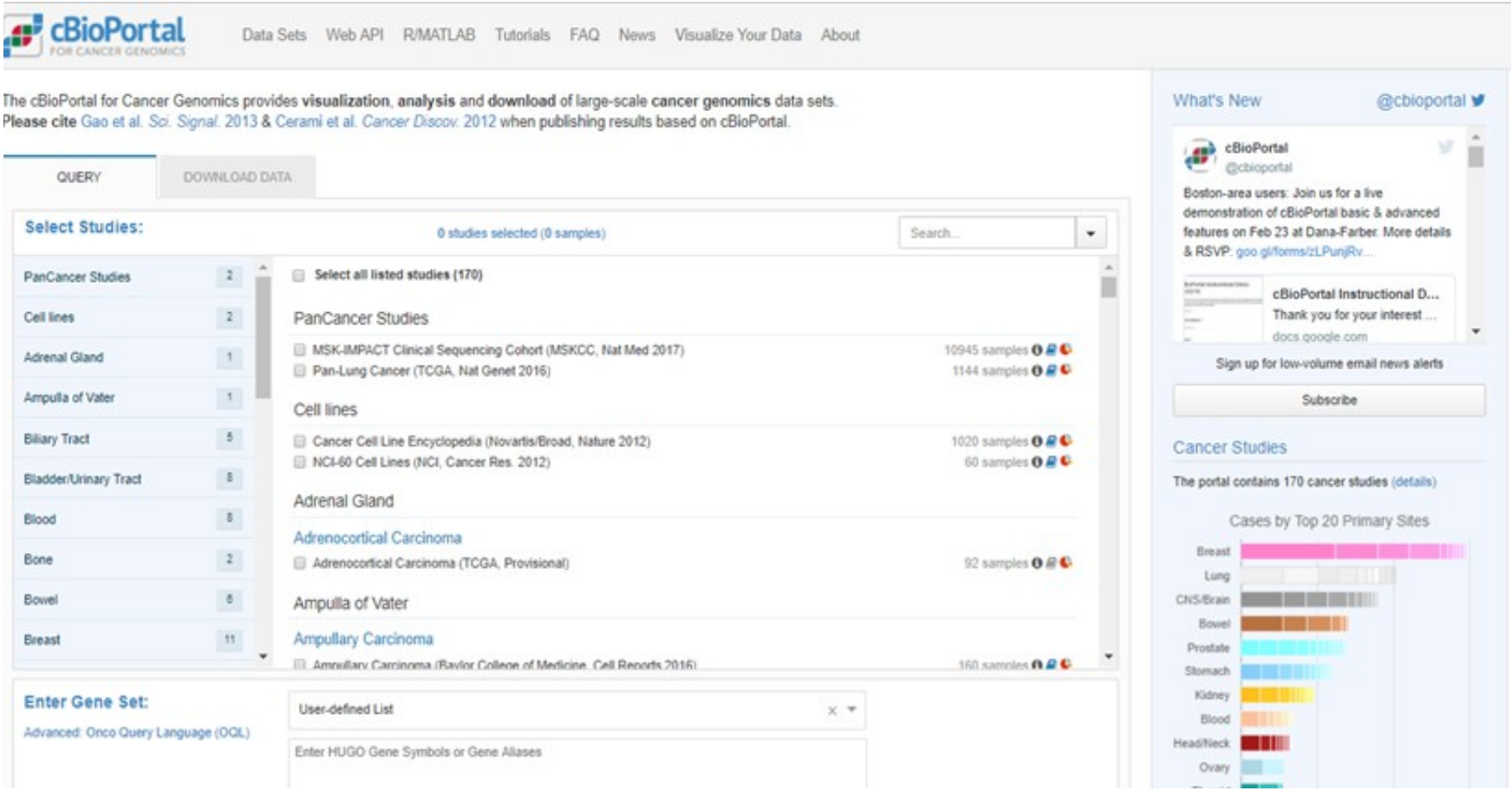
- Originally developed at Memorial Sloan Kettering Cancer Center (MSK) now developed and maintained by a multi-institutional team.
- Available information:
  - non-synonymous mutations
  - DNA copy-number data
  - mRNA and microRNA expression data
  - protein-level and phosphoprotein level data (RPPA or mass spectrometry)
  - DNA methylation data
  - clinical data (de-identified)
- Free to use in publications or presentations (with citation)
- No normal tissue data
- 275 cancer data sets, including TCGA and other public data

2. cBioPortal

Query studies:

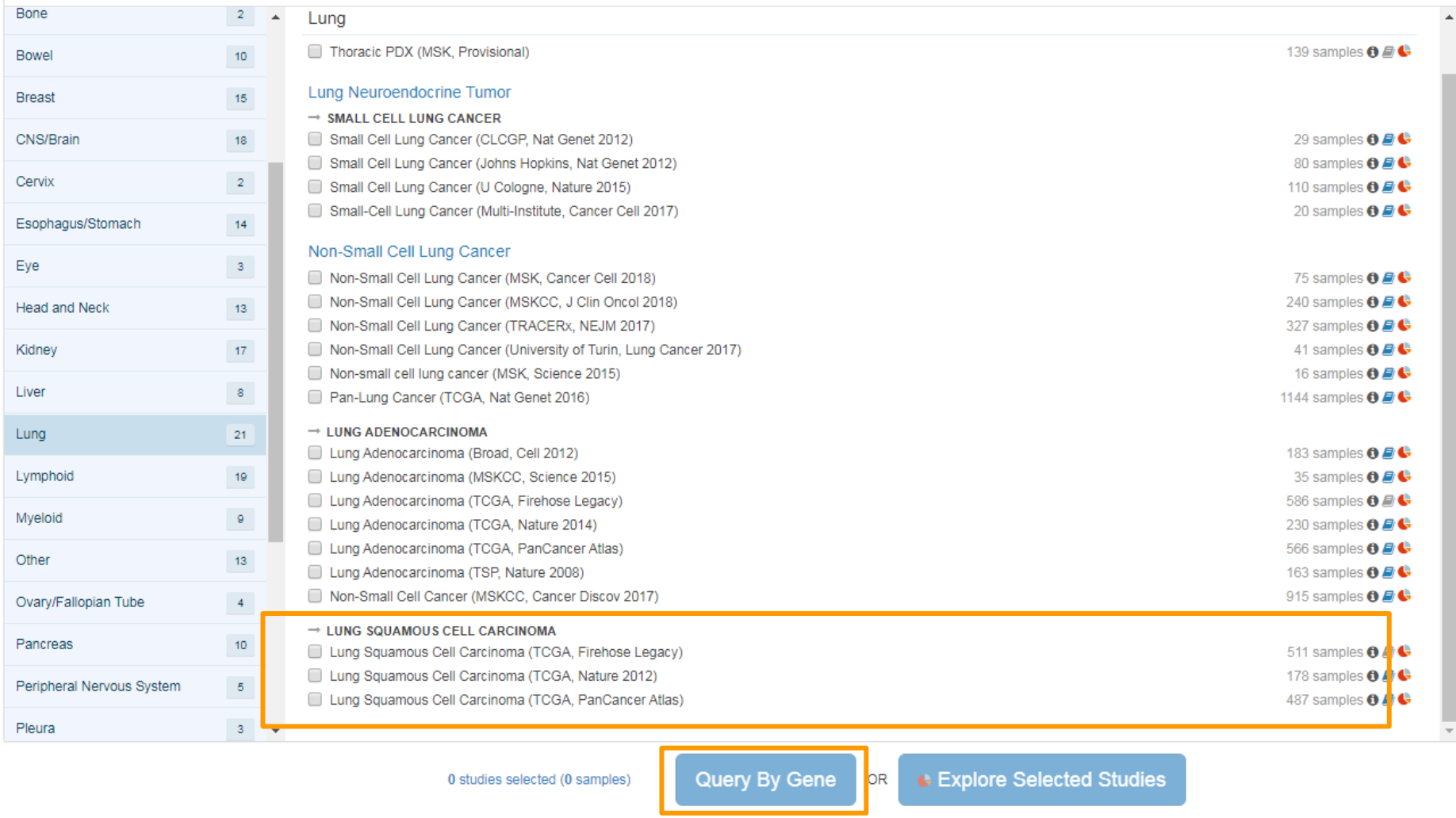
- All (or case/sample selection)
  - Summary
    - Survival
    - Mutations
    - CN
  - Clinical
  - Heat maps
  - CN segments
- By gene
- Allows filtering by clinical variables
- Visualize your own data

2. cBioPortal

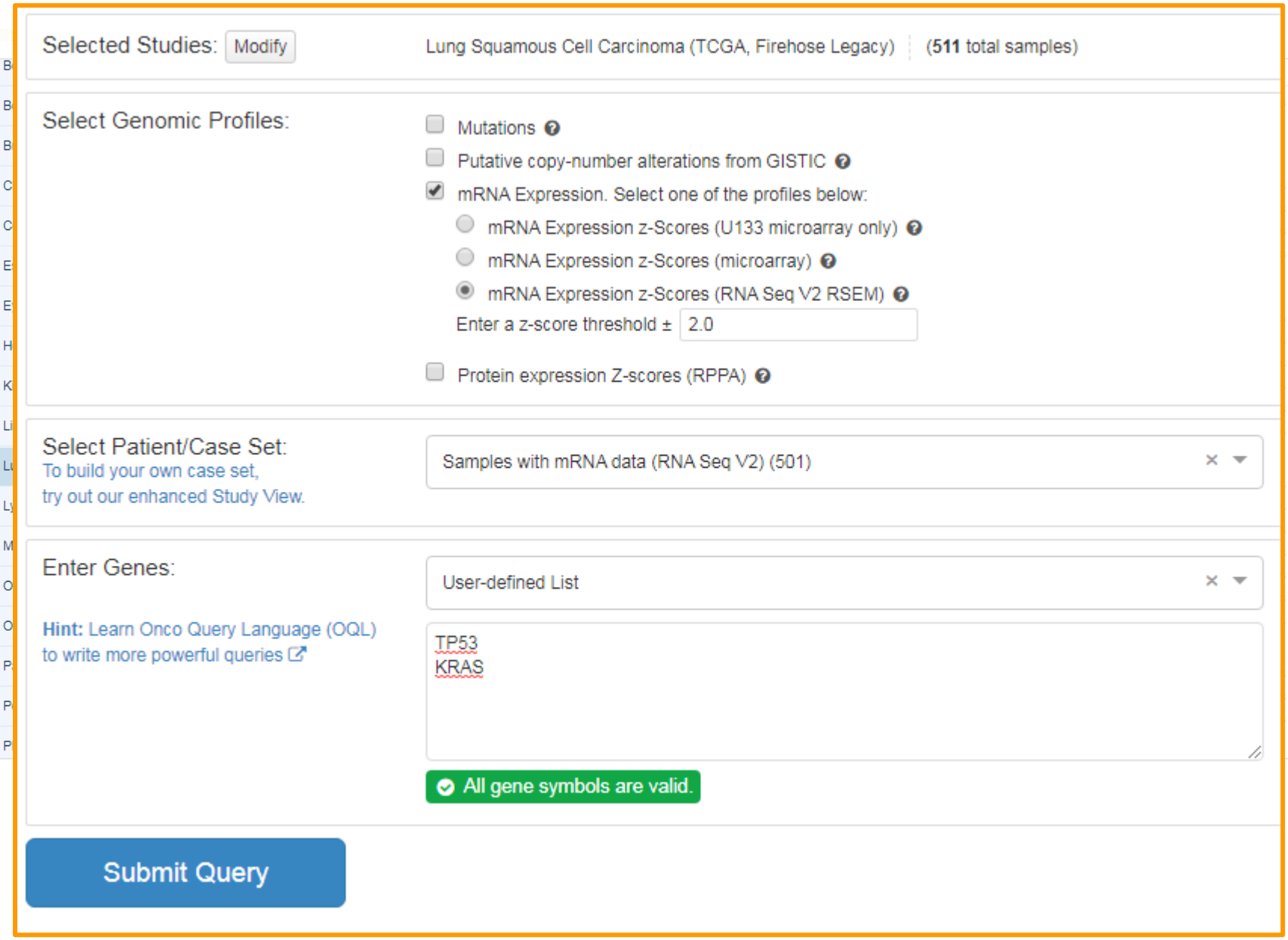


https://www.cbioportal.org/

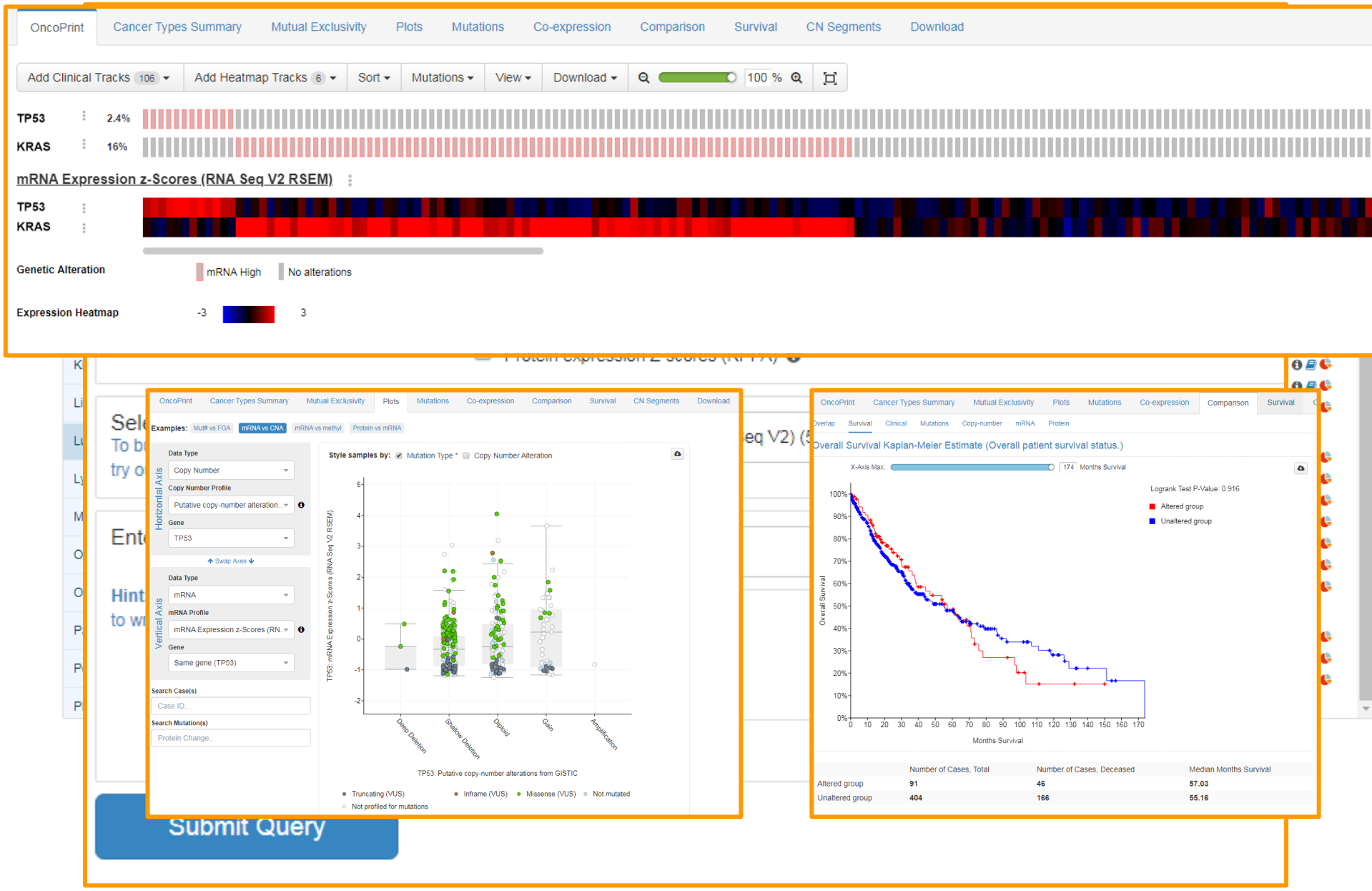
2. cBioPortal. LUSC



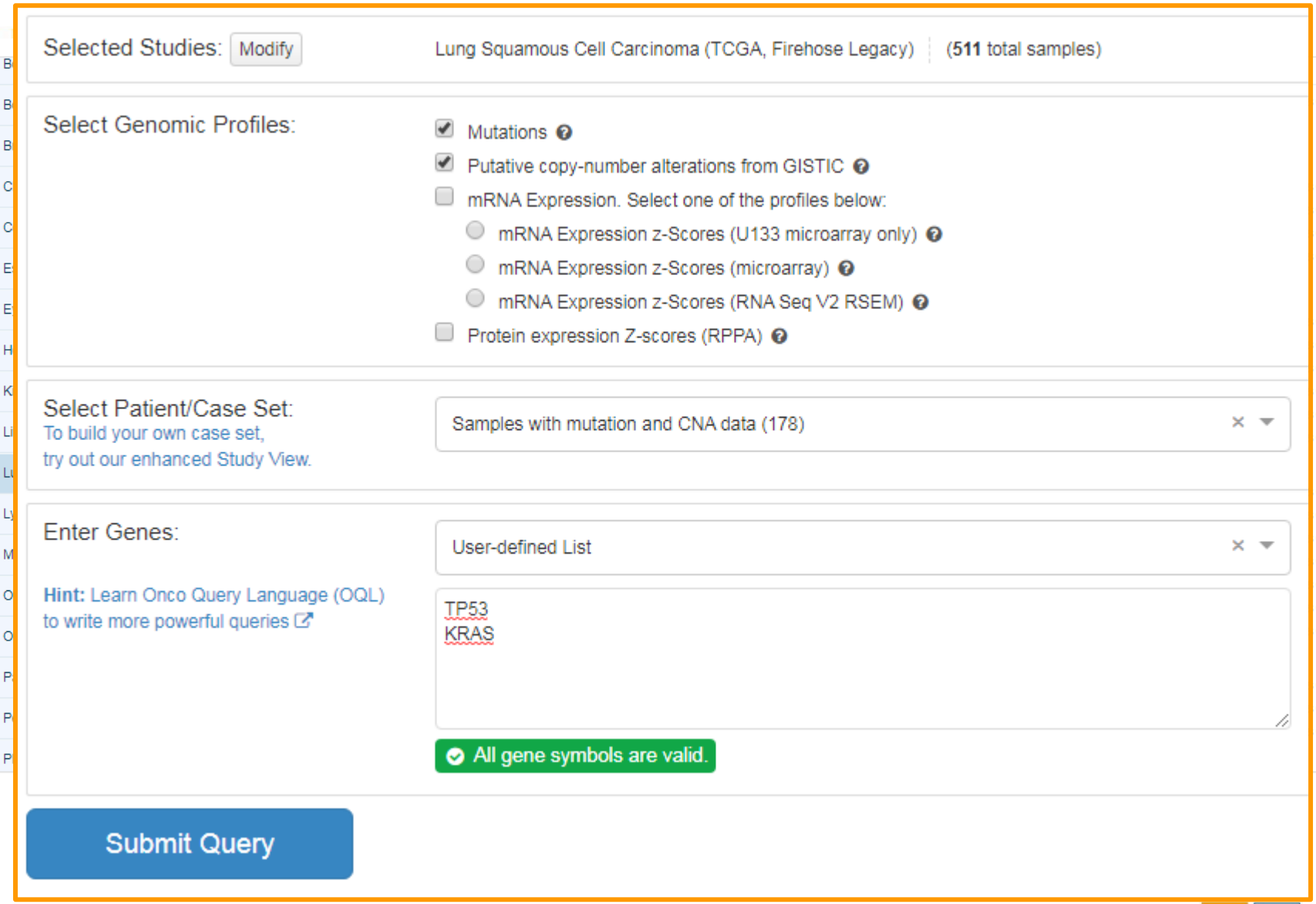
2. cBioPortal. LUSC



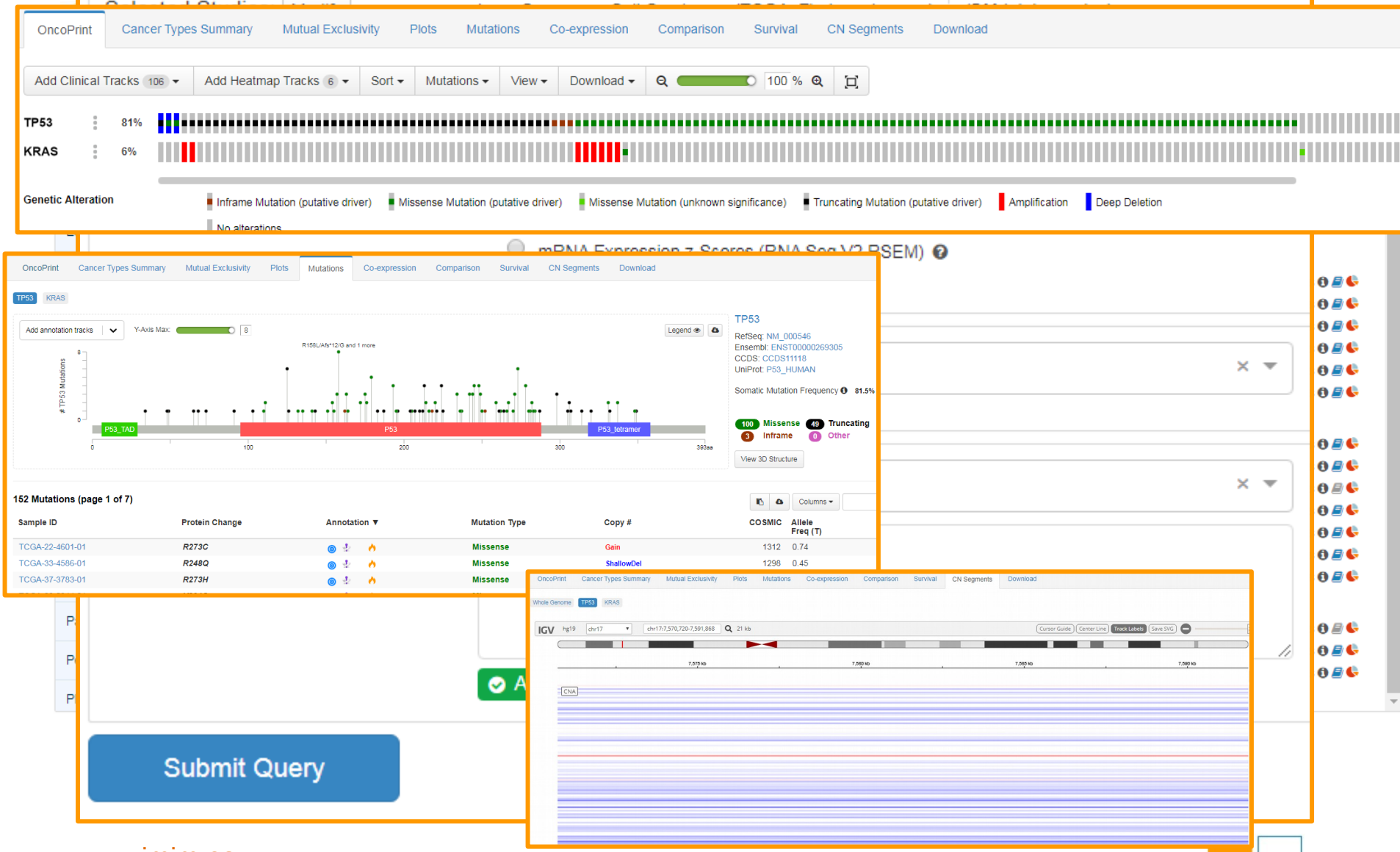
2. cBioPortal. LUSC



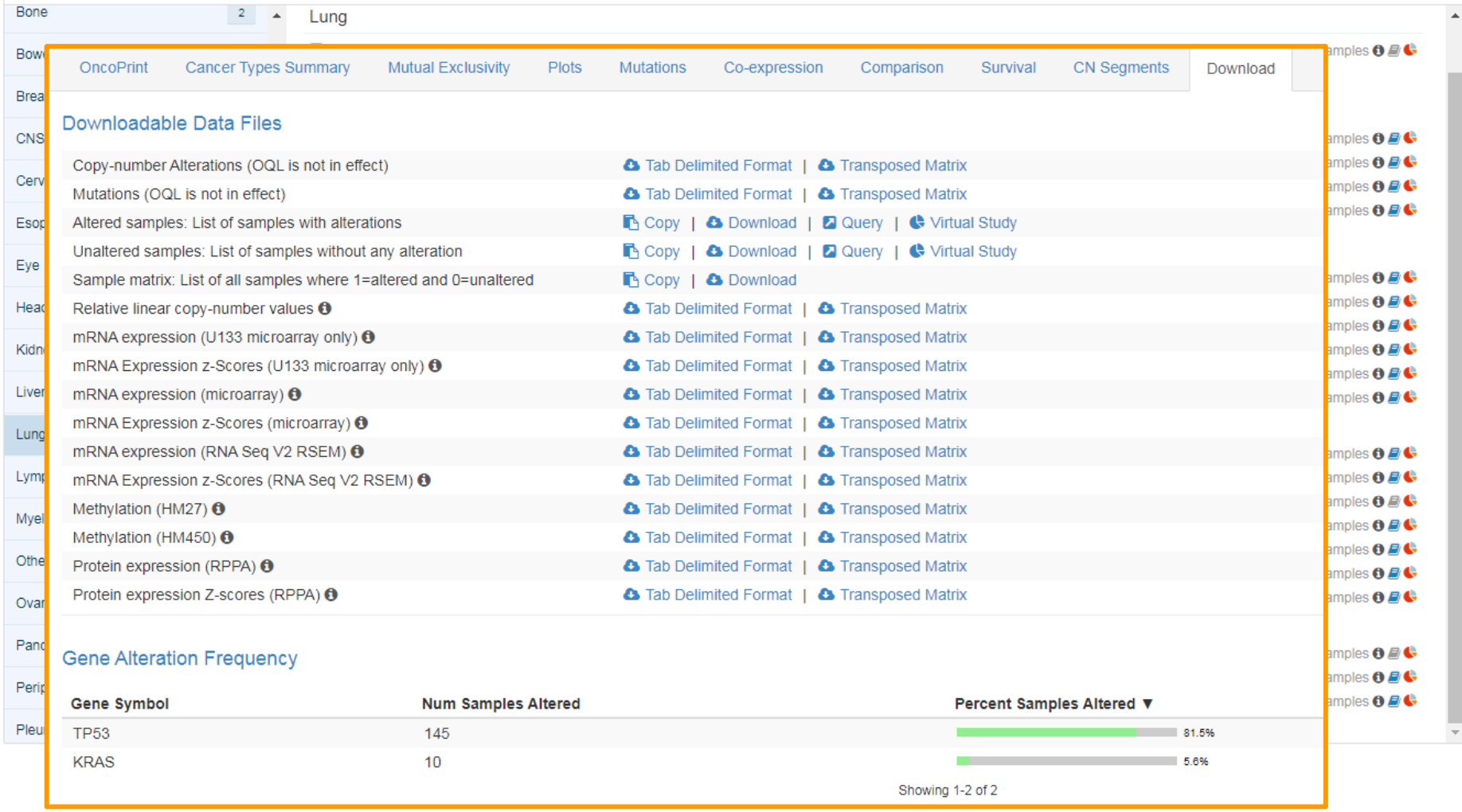
2. cBioPortal. LUSC



2. cBioPortal. LUSC



2. cBioPortal. LUSC



2. GEPIA

Gene Expression Profiling Interactive Analysis (GEPIA)

- Interactive web server for analyzing the RNA sequencing expression data
- 9,736 tumors (TCGA)
- 8,587 normal samples (TCGA and GTEx)
- Standard processing pipeline
- Customizable functions:
  - tumor/normal differential expression analysis
  - profiling according to cancer types or pathological stages
  - patient survival analysis
  - similar gene detection
  - correlation analysis
  - dimensionality reduction analysis.
- GEPIA2 (beta) available
  - Custom data

W98–W102 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 12 April 2017*
doi: 10.1093/nar/gkx247

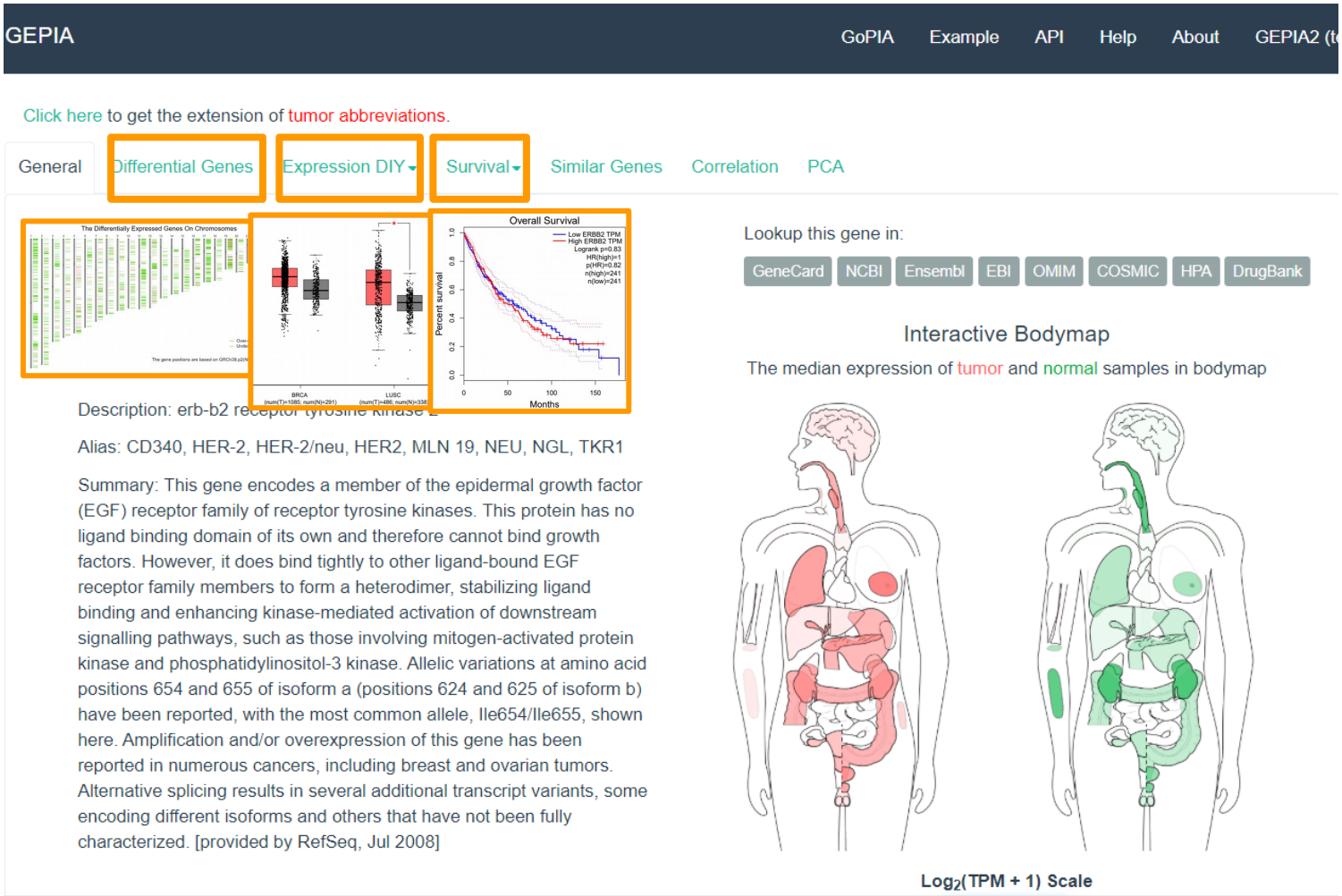
GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses

Zefang Tang[1,†], Chenwei Li[1,2,†], Boxi Kang[1], Ge Gao[3], Cheng Li[2,3] and Zemin Zhang[1,2,4,5,*]

2. GEPIA

http://gepia.cancer-pku.cn/

2. GEPIA

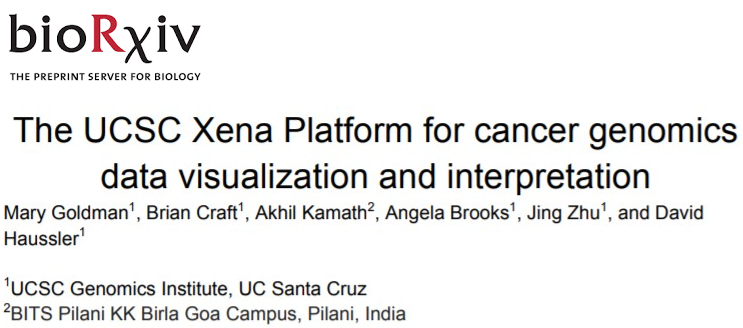


http://gepia.cancer-pku.cn/

2. UCSC Xena

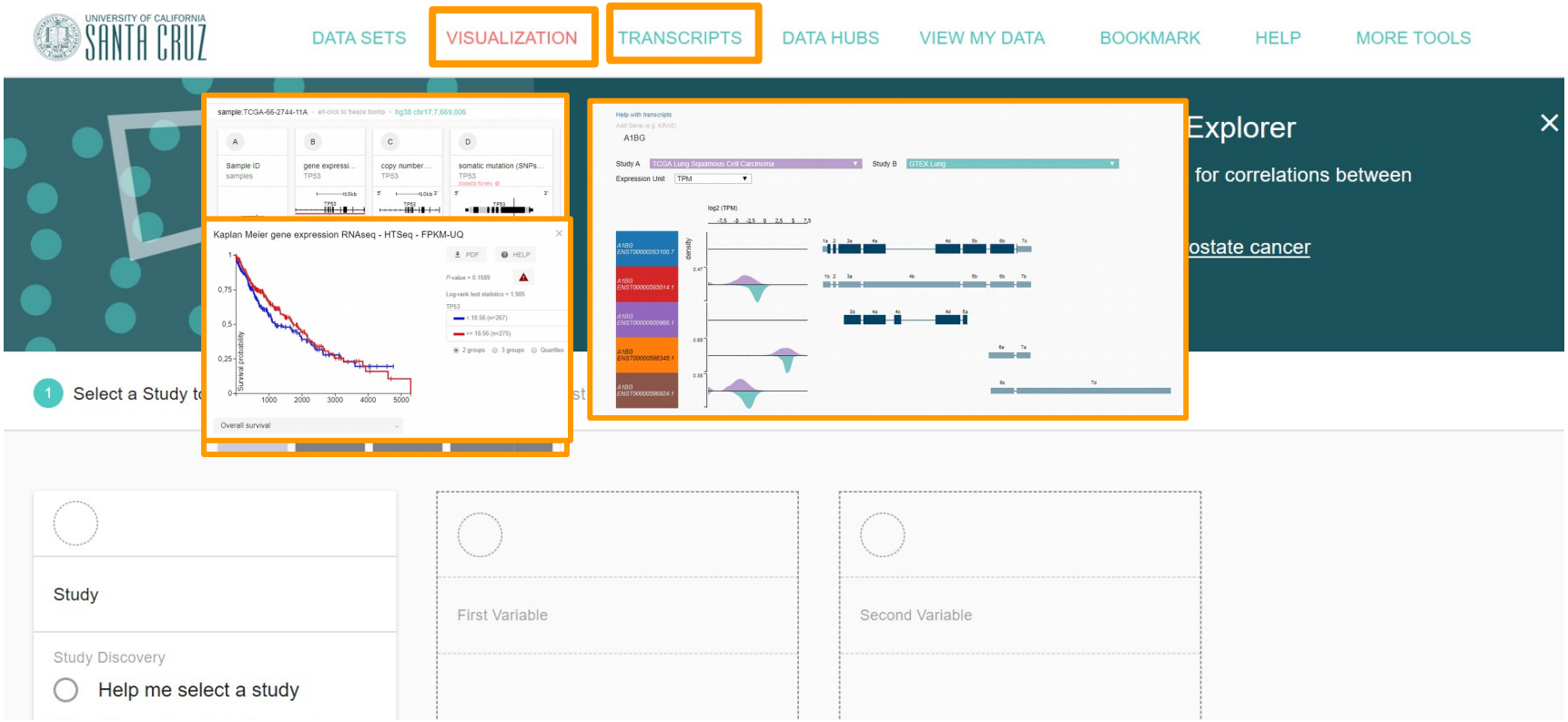
A platform for functional genomics visualization and analysis

- UCSC
- More than 1500 datasets across 50 cancer types
- Resources:
  - TCGA (including Pan-Cancer Atlas)
  - ICGC
  - Cell line encyclopedia
  - GTEx
- Private or your own data



https://xena.ucsc.edu/

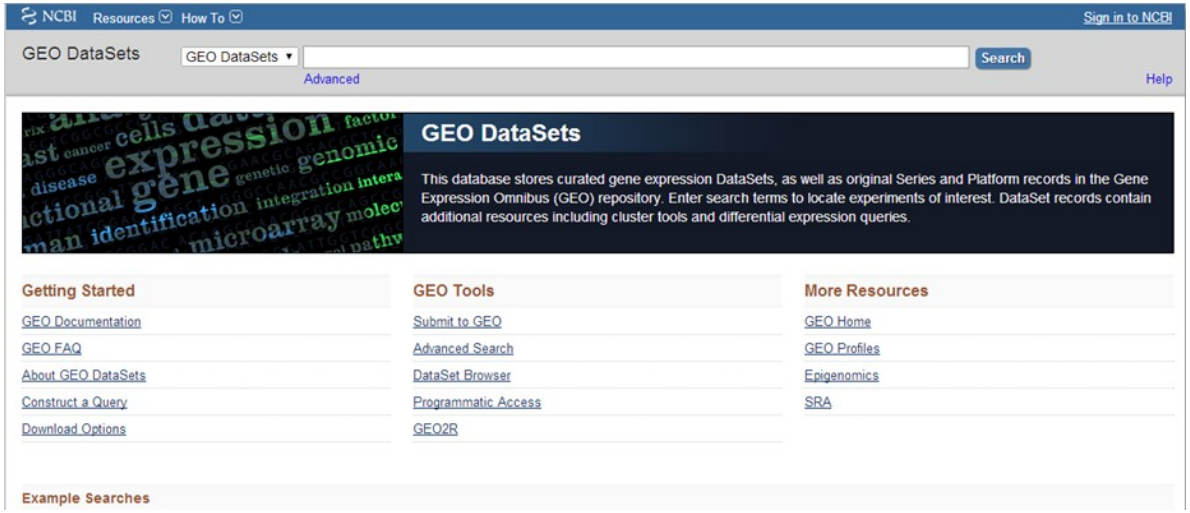
2. UCSC Xena


UNIVERSITY OF CALIFORNIA SANTA CRUZ
DATA SETS
VISUALIZATION
TRANSCRIPTS
DATA HUBS
VIEW MY DATA
BOOKMARK
HELP
MORE TOOLS
Explorer
for correlations between
ostate cancer
Kaplan Meier gene expression RNAseq - HTSeq - FPKM-UQ
Select a Study
Study
Study Discovery
Help me select a study
First Variable
Second Variable

2. GEO

Gene Expression Omnibus (GEO) is a public functional high-throughput genomics data repository supporting MIAME-compliant data submissions.

- >120.000 experiment data sets
- “Needed” to publish high-throughput data
- SRA for raw sequencing files

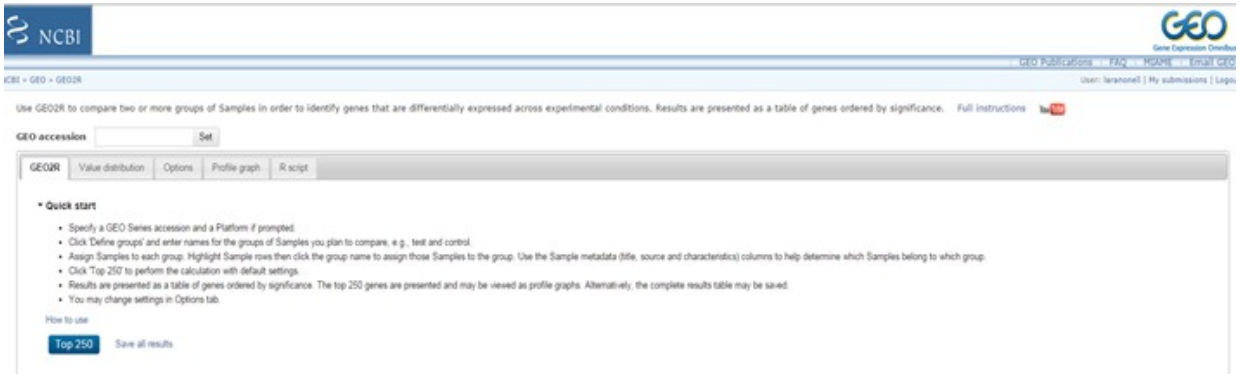


2. GEO

- Platform: (GPLxxx) describes the list of elements on the array (e.g., cDNAs, oligonucleotide probesets, ORFs, antibodies) or the list of elements that may be detected and quantified in that experiment (e.g., SAGE tags, peptides).
- Series (GSExxx): Set of related Samples considered to be part of a group and provides a focal point and description of the experiment as a whole. Series records may also contain tables describing extracted data, summary conclusions, or analyses
- GEO DataSets (GDSxxx): Curated sets of GEO Sample data. Samples within a GDS refer to the same Platform.
- Sample (GSMxxx): Describes the conditions of a Sample.

2. GEO2R

- **GEO2R** is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions based on the original submitter-supplied processed data tables and using the GEOquery and limma R packages from the Bioconductor project.
- Results are presented as a table of genes ordered by significance.
- Is available for each GEO Series and interrogates the original Series Matrix data file directly



2. GEO2R. Data

SCIENTIFIC REPORTS

OPEN

Transcriptomic and functional network features of lung squamous cell carcinoma through integrative analysis of GEO and TCGA data

Received: 31 May 2018
Accepted: 12 October 2018
Published online: 26 October 2018

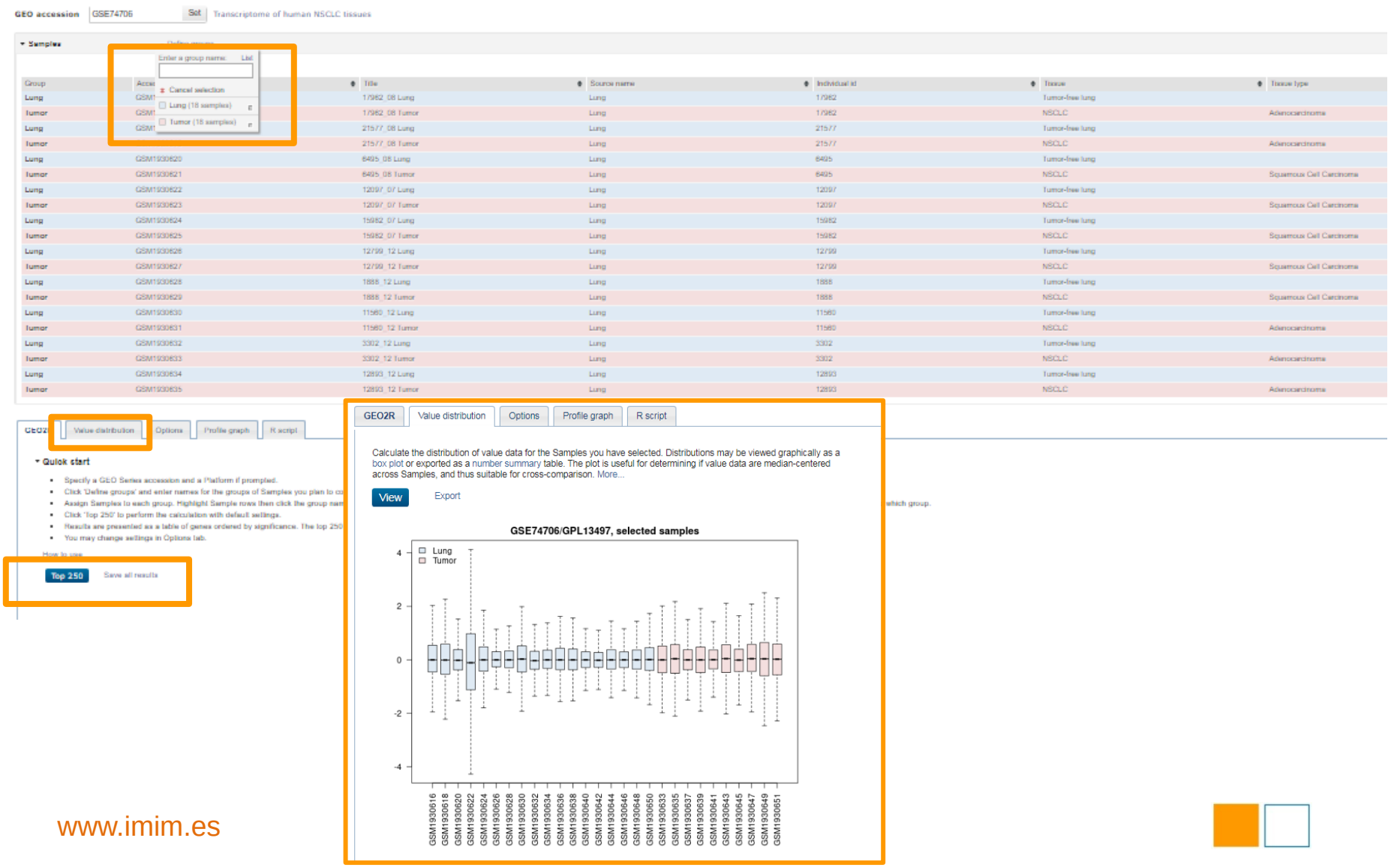
Yin Li, Jie Gu, Fengkai Xu, Qiaoliang Zhu, Di Ge & Chunlai Lu

Lung squamous cell carcinoma (LUSC) is associated with poor clinical prognosis and lacks available targeted therapy. Novel molecules are urgently required for the diagnosis and prognosis of LUSC. Here, we conducted our data mining analysis for LUSC by integrating the differentially expressed genes acquired from Gene Expression Omnibus (GEO) database by comparing tumor tissues versus normal tissues (GSE8569, GSE21933, GSE33479, GSE33532, GSE40275, GSE62113, GSE74706) into The Cancer Genome Atlas (TCGA) database which includes 502 tumors and 49 adjacent non-tumor lung tissues. We identified intersections of 129 genes (91 up-regulated and 38 down-regulated) between GEO data and TCGA data. Based on these genes, we conducted our downstream analysis including functional enrichment analysis, protein-protein interaction, competing endogenous RNA (ceRNA) network and survival analysis. This study may provide more insight into the transcriptomic and functional features of LUSC through integrative analysis of GEO and TCGA data and suggests therapeutic targets and biomarkers for LUSC.

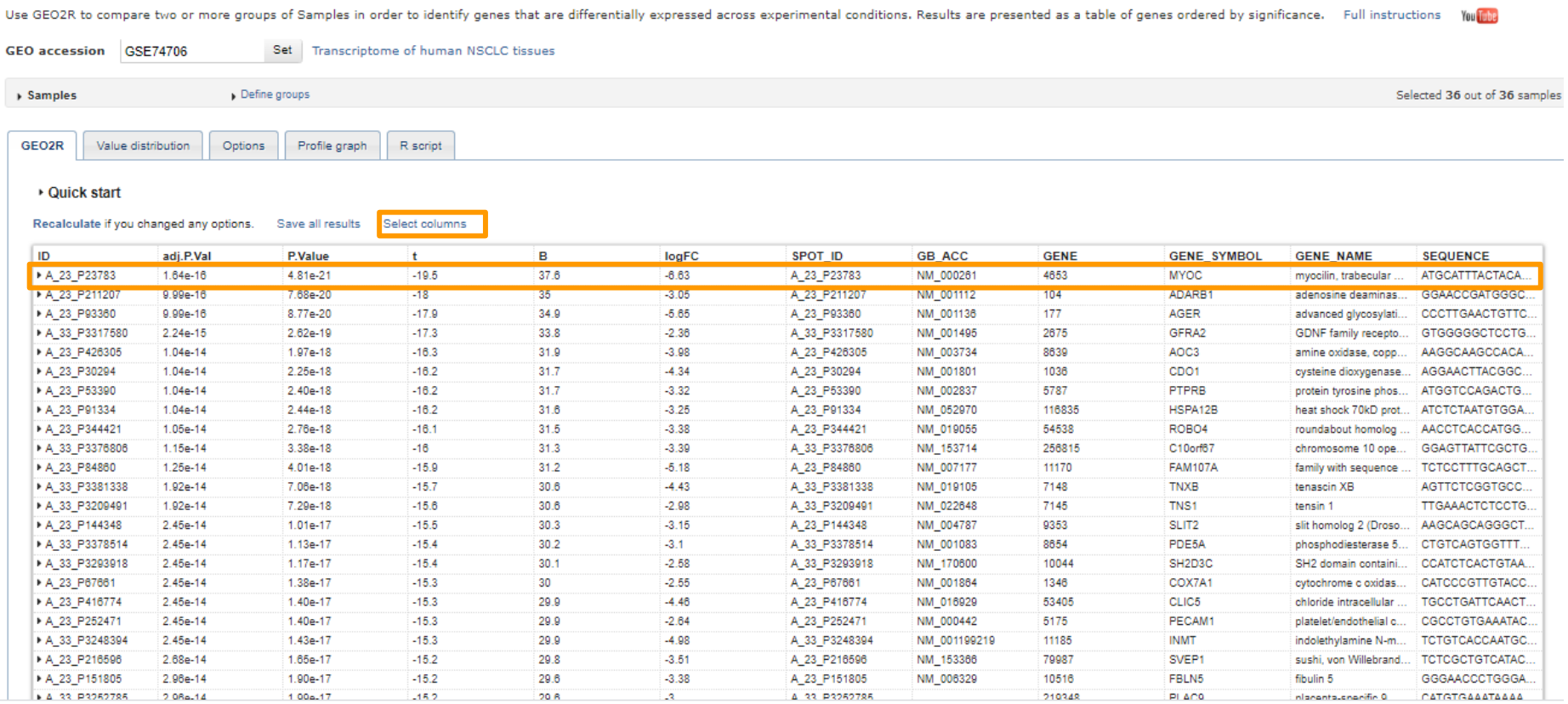
2. GEO2R. Data

GSE74706

2. GEO2R. Analysis



2. GEO2R. Results



Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions

GEO accession GSE74706 Set Transcriptome of human NSCLC tissues

Samples Define groups Selected 36 out of 36 samples

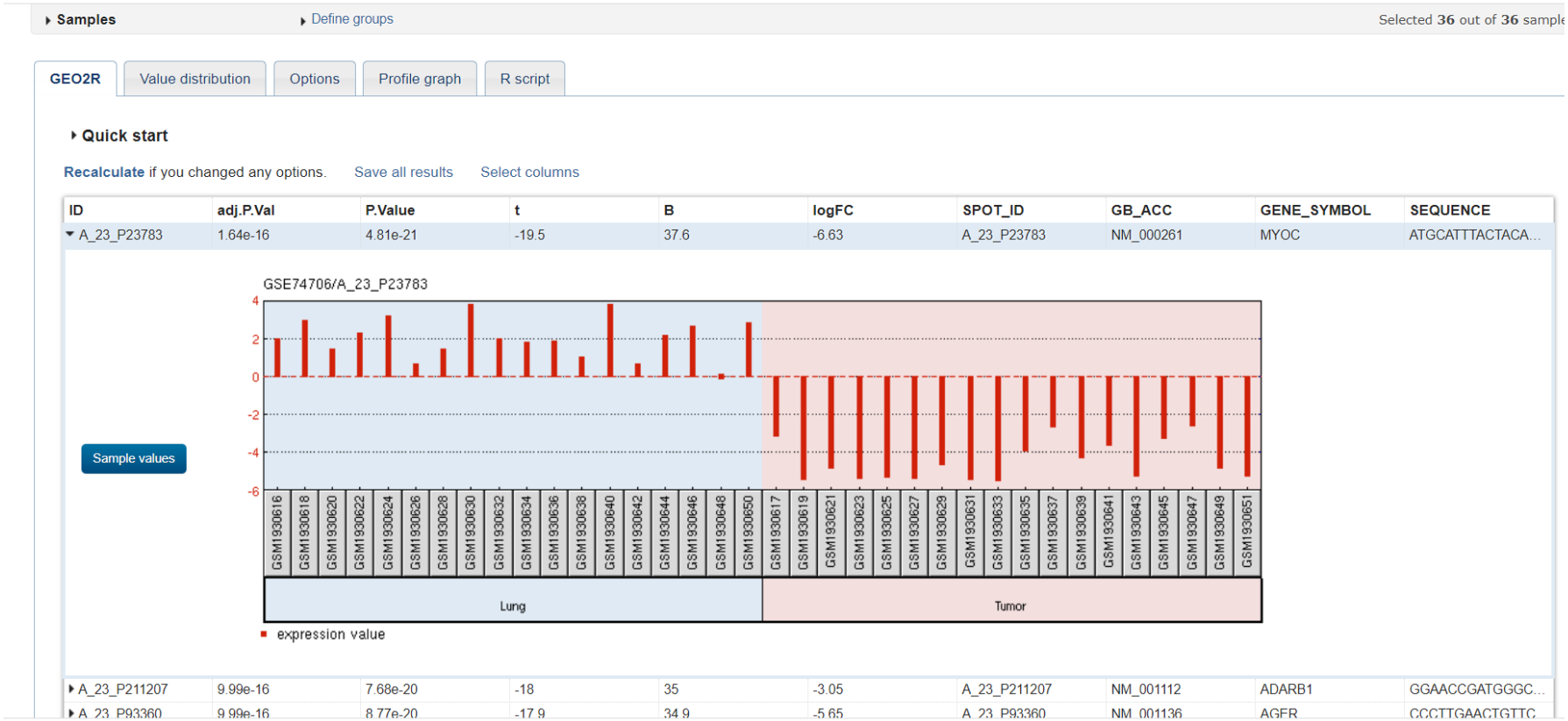
GEO2R | Value distribution | Options | Profile graph | R script

Quick start

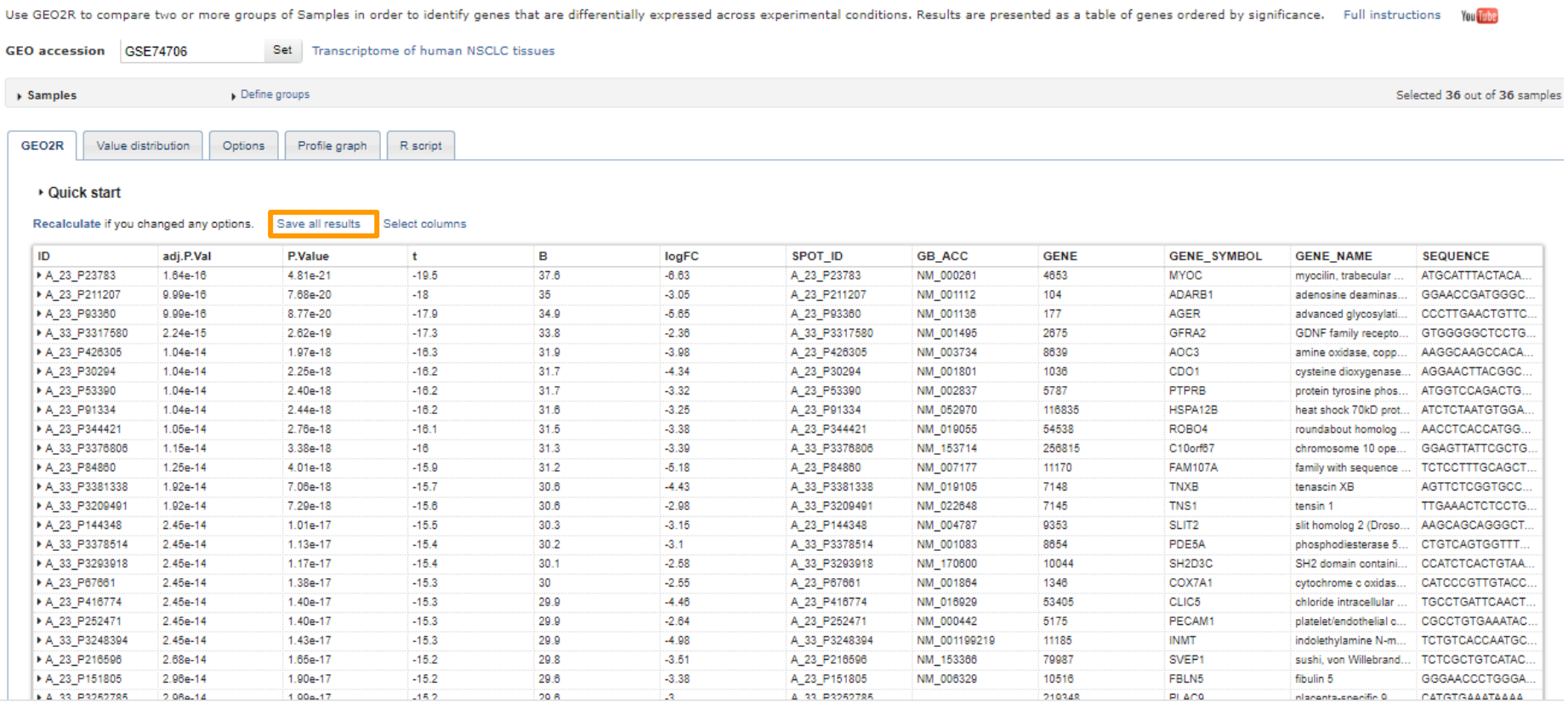
Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	SPOT_ID	GB_ACC	GENE	GENE_SYMBOL	GENE_NAME	SEQUENCE
A_23_P23783	1.64e-16	4.81e-21	-19.5	37.6	-6.63	A_23_P23783	NM_000261	4653	MYOC	myocilin, trabecular ...	ATGCATTTACTACA...
A_23_P211207	9.99e-16	7.68e-20	-18	35	-3.05	A_23_P211207	NM_001112	104	ADARB1	adenosine deaminas...	GGAACCGATGGGC...
A_23_P93360	9.99e-16	8.77e-20	-17.9	34.9	-5.65	A_23_P93360	NM_001136	177	AGER	advanced glycosylati...	CCCTTGAACTGTTC...
A_33_P3317580	2.24e-15	2.62e-19	-17.3	33.8	-2.36	A_33_P3317580	NM_001495	2675	GFRA2	GDNF family recepto...	GTGGGGGCTCCTG...
A_23_P426305	1.04e-14	1.97e-18	-16.3	31.9	-3.98	A_23_P426305	NM_003734	8639	AOC3	amine oxidase, copp...	AAGGCAAGCCACA...
A_23_P30294	1.04e-14	2.25e-18	-16.2	31.7	-4.34	A_23_P30294	NM_001801	1036	CDO1	cysteine dioxygenase...	AGGAACTTACGGC...
A_23_P53390	1.04e-14	2.40e-18	-16.2	31.7	-3.32	A_23_P53390	NM_002837	5787	PTPRB	protein tyrosine phos...	ATGGTCCAGACTG...
A_23_P91334	1.04e-14	2.44e-18	-16.2	31.6	-3.25	A_23_P91334	NM_052970	116835	HSPA12B	heat shock 70kD prot...	ATCTCTAATGTGGA...
A_23_P344421	1.05e-14	2.76e-18	-16.1	31.5	-3.38	A_23_P344421	NM_019055	54538	ROBO4	roundabout homolog ...	AACCTCACCATGG...
A_33_P3376806	1.15e-14	3.38e-18	-16	31.3	-3.39	A_33_P3376806	NM_153714	256815	C10orf67	chromosome 10 ope...	GGAGTTATTCGCTG...
A_23_P84860	1.25e-14	4.01e-18	-15.9	31.2	-5.18	A_23_P84860	NM_007177	11170	FAM107A	family with sequence ...	TCTCCTTTGCAGCT...
A_33_P3381338	1.92e-14	7.06e-18	-15.7	30.6	-4.43	A_33_P3381338	NM_019105	7148	TNXB	tenascin XB	AGTTCTCGGTGCC...
A_33_P3209491	1.92e-14	7.29e-18	-15.6	30.6	-2.98	A_33_P3209491	NM_022648	7145	TNS1	tensin 1	TTGAAACTCTCCTG...
A_23_P144348	2.45e-14	1.01e-17	-15.5	30.3	-3.15	A_23_P144348	NM_004787	9353	SLIT2	slit homolog 2 (Droso...	AAGCAGCAGGGCT...
A_33_P3378514	2.45e-14	1.13e-17	-15.4	30.2	-3.1	A_33_P3378514	NM_001083	8654	PDE5A	phosphodiesterase 5...	CTGTCAGTGGTTT...
A_33_P3293918	2.45e-14	1.17e-17	-15.4	30.1	-2.58	A_33_P3293918	NM_170600	10044	SH2D3C	SH2 domain containi...	CCATCTCACTGTAA...
A_23_P67661	2.45e-14	1.38e-17	-15.3	30	-2.55	A_23_P67661	NM_001864	1346	COX7A1	cytochrome c oxidas...	CATCCCGTTGTACC...
A_23_P416774	2.45e-14	1.40e-17	-15.3	29.9	-4.46	A_23_P416774	NM_016929	53405	CLIC5	chloride intracellular ...	TGCCTGATTCAACT...
A_23_P252471	2.45e-14	1.40e-17	-15.3	29.9	-2.64	A_23_P252471	NM_000442	5175	PECAM1	platelet/endothelial c...	CGCCTGTGAAATAC...
A_33_P3248394	2.45e-14	1.43e-17	-15.3	29.9	-4.98	A_33_P3248394	NM_001199219	11185	INMT	indolethylamine N-m...	TCTGTCACCAATGC...
A_23_P216596	2.68e-14	1.65e-17	-15.2	29.8	-3.51	A_23_P216596	NM_153366	79987	SVEP1	sushi, von Willebrand...	TCTCGCTGTCATAC...
A_23_P151805	2.96e-14	1.90e-17	-15.2	29.6	-3.38	A_23_P151805	NM_006329	10516	FBLN5	fibulin 5	GGGAACCCTGGGA...
A_33_P3252785	2.96e-14	1.99e-17	-15.2	29.6	-3	A_33_P3252785		219348	PLAC9	placenta-specific 9	CATGTGAAATAAAA...

2. GEO2R. Results



2. GEO2R. Results



Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions

GEO accession GSE74706 Set Transcriptome of human NSCLC tissues

Samples Define groups Selected 36 out of 36 samples

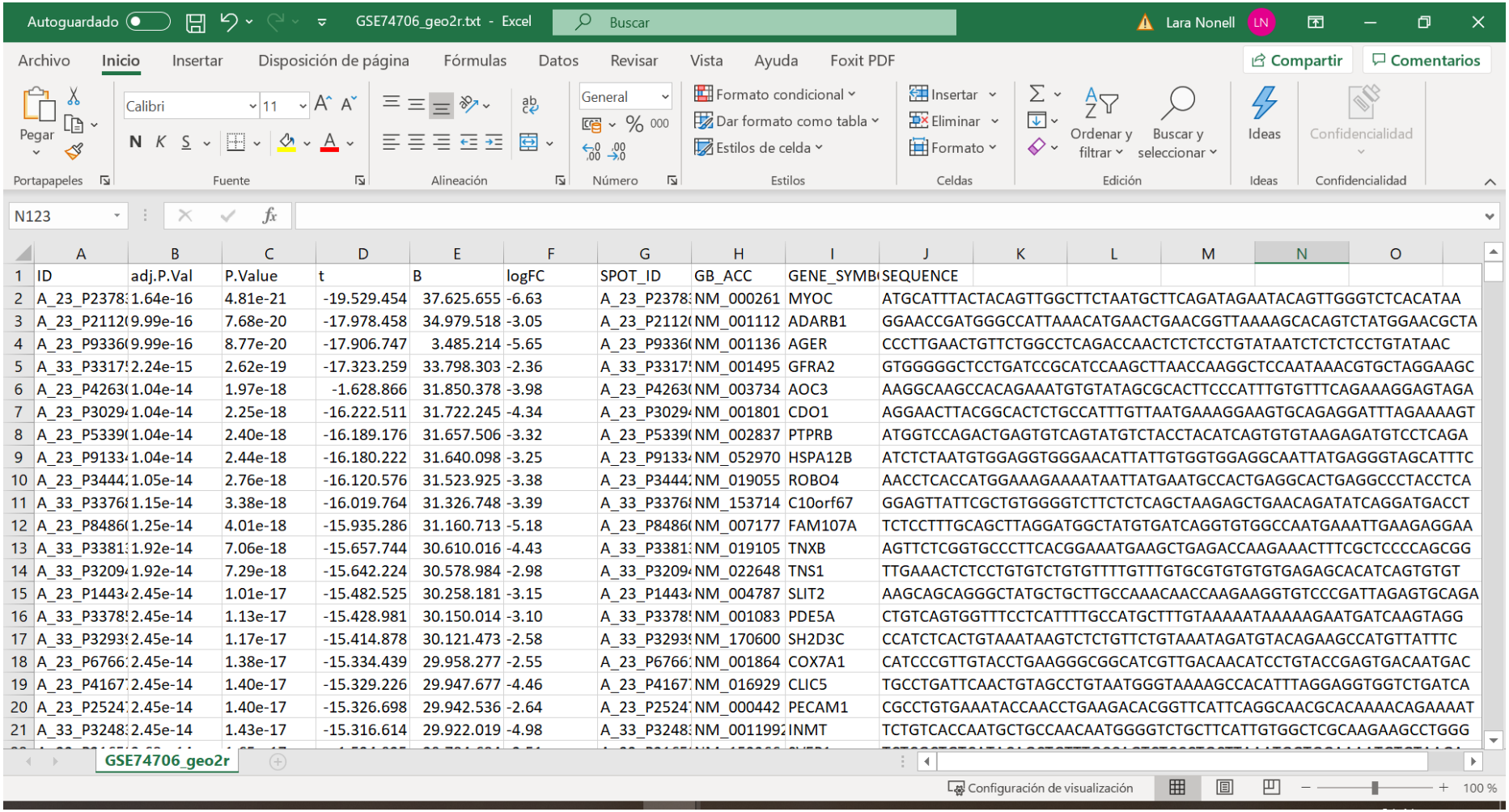
GEO2R | Value distribution | Options | Profile graph | R script

Quick start

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	SPOT_ID	GB_ACC	GENE	GENE_SYMBOL	GENE_NAME	SEQUENCE
A_23_P23783	1.64e-16	4.81e-21	-19.5	37.6	-6.63	A_23_P23783	NM_000261	4653	MYOC	myocilin, trabecular ...	ATGCATTTACTACA...
A_23_P211207	9.99e-16	7.68e-20	-18	35	-3.05	A_23_P211207	NM_001112	104	ADARB1	adenosine deaminas...	GGAACCGATGGGC...
A_23_P93360	9.99e-16	8.77e-20	-17.9	34.9	-5.65	A_23_P93360	NM_001136	177	AGER	advanced glycosylati...	CCCTTGAACTGTTC...
A_33_P3317580	2.24e-15	2.62e-19	-17.3	33.8	-2.36	A_33_P3317580	NM_001495	2675	GFRA2	GDNF family recepto...	GTGGGGGCTCCTG...
A_23_P426305	1.04e-14	1.97e-18	-16.3	31.9	-3.98	A_23_P426305	NM_003734	8639	AOC3	amine oxidase, copp...	AAGGCAAGCCACA...
A_23_P30294	1.04e-14	2.25e-18	-16.2	31.7	-4.34	A_23_P30294	NM_001801	1036	CDO1	cysteine dioxygenase...	AGGAACTTACGGC...
A_23_P53390	1.04e-14	2.40e-18	-16.2	31.7	-3.32	A_23_P53390	NM_002837	5787	PTPRB	protein tyrosine phos...	ATGGTCCAGACTG...
A_23_P91334	1.04e-14	2.44e-18	-16.2	31.6	-3.25	A_23_P91334	NM_052970	116835	HSPA12B	heat shock 70kD prot...	ATCTCTAATGTGGA...
A_23_P344421	1.05e-14	2.76e-18	-16.1	31.5	-3.38	A_23_P344421	NM_019055	54538	ROBO4	roundabout homolog ...	AACCTCACCATGG...
A_33_P3376806	1.15e-14	3.38e-18	-16	31.3	-3.39	A_33_P3376806	NM_153714	256815	C10orf67	chromosome 10 ope...	GGAGTTATTCGCTG...
A_23_P84860	1.25e-14	4.01e-18	-15.9	31.2	-5.18	A_23_P84860	NM_007177	11170	FAM107A	family with sequence ...	TCTCCTTTGCAGCT...
A_33_P3381338	1.92e-14	7.06e-18	-15.7	30.6	-4.43	A_33_P3381338	NM_019105	7148	TNXB	tenascin XB	AGTTCTCGGTGCC...
A_33_P3209491	1.92e-14	7.29e-18	-15.6	30.6	-2.98	A_33_P3209491	NM_022648	7145	TNS1	tensin 1	TTGAAACTCTCCTG...
A_23_P144348	2.45e-14	1.01e-17	-15.5	30.3	-3.15	A_23_P144348	NM_004787	9353	SLIT2	slit homolog 2 (Droso...	AAGCAGCAGGGCT...
A_33_P3378514	2.45e-14	1.13e-17	-15.4	30.2	-3.1	A_33_P3378514	NM_001083	8654	PDE5A	phosphodiesterase 5...	CTGTCAGTGGTTT...
A_33_P3293918	2.45e-14	1.17e-17	-15.4	30.1	-2.58	A_33_P3293918	NM_170600	10044	SH2D3C	SH2 domain containi...	CCATCTCACTGTAA...
A_23_P67661	2.45e-14	1.38e-17	-15.3	30	-2.55	A_23_P67661	NM_001864	1346	COX7A1	cytochrome c oxidas...	CATCCCGTTGTACC...
A_23_P416774	2.45e-14	1.40e-17	-15.3	29.9	-4.46	A_23_P416774	NM_016929	53405	CLIC5	chloride intracellular ...	TGCCTGATTCAACT...
A_23_P252471	2.45e-14	1.40e-17	-15.3	29.9	-2.64	A_23_P252471	NM_000442	5175	PECAM1	platelet/endothelial c...	CGCCTGTGAAATAC...
A_33_P3248394	2.45e-14	1.43e-17	-15.3	29.9	-4.98	A_33_P3248394	NM_001199219	11185	INMT	indolethylamine N-m...	TCTGTCACCAATGC...
A_23_P216596	2.68e-14	1.65e-17	-15.2	29.8	-3.51	A_23_P216596	NM_153366	79987	SVEP1	sushi, von Willebrand...	TCTCGCTGTCATAC...
A_23_P151805	2.96e-14	1.90e-17	-15.2	29.6	-3.38	A_23_P151805	NM_006329	10516	FBLN5	fibulin 5	GGGAACCCTGGGA...
A_33_P3252785	2.96e-14	1.99e-17	-15.2	29.6	-3	A_33_P3252785		219348	PLAC9	placenta-specific 9	CATGTGAAATAAAA...

2. GEO2R. Results



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	adj.P.Val	P.Value	t	B	logFC	SPOT_ID	GB_ACC	GENE_SYMB	SEQUENCE					
2	A_23_P2378	1.64e-16	4.81e-21	-19.529.454	37.625.655	-6.63	A_23_P2378	NM_000261	MYOC	ATGCATTTACTACAGTTGGCTTCTAATGCTTCAGATAGAATACAGTTGGGTCTCACATAA					
3	A_23_P2112	9.99e-16	7.68e-20	-17.978.458	34.979.518	-3.05	A_23_P2112	NM_001112	ADARB1	GGAACCGATGGGCCATTAAACATGAACTGAACGGTTAAAAGCACAGTCTATGGAACGCTA					
4	A_23_P9336	9.99e-16	8.77e-20	-17.906.747	3.485.214	-5.65	A_23_P9336	NM_001136	AGER	CCCTTGAACTGTTCTGGCCTCAGACCAACTCTCTCCTGTATAATCTCTCTCCTGTATAAC					
5	A_33_P3317	2.24e-15	2.62e-19	-17.323.259	33.798.303	-2.36	A_33_P3317	NM_001495	GFRA2	GTGGGGGCTCCTGATCCGCATCCAAGCTTAACCAAGGCTCCAATAAACGTGCTAGGAAGC					
6	A_23_P4263	1.04e-14	1.97e-18	-1.628.866	31.850.378	-3.98	A_23_P4263	NM_003734	AOC3	AAGGCAAGCCACAGAAATGTGTATAGCGCACTTCCCATTTGTGTTTCAGAAAGGAGTAGA					
7	A_23_P3029	1.04e-14	2.25e-18	-16.222.511	31.722.245	-4.34	A_23_P3029	NM_001801	CDO1	AGGAACTTACGGCACTCTGCCATTTGTTAATGAAAGGAAGTGCAGAGGATTTAGAAAAGT					
8	A_23_P5339	1.04e-14	2.40e-18	-16.189.176	31.657.506	-3.32	A_23_P5339	NM_002837	PTPRB	ATGGTCCAGACTGAGTGTCAGTATGTCTACCTACATCAGTGTGTAAGAGATGTCCTCAGA					
9	A_23_P9133	1.04e-14	2.44e-18	-16.180.222	31.640.098	-3.25	A_23_P9133	NM_052970	HSPA12B	ATCTCTAATGTGGAGGTGGGAACATTATTGTGGTGGAGGCAATTATGAGGGTAGCATTTC					
10	A_23_P3444	1.05e-14	2.76e-18	-16.120.576	31.523.925	-3.38	A_23_P3444	NM_019055	ROBO4	AACCTCACCATGGAAAGAAAATAATTATGAATGCCACTGAGGCACTGAGGCCCTACCTCA					
11	A_33_P3376	1.15e-14	3.38e-18	-16.019.764	31.326.748	-3.39	A_33_P3376	NM_153714	C10orf67	GGAGTTATTCGCTGTGGGGTCTTCTCTCAGCTAAGAGCTGAACAGATATCAGGATGACCT					
12	A_23_P8486	1.25e-14	4.01e-18	-15.935.286	31.160.713	-5.18	A_23_P8486	NM_007177	FAM107A	TCTCCTTTGCAGCTTAGGATGGCTATGTGATCAGGTGTGGCCAATGAAATTGAAGAGGAA					
13	A_33_P3381	1.92e-14	7.06e-18	-15.657.744	30.610.016	-4.43	A_33_P3381	NM_019105	TNXB	AGTTCTCGGTGCCCTTCACGGAAATGAAGCTGAGACCAAGAAACTTTCGCTCCCCAGCGG					
14	A_33_P3209	1.92e-14	7.29e-18	-15.642.224	30.578.984	-2.98	A_33_P3209	NM_022648	TNS1	TTGAAACTCTCCTGTGTCTGTGTTTTGTTTGTGCGTGTGTGTGAGAGCACATCAGTGTGT					
15	A_23_P1443	2.45e-14	1.01e-17	-15.482.525	30.258.181	-3.15	A_23_P1443	NM_004787	SLIT2	AAGCAGCAGGGCTATGCTGCTTGCCAAACAACCAAGAAGGTGTCCCGATTAGAGTGCAGA					
16	A_33_P3378	2.45e-14	1.13e-17	-15.428.981	30.150.014	-3.10	A_33_P3378	NM_001083	PDE5A	CTGTCAGTGGTTTCCTCATTTTGCCATGCTTTGTAAAAATAAAAAGAATGATCAAGTAGG					
17	A_33_P3293	2.45e-14	1.17e-17	-15.414.878	30.121.473	-2.58	A_33_P3293	NM_170600	SH2D3C	CCATCTCACTGTAAATAAGTCTCTGTTCTGTAAATAGATGTACAGAAGCCATGTTATTTC					
18	A_23_P6766	2.45e-14	1.38e-17	-15.334.439	29.958.277	-2.55	A_23_P6766	NM_001864	COX7A1	CATCCCGTTGTACCTGAAGGGCGGCATCGTTGACAACATCCTGTACCGAGTGACAATGAC					
19	A_23_P4167	2.45e-14	1.40e-17	-15.329.226	29.947.677	-4.46	A_23_P4167	NM_016929	CLIC5	TGCCTGATTCAACTGTAGCCTGTAATGGGTAAAAGCCACATTTAGGAGGTGGTCTGATCA					
20	A_23_P2524	2.45e-14	1.40e-17	-15.326.698	29.942.536	-2.64	A_23_P2524	NM_000442	PECAM1	CGCCTGTGAAATACCAACCTGAAGACACGGTTCATTCAGGCAACGCACAAAACAGAAAAT					
21	A_33_P3248	2.45e-14	1.43e-17	-15.316.614	29.922.019	-4.98	A_33_P3248	NM_001199	INMT	TCTGTCACCAATGCTGCCAACAATGGGGTCTGCTTCATTGTGGCTCGCAAGAAGCCTGGG					

2. Practical exercises

1. Perform GEO2R analysis on the GSE33532 between tissue site A and normal lung tissue
2. Take top DEG and perform the following subanalyses:
   a. Are there differences in gene expression between TCGA LUSC and normal samples?
   b. Are there differences between highly expressed LUSC samples and low expressed samples in terms of survival
   c. Is these gene expressed in other tissues?
   d. What about its corresponding protein?
   e. Is it expressed in NK-cells?
3. Which are the most frequently regions lost in breast cancer?

Summary

1. Omics
2. Public sources
3. Functional analysis

3. Functional analysis

GOAL: Get global functional information from our data to help in the interpretation of the results beyond a list of genes

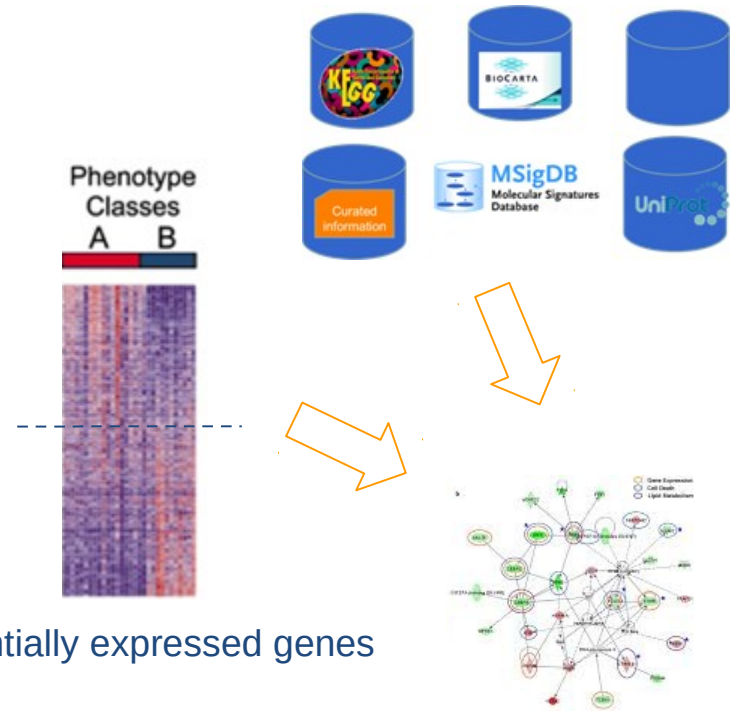
Several databases:

- GO
- Pathways (KEGG, BioCarta, etc..)
- mSigDB

Several statistical approaches

Several tools:

- DAVID
- GSEA
- PANTHER
- Cytoscape
- ...

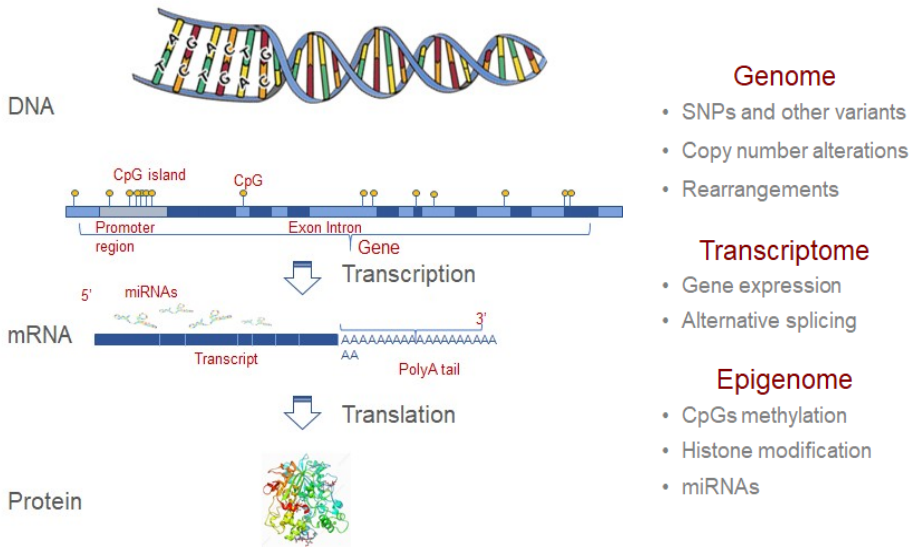


Differentially expressed genes

3. Functional analysis

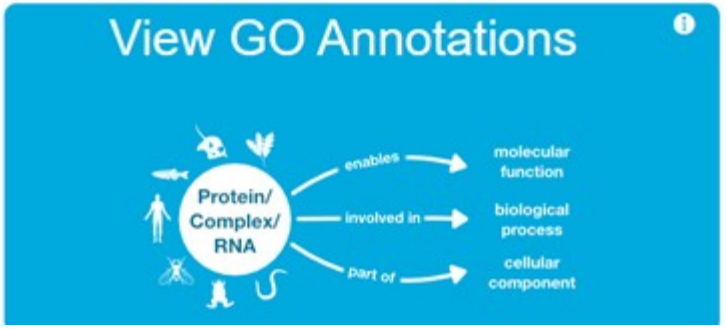
Omics functional approach

- Sequential analysis and functional exploitation of all results
  a. Convert all data into a common basic unit -> genes
  b. Apply any known functional methodology
- Perform functional analysis on the results obtained by an integration approach



3. Gene Ontology

Three ontologies



- Molecular Function

Gene functions

Ex: Transcription factor, membrane receptor

- Biological Process

Biological process in which gene is involved

Ex: mitosis, apoptosis

- Cellular Component

Cell location of gen depending on functions

Ex: nucleous, telomerus

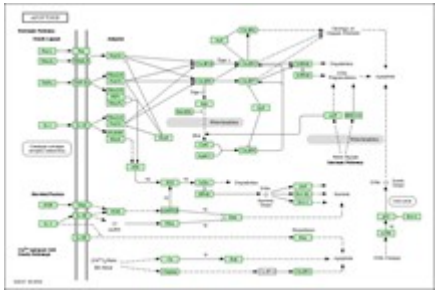
- **GO:0003673 : Gene_Ontology (65883)**
  - **GO:0008150 : biological_process (44405)**
    - GO:0007610 : behavior (357)
    - GO:0000004 : biological_process unknown (7877)
    - **GO:0009987 : cellular process (32672)**
      - GO:0007154 : cell communication (5384)
      - GO:0008219 : cell death (744)
      - GO:0030154 : cell differentiation (464)
      - GO:0008151 : cell growth and/or maintenance (28802)
      - GO:0006928 : cell motility (911)
      - GO:0006944 : membrane fusion (257)
    - GO:0016265 : death (793)
    - GO:0007275 : development (4615)
    - GO:0008371 : obsolete (1581)
    - GO:0007582 : physiological processes (31124)
    - GO:0016032 : viral life cycle (115)
  - GO:0005575 : cellular_component (32869)
  - GO:0003674 : molecular_function (53910)

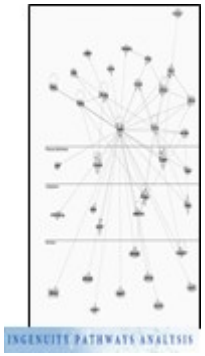
http://geneontology.org/

3. Pathways

Pathway is the term from molecular biology which depicts an artificial simplified model of a process within a cell or tissue.

- KEGG: http://www.genome.jp/kegg/pathway.html
- BioCarta
- Reactome: http://www.reactome.org/
- Wikipathways: http://www.wikipathways.org/index.php/WikiPathways
- ...



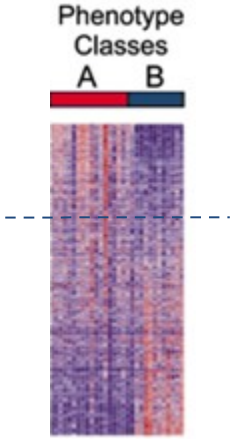


3. Molecular Signatures Database

https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

3. Statistical approaches

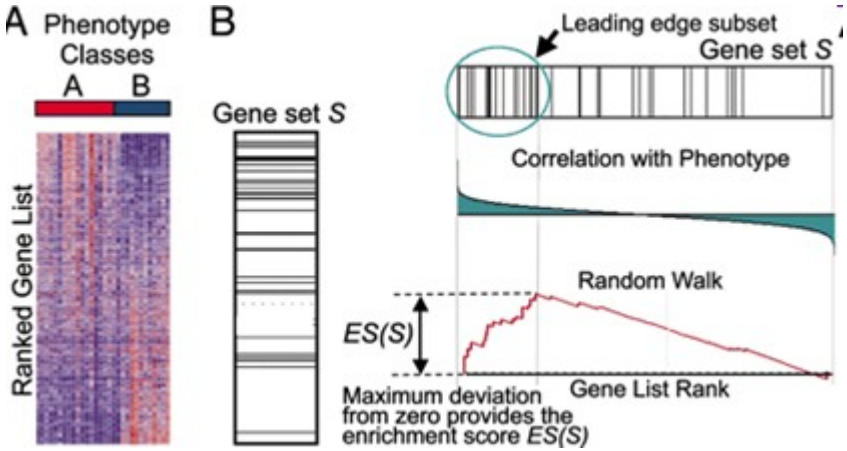
- Specific terms:
  - Gene set
  - (gene) Universe
- How to detect enrichment
  - Mere overlapping with a gene set
  - Statistical overrepresentation of a gene list in a gene set
    - Fisher's exact test (or hypergeometric)
      - List of differentially expressed genes (DEG)
      - Human genome (background/universe)
    - Enrichment analysis (the whole list)
    - Redundancy (some solutions in GSEA tool)



Differentially expressed genes

3. Statistical approaches

- GSEA: GeneSet Enrichment Analysis (A. Subramanian et al 2005)
  - Complete list of genes (ranked)
  a. Calculation of Enrichment Scores (NES)
  b. Estimation of Significance Level of ES
  c. Adjustment for Multiple Hypothesis Testing. FDR



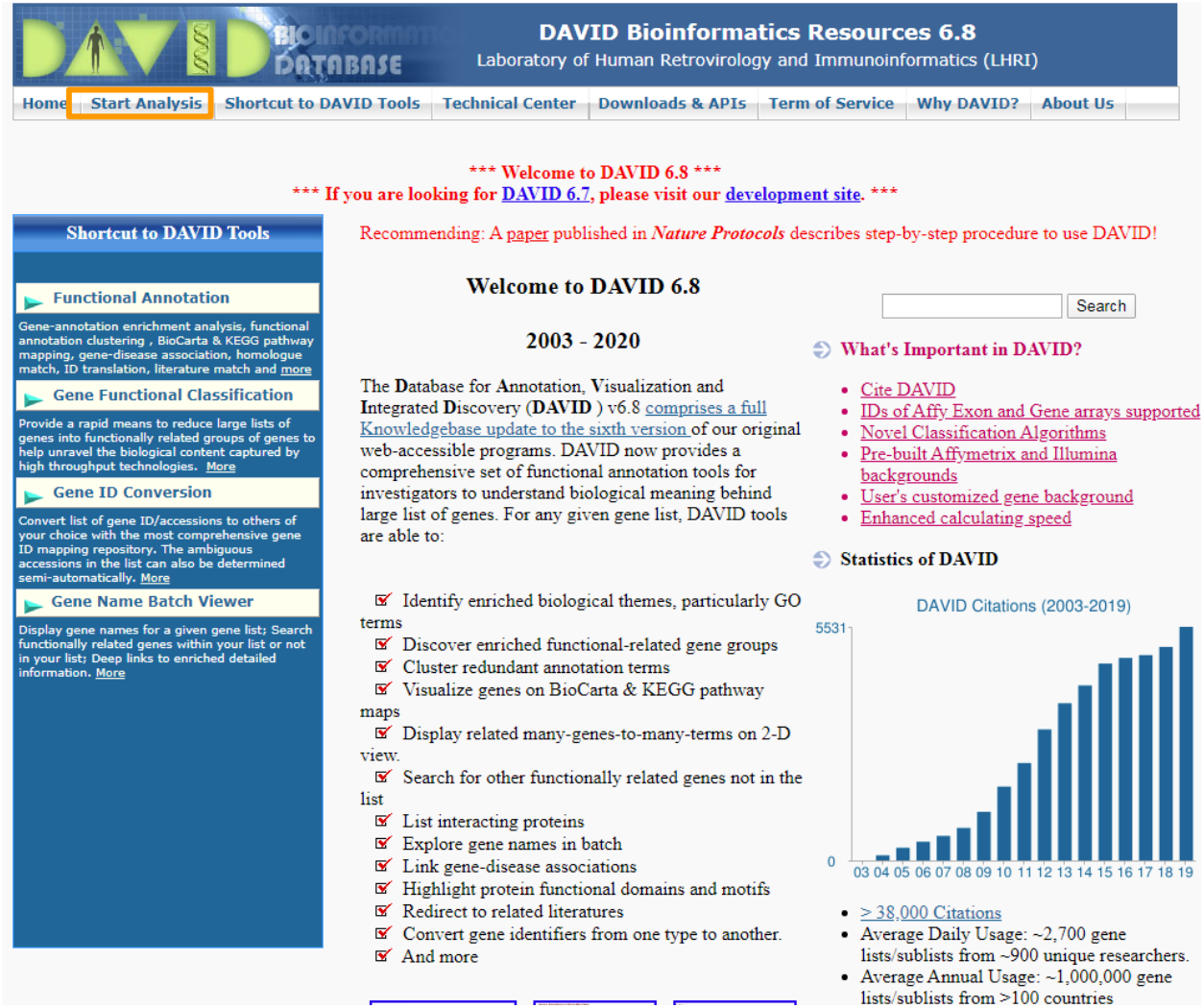
3. DAVID

The Database for Annotation, Visualization and Integrated Discovery (DAVID) provides a comprehensive set of functional annotation tools to understand biological meaning behind a list of genes

- ID conversion
- Functional annotation
- Functional knowledge categories:
  - Disease (1 selected)
  - Functional_Categories (3 selected)
  - Gene_Ontology (3 selected)
  - General_Annotations (0 selected)
  - Literature (0 selected)
  - Main_Accessions (0 selected)
  - Pathways (2 selected)
  - Protein_Domains (3 selected)
  - Protein_Interactions (0 selected)
  - Tissue_Expression (0 selected)

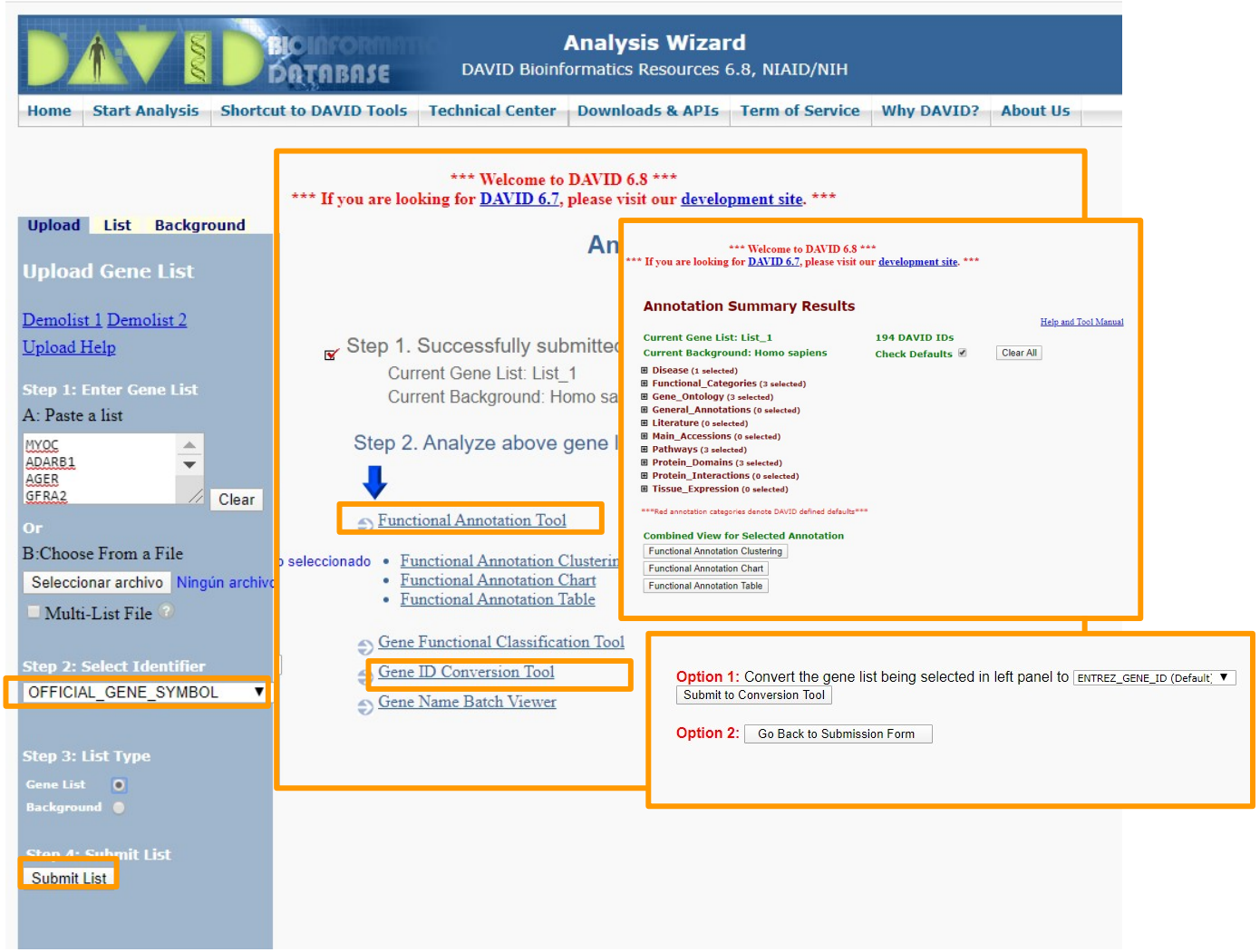
Statistics: EASE score (a modified Fisher's exact test)

3. DAVID



https://david.ncifcrf.gov/

3. DAVID


Analysis Wizard
DAVID Bioinformatics Resources 6.8, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***
Upload
List
Background
Upload Gene List
Demolist 1 Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
MYOC
ADARB1
AGER
GFRA2
Clear
Or
B:Choose From a File
Seleccionar archivo
Ningún archivo
Multi-List File
Step 2: Select Identifier
OFFICIAL_GENE_SYMBOL
Step 3: List Type
Gene List
Background
Submit List
Step 1. Successfully submitted
Current Gene List: List_1
Step 2. Analyze above gene
Functional Annotation Tool
Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table
Gene Functional Classification Tool
Gene ID Conversion Tool
Gene Name Batch Viewer
Annotation Summary Results
Help and Tool Manual
Current Gene List: List_1
Current Background: Homo sapiens
194 DAVID IDs
Check Defaults
Clear All
Disease (1 selected)
Functional_Categories (3 selected)
Gene_Ontology (3 selected)
General_Annotations (0 selected)
Literature (0 selected)
Main_Accessions (0 selected)
Pathways (3 selected)
Protein_Domains (3 selected)
Protein_Interactions (0 selected)
Tissue_Expression (0 selected)
Red annotation categories denote DAVID defined defaults
Combined View for Selected Annotation
Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table
Option 1: Convert the gene list being selected in left panel to
ENTREZ_GENE_ID (Default)
Submit to Conversion Tool
Option 2:
Go Back to Submission Form

3. Practical exercises

1. Take top 100 DEG obtained by GEO2R analysis on the GSE33532 between tissue site A and normal lung tissue and perform functional analysis using DAVID
   a. What is the top enriched KEGG pathway?
   b. Convert the gene ids to the Ensembl gene ids