

# Functional analysis of omics data

february 2020

**MARGenomics**



# Summary

---

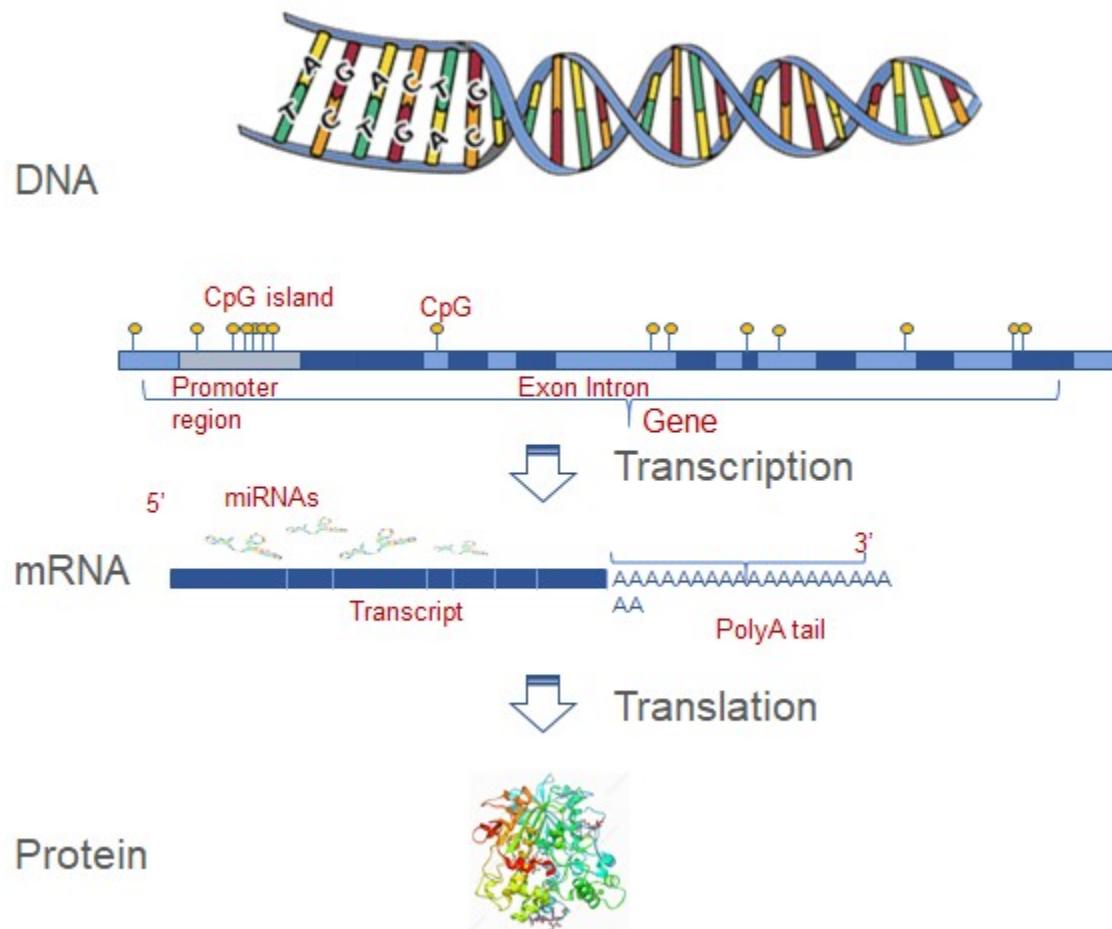
1. Omics
2. Public sources
3. Functional analysis

# 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  
...

### Databases

GO  
REACTOME  
KEGG  
BioCarta  
mSigDB  
...

### Web sources

GDAC iCGC  
cBioPortal  
GEPIA XENA  
GEO SRA  
dbGaP  
...

## 2. Public sources

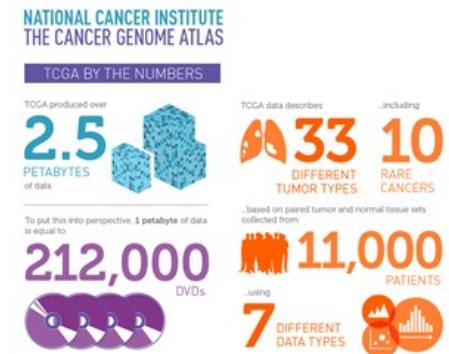
---

- 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

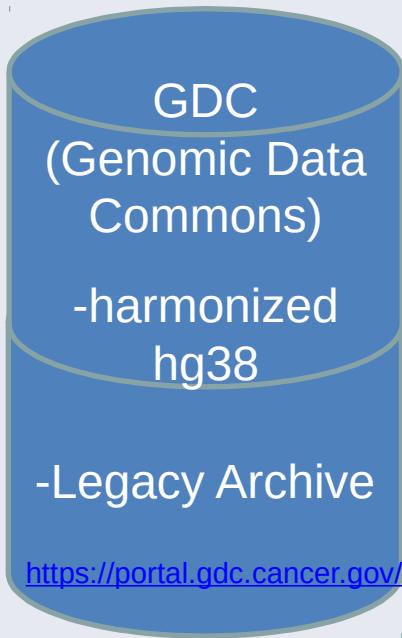


<http://cancergenome.nih.gov/>



## 2. TCGA

### Raw data



Some files can also be found in the sra database (ncbi)

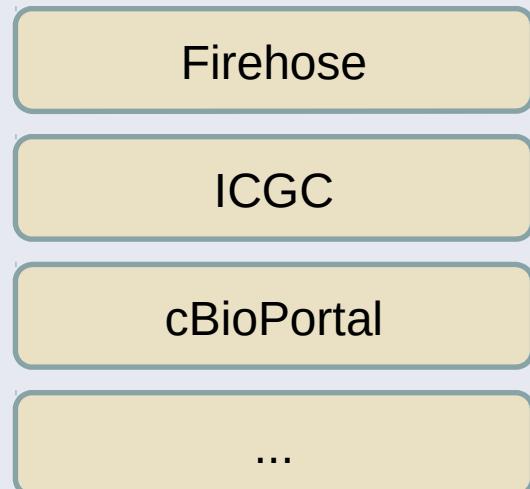
### Processed data



### Visualization



### Web access and analysis



## 2. TCGA Firehose

 FIREHOSE  
Broad GDAC

Dashboards Data Analyses Software Documentation FAQ Download Contact Us

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	<a href="#">Browse</a>	<a href="#">Browse</a>
Bladder urothelial carcinoma	BLCA	412	<a href="#">Browse</a>	<a href="#">Browse</a>
Breast invasive carcinoma	BRCA	1098	<a href="#">Browse</a>	<a href="#">Browse</a>
Cervical and endocervical cancers	CESC	307	<a href="#">Browse</a>	<a href="#">Browse</a>
Cholangiocarcinoma	CHOL	51	<a href="#">Browse</a>	<a href="#">Browse</a>
Colon adenocarcinoma	COAD	460	<a href="#">Browse</a>	<a href="#">Browse</a>
Colorectal adenocarcinoma	COADREAD	631	<a href="#">Browse</a>	<a href="#">Browse</a>
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	<a href="#">Browse</a>	<a href="#">Browse</a>
Esophageal carcinoma	ESCA	185	<a href="#">Browse</a>	<a href="#">Browse</a>
FFPE Pilot Phase II	FPPP	38	None	<a href="#">Browse</a>
Glioblastoma multiforme	GBM	613	<a href="#">Browse</a>	<a href="#">Browse</a>
Glioma	GBMLGG	1129	<a href="#">Browse</a>	<a href="#">Browse</a>
Head and Neck squamous cell carcinoma	HNSC	528	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney Chromophobe	KICH	113	<a href="#">Browse</a>	<a href="#">Browse</a>
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney renal clear cell carcinoma	KIRC	537	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney renal papillary cell carcinoma	KIRP	323	<a href="#">Browse</a>	<a href="#">Browse</a>
Acute Myeloid Leukemia	LAML	200	<a href="#">Browse</a>	<a href="#">Browse</a>
Brain Lower Grade Glioma	LGG	516	<a href="#">Browse</a>	<a href="#">Browse</a>
Liver hepatocellular carcinoma	LIHC	372	<a href="#">Browse</a>	<a href="#">Browse</a>
Lung adenocarcinoma	LIAD	585	<a href="#">Browse</a>	<a href="#">Browse</a>
Lung squamous cell carcinoma	LUSC	504	<a href="#">Browse</a>	<a href="#">Browse</a>
Mesothelioma	MESO	9	<a href="#">Browse</a>	<a href="#">Browse</a>
Ovarian serous cystadenocarcinoma	OV	602	<a href="#">Browse</a>	<a href="#">Browse</a>
Pancreatic adenocarcinoma	PAAD	185	<a href="#">Browse</a>	<a href="#">Browse</a>

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

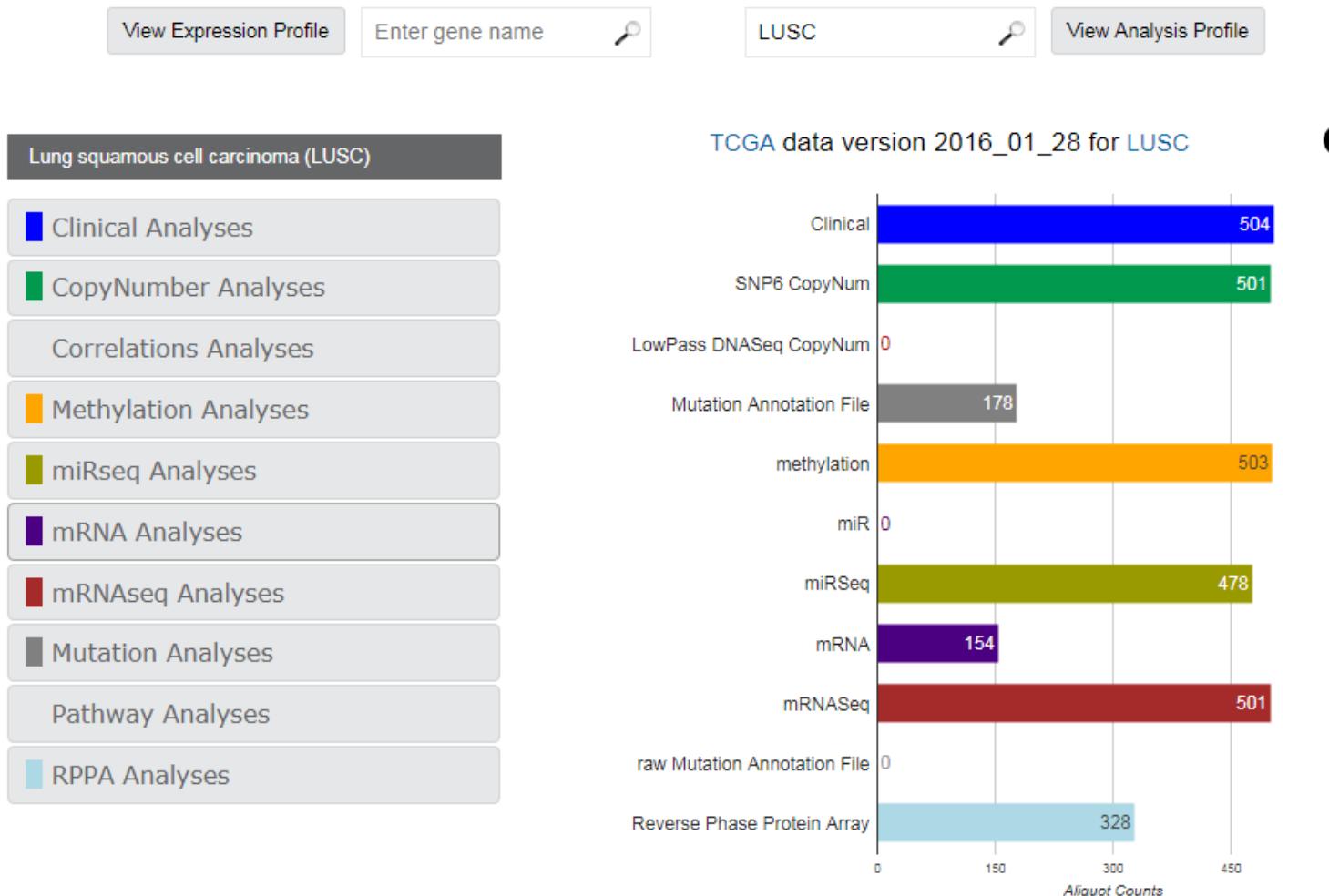
[Browse Samples](#) [Browse Workflow Graph](#) [Browse Analyses](#)

Thyroid carcinoma	THCA	503	<a href="#">Browse</a>	<a href="#">Browse</a>
Thymoma	THYM	124	<a href="#">Browse</a>	<a href="#">Browse</a>
Uterine Corpus Endometrial Carcinoma	UCEC	560	<a href="#">Browse</a>	<a href="#">Browse</a>
Uterine Carcinosarcoma	UCS	57	<a href="#">Browse</a>	<a href="#">Browse</a>
Uveal Melanoma	UVM	80	<a href="#">Browse</a>	<a href="#">Browse</a>

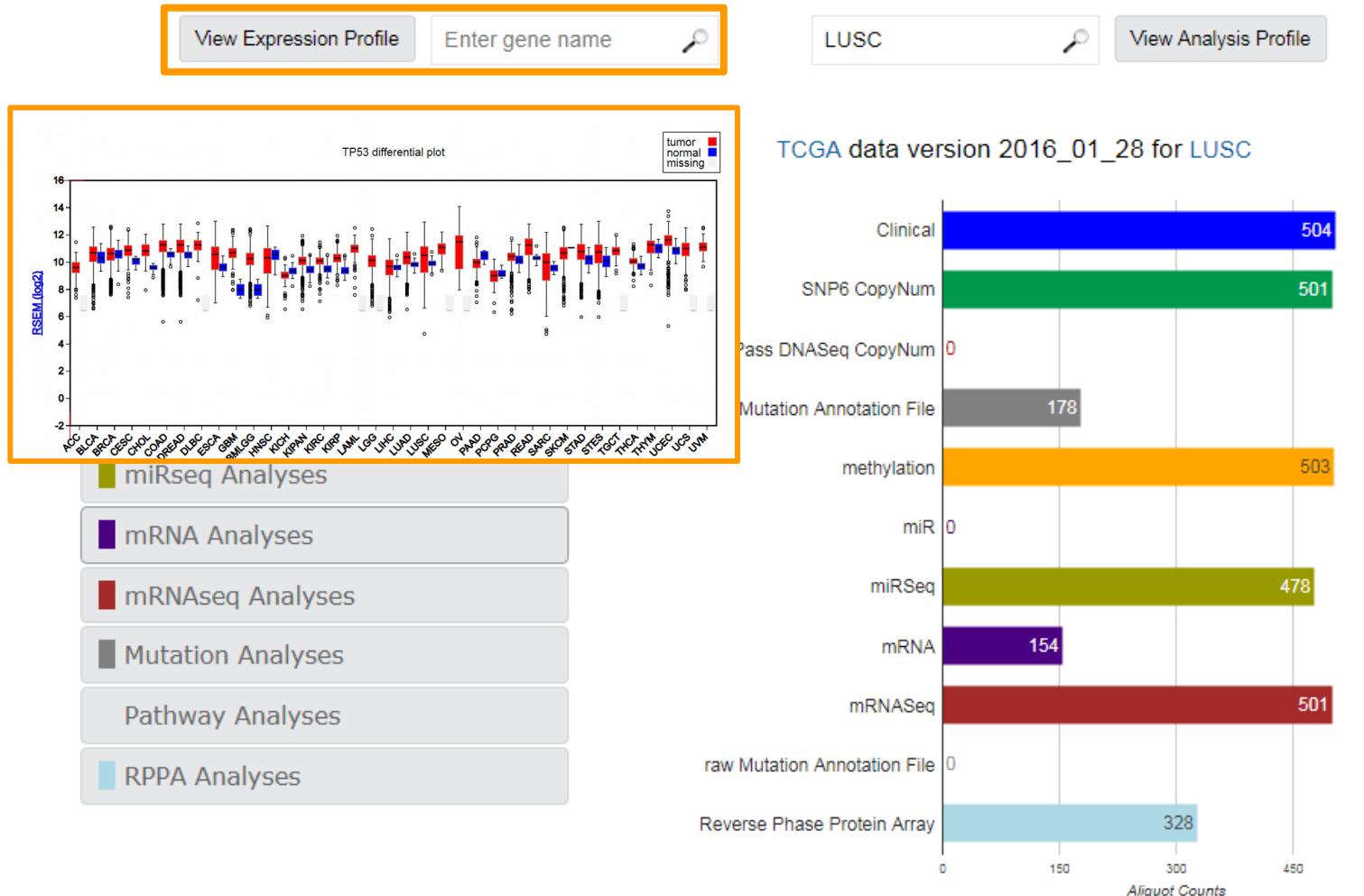
<https://gdac.broadinstitute.org/>



## 2. TCGA Firehose. LUSC



## 2. TCGA Firehose. LUSC



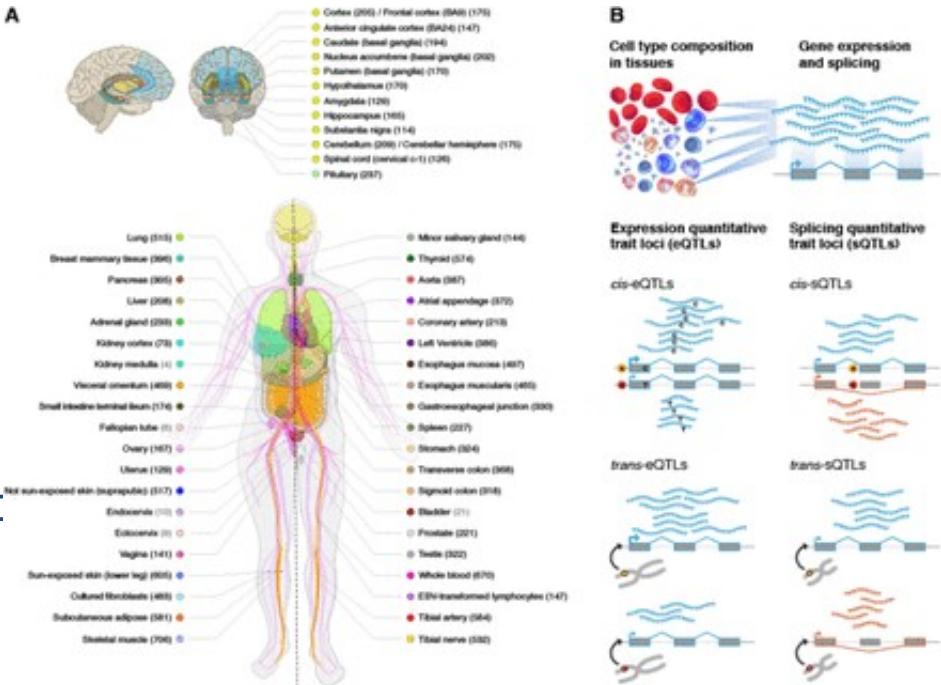
## 2. TCGA Firehose. LUSC



## 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.ncbi.nlm.nih.gov/pmc/articles/PMC787903/>
- [v1.full](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC787903/v1.full)
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC787903/>



# 2. GTEx

The screenshot shows the GTEx Portal homepage. At the top, there's a banner for 'GTEx Portal V8 Release' dated 2019-08-26, highlighting 17,382 RNA-Seq samples from 948 donors. Below the banner, the main navigation bar includes links for Home, Datasets, Expression, QTLs & Browsers, Sample Data, Documentation, About GTEx, Publications, Access Biospecimens, FAQs, and Contact. A search bar and a sign-in button are also present.

**Resource Overview**

- Current Release (V8)
  - Tissue & Sample Statistics
  - Tissue Sampling Info (Anatomogram)
  - Access & Download Data
  - Release History
  - How to cite GTEx?
- News and Events
  - 2020-01-06 New Functional Annotation Track in Locus Browser
  - 2020-01-06 RNA-seq Coverage Tracks
  - 2019-08-25 GTEx Portal V8 Release
  - 2017-10-18 ASHG GTEx Workshop Materials
- Documentation
  - Publication Policy
  - Consortium
  - Analysis Methods

**Explore GTEx**

**Browse**

- By gene ID
- By variant or rs ID
- By Tissue
- Histology Image Viewer

Browse and search all data by gene  
Browse and search all data by variant  
Browse and search all data by tissue  
Browse and search GTEx histology images

**Expression**

- Multi-Gene Query
- Top 50 Expressed Genes
- Transcript Browser

Browse and search expression by gene and tissue  
Visualize the top 50 expressed genes in each tissue  
Visualize transcript expression and isoform structures

**QTL**

- Locus Browser
- IGV Browser
- eQTL Dashboard
- eQTL Calculator

Visualize QTLs by gene in the Locus Browser  
Visualize tissue-specific eQTLs and coverage data in the IGV Browser  
Batch query eQTLs by gene and tissue  
Test your own eQTLs

**eGTEx**

- H3K27ac

Browse H3K27ac ChIP-seq data in IGV Browser

<https://gtexportal.org/>

[www.imim.es](http://www.imim.es)

## 2. GTEx

Explore GTEx

**Browse**

- By gene ID
- By variant or rs ID
- By Tissue
- Histology Image Viewer

**Expression**

- Multi-Gene Query
- Top 50 Expressed Genes
- Transcript Browser

**QTL**

- Locus Browser
- IGV Browser
- eQTL Dashboard
- eQTL Calculator

**eGTEx**

- H3K27ac

Browse and search all data by gene

Browse and search all data by variant

Browse and search all data by tissue

Browse and search GTEx histology images

Browse and search expression by gene and tissue

Visualize the top 50 expressed genes in each tissue

Visualize transcript expression and isoform structures

Visualize QTLs by gene in the Locus Browser

Visualize tissue-specific eQTLs and coverage data in the IGV Browser

Batch query eQTLs by gene and tissue

Test your own eQTLs

Browse H3K27ac ChIP-seq data in IGV Browser



<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

# THE HUMAN PROTEIN ATLAS



≡ MENU HELP NEWS

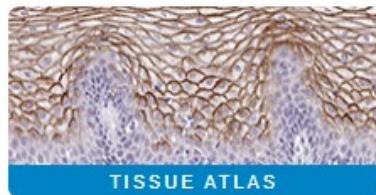
SEARCH<sup>i</sup>

lung cancer

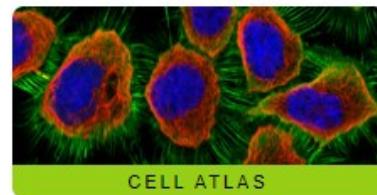
e.g. RBM3, insulin, CD36

Search

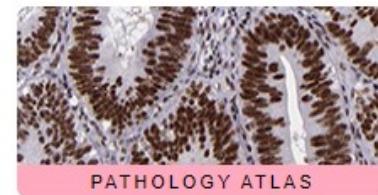
Fields »



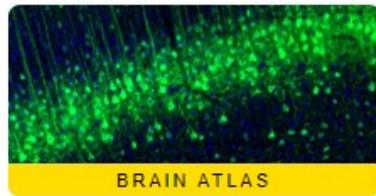
TISSUE ATLAS



CELL ATLAS



PATHOLOGY ATLAS



BRAIN ATLAS



BLOOD ATLAS



METABOLIC ATLAS

<https://www.proteinatlas.org/>

## 2. The Human Protein Atlas

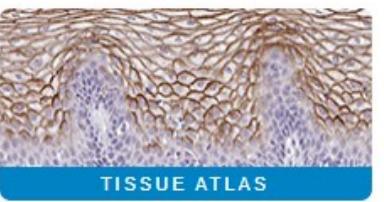
**THE HUMAN PROTEIN ATLAS** 

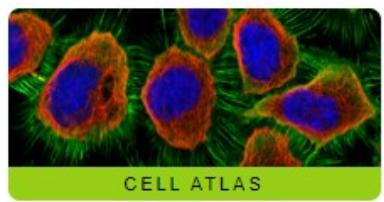
≡ MENU HELP NEWS

SEARCH<sup>i</sup>

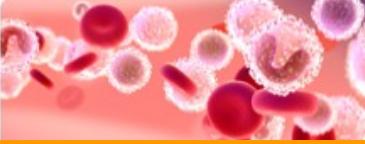
lung cancer  
e.g. RBM3, insulin, CD36

Search Fields »

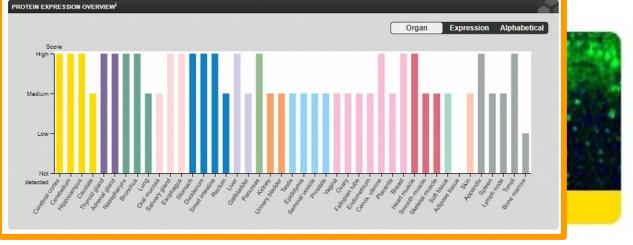
**TISSUE ATLAS** 

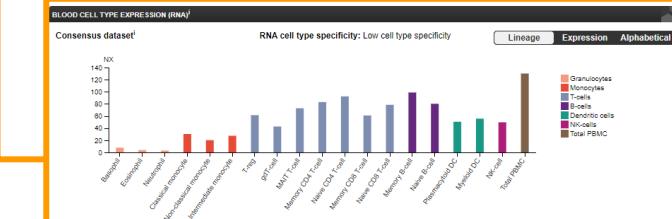
**CELL ATLAS** 

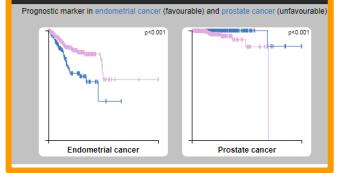
**PATHOLOGY ATLAS** 

**BLOOD ATLAS** 

**METABOLIC ATLAS** 

**PROTEIN EXPRESSION OVERVIEW** 

**BLOOD CELL TYPE EXPRESSION (RNA)<sup>j</sup>** 

Prognostic marker in endometrial cancer (favourable) and prostate cancer (unfavourable)  


[www.proteinatlas.org/](http://www.proteinatlas.org/)

[www.imim.es](http://www.imim.es)

## 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

The cBioPortal for Cancer Genomics provides visualization, analysis and download of large-scale cancer genomics data sets. Please cite Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012 when publishing results based on cBioPortal.

**QUERY**    DOWNLOAD DATA

**Select Studies:** 0 studies selected (0 samples) Search...

PanCancer Studies	2	<input type="checkbox"/> Select all listed studies (170)
Cell lines	2	<input type="checkbox"/> PanCancer Studies
Adrenal Gland	1	<input type="checkbox"/> MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017) 10945 samples   
Ampulla of Vater	1	<input type="checkbox"/> Pan-Lung Cancer (TCGA, Nat Genet 2016) 1144 samples   
Biliary Tract	5	<input type="checkbox"/> Cell lines
Bladder/Urinary Tract	8	<input type="checkbox"/> Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012) 1020 samples   
Blood	8	<input type="checkbox"/> NCI-60 Cell Lines (NCI, Cancer Res. 2012) 60 samples   
Bone	2	<input type="checkbox"/> Adrenal Gland
Bowel	6	<input type="checkbox"/> Adrenocortical Carcinoma
Breast	11	<input type="checkbox"/> Adrenocortical Carcinoma (TCGA, Provisional) 92 samples   
		<input type="checkbox"/> Ampulla of Vater
		<input type="checkbox"/> Ampullary Carcinoma
		<input type="checkbox"/> Ampullary Carcinoma (Baylor College of Medicine, Cell Reports 2016) 160 samples   

**Enter Gene Set:** User-defined List Advanced: Onco Query Language (OQL)  
Enter HUGO Gene Symbols or Gene Aliases

**What's New** @cbioportal

 cBioPortal  
Boston-area users: Join us for a live demonstration of cBioPortal basic & advanced features on Feb 23 at Dana-Farber. More details & RSVP: goo.gl/forms/zLPunjRv...

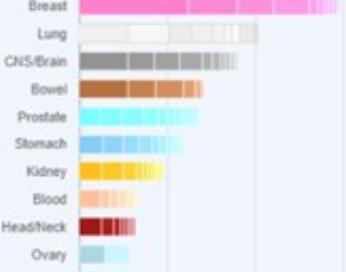
cBioPortal Instructional D... Thank you for your interest ... docs.google.com

Sign up for low-volume email news alerts

Subscribe

**Cancer Studies**  
The portal contains 170 cancer studies (details)

Cases by Top 20 Primary Sites



Primary Site	Number of Cases
Breast	10945
Lung	1144
CNS-Brain	1020
Bowel	60
Prostate	92
Stomach	160
Kidney	5
Blood	8
Head/Neck	11
Ovary	2

<https://www.cbioportal.org/>



## 2. cBioPortal. LUSC

Bone 2 ▲

Bowel 10

Breast 15

CNS/Brain 18

Cervix 2

Esophagus/Stomach 14

Eye 3

Head and Neck 13

Kidney 17

Liver 8

Lung 21

Lymphoid 19

Myeloid 9

Other 13

Ovary/Fallopian Tube 4

Pancreas 10

Peripheral Nervous System 5

Pleura 3 ▼

Lung

Thoracic PDX (MSK, Provisional) 139 samples ⓘ ⓘ ⓘ

Lung Neuroendocrine Tumor

→ SMALL CELL LUNG CANCER

Small Cell Lung Cancer (CLCGP, Nat Genet 2012) 29 samples ⓘ ⓘ ⓘ

Small Cell Lung Cancer (Johns Hopkins, Nat Genet 2012) 80 samples ⓘ ⓘ ⓘ

Small Cell Lung Cancer (U Cologne, Nature 2015) 110 samples ⓘ ⓘ ⓘ

Small-Cell Lung Cancer (Multi-Institute, Cancer Cell 2017) 20 samples ⓘ ⓘ ⓘ

Non-Small Cell Lung Cancer

Non-Small Cell Lung Cancer (MSK, Cancer Cell 2018) 75 samples ⓘ ⓘ ⓘ

Non-Small Cell Lung Cancer (MSKCC, J Clin Oncol 2018) 240 samples ⓘ ⓘ ⓘ

Non-Small Cell Lung Cancer (TRACERx, NEJM 2017) 327 samples ⓘ ⓘ ⓘ

Non-Small Cell Lung Cancer (University of Turin, Lung Cancer 2017) 41 samples ⓘ ⓘ ⓘ

Non-small cell lung cancer (MSK, Science 2015) 16 samples ⓘ ⓘ ⓘ

Pan-Lung Cancer (TCGA, Nat Genet 2016) 1144 samples ⓘ ⓘ ⓘ

→ LUNG ADENOCARCINOMA

Lung Adenocarcinoma (Broad, Cell 2012) 183 samples ⓘ ⓘ ⓘ

Lung Adenocarcinoma (MSKCC, Science 2015) 35 samples ⓘ ⓘ ⓘ

Lung Adenocarcinoma (TCGA, Firehose Legacy) 586 samples ⓘ ⓘ ⓘ

Lung Adenocarcinoma (TCGA, Nature 2014) 230 samples ⓘ ⓘ ⓘ

Lung Adenocarcinoma (TCGA, PanCancer Atlas) 566 samples ⓘ ⓘ ⓘ

Lung Adenocarcinoma (TSP, Nature 2008) 163 samples ⓘ ⓘ ⓘ

Non-Small Cell Cancer (MSKCC, Cancer Discov 2017) 915 samples ⓘ ⓘ ⓘ

→ LUNG SQUAMOUS CELL CARCINOMA

Lung Squamous Cell Carcinoma (TCGA, Firehose Legacy) 511 samples ⓘ ⓘ ⓘ

Lung Squamous Cell Carcinoma (TCGA, Nature 2012) 178 samples ⓘ ⓘ ⓘ

Lung Squamous Cell Carcinoma (TCGA, PanCancer Atlas) 487 samples ⓘ ⓘ ⓘ

0 studies selected (0 samples)

Query By Gene OR Explore Selected Studies



## 2. cBioPortal. LUSC

Selected Studies: [Modify](#)

Lung Squamous Cell Carcinoma (TCGA, Firehose Legacy) | (511 total samples)

Select Genomic Profiles:

- Mutations [?](#)
- Putative copy-number alterations from GISTIC [?](#)
- mRNA Expression. Select one of the profiles below:
  - mRNA Expression z-Scores (U133 microarray only) [?](#)
  - mRNA Expression z-Scores (microarray) [?](#)
  - mRNA Expression z-Scores (RNA Seq V2 RSEM) [?](#)

Enter a z-score threshold ±

Protein expression Z-scores (RPPA) [?](#)

Select Patient/Case Set:  
To build your own case set,  
try out our enhanced Study View.

Samples with mRNA data (RNA Seq V2) (501) [X](#) [▼](#)

Enter Genes:

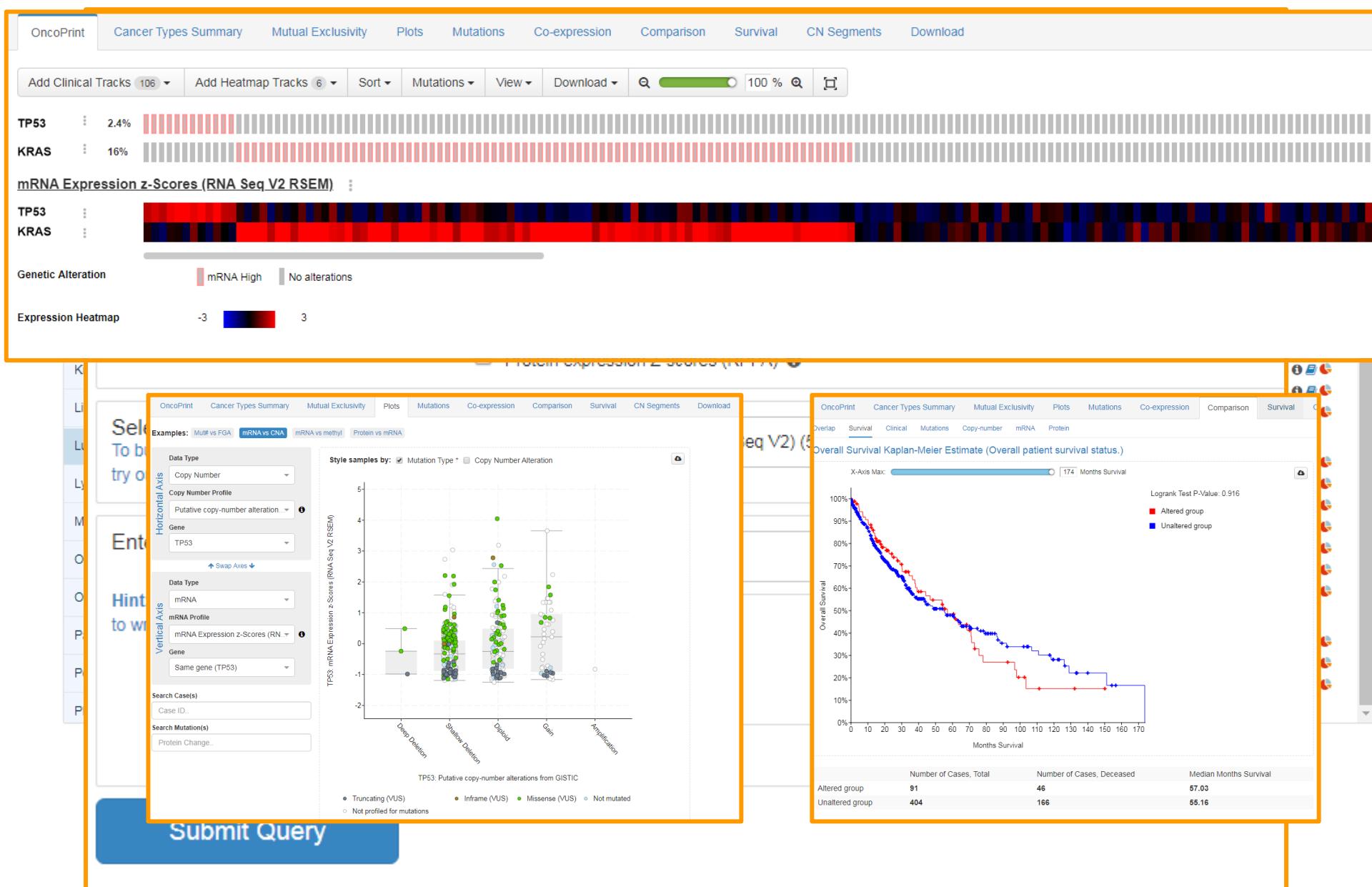
User-defined List [X](#) [▼](#)

TP53  
KRAS

All gene symbols are valid.

[Submit Query](#)

## 2. cBioPortal. LUSC



## 2. cBioPortal. LUSC

Selected Studies: [Modify](#) Lung Squamous Cell Carcinoma (TCGA, Firehose Legacy) | (511 total samples)

Select Genomic Profiles:

- Mutations [?](#)
- Putative copy-number alterations from GISTIC [?](#)
- mRNA Expression. Select one of the profiles below:
  - mRNA Expression z-Scores (U133 microarray only) [?](#)
  - mRNA Expression z-Scores (microarray) [?](#)
  - mRNA Expression z-Scores (RNA Seq V2 RSEM) [?](#)
- Protein expression Z-scores (RPPA) [?](#)

Select Patient/Case Set:  
To build your own case set,  
try out our enhanced Study View.

Samples with mutation and CNA data (178)

Enter Genes:

Hint: Learn Onco Query Language (OQL)  
to write more powerful queries [?](#)

User-defined List

TP53  
KRAS

All gene symbols are valid.

[Submit Query](#)

## 2. cBioPortal. LUSC

The screenshot displays the cBioPortal interface for Lung Squamous Cell Carcinoma (LUSC). The top navigation bar includes links for OncoPrint, Cancer Types Summary, Mutual Exclusivity, Plots, Mutations, Co-expression, Comparison, Survival, CN Segments, and Download.

Key features shown include:

- Genetic Alteration Analysis:** A horizontal bar shows mutation types for TP53 (81% mutations) and KRAS (6% mutations). The legend indicates: Inframe Mutation (putative driver) (dark blue), Missense Mutation (putative driver) (green), Missense Mutation (unknown significance) (light green), Truncating Mutation (putative driver) (black), Amplification (red), and Deep Deletion (blue).
- mRNA Expression z-Scores (RNA Seq V2 RSEM):** A plot comparing mRNA expression levels between TP53 and KRAS across various samples.
- Mutation Analysis for TP53:** A detailed view of TP53 mutations, including a scatter plot of mutation frequency (0-8) against genomic position (0-390aa), highlighting R158L/A162G and more. It also shows TP53 domain annotations (P53\_TAD, P53, P53\_tetramer) and mutation statistics (100 Missense, 40 Truncating, 3 Inframe, 0 Other).
- Sample ID and Protein Change Table:** A table listing 152 mutations (page 1 of 7) with columns for Sample ID, Protein Change, Annotation, Mutation Type, Copy #, and COSMIC Allele Freq (T).
- Copy Number Analysis (CNA):** An IGV-based visualization showing copy number changes across the TP53 gene region (hg19, chr17, chr17:7,570,720-7,591,868).
- Submit Query:** A large blue button at the bottom left.

## 2. cBioPortal. LUSC

Bone      2      Lung

Bow ...  
Brea ...  
Bre ...  
CNS ...  
Cerv ...  
Esop ...  
Eye ...  
Head ...  
Kidn ...  
Liver ...  
Lung ...  
Lym ...  
Myel ...  
Othe ...  
Ovar ...  
Pan ...  
Perip ...  
Pleu ...

Oncoprint   Cancer Types Summary   Mutual Exclusivity   Plots   Mutations   Co-expression   Comparison   Survival   CN Segments   Download

samples

Downloadable Data Files

Copy-number Alterations (OQL is not in effect)   Tab Delimited Format | Transposed Matrix  
Mutations (OQL is not in effect)   Tab Delimited Format | Transposed Matrix  
Altered samples: List of samples with alterations   Copy | Download | Query | Virtual Study  
Unaltered samples: List of samples without any alteration   Copy | Download | Query | Virtual Study  
Sample matrix: List of all samples where 1=altered and 0=unaltered   Copy | Download  
Relative linear copy-number values   Tab Delimited Format | Transposed Matrix  
mRNA expression (U133 microarray only)   Tab Delimited Format | Transposed Matrix  
mRNA Expression z-Scores (U133 microarray only)   Tab Delimited Format | Transposed Matrix  
mRNA expression (microarray)   Tab Delimited Format | Transposed Matrix  
mRNA Expression z-Scores (microarray)   Tab Delimited Format | Transposed Matrix  
mRNA expression (RNA Seq V2 RSEM)   Tab Delimited Format | Transposed Matrix  
mRNA Expression z-Scores (RNA Seq V2 RSEM)   Tab Delimited Format | Transposed Matrix  
Methylation (HM27)   Tab Delimited Format | Transposed Matrix  
Methylation (HM450)   Tab Delimited Format | Transposed Matrix  
Protein expression (RPPA)   Tab Delimited Format | Transposed Matrix  
Protein expression Z-scores (RPPA)   Tab Delimited Format | Transposed Matrix

Gene Alteration Frequency

Gene Symbol	Num Samples Altered	Percent Samples Altered ▾
TP53	145	81.5%
KRAS	10	5.6%

Showing 1-2 of 2



## 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  
doi: 10.1093/nar/gkx247

Published online 12 April 2017

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

Zefang Tang<sup>1,†</sup>, Chenwei Li<sup>1,2,†</sup>, Boxi Kang<sup>1</sup>, Ge Gao<sup>3</sup>, Cheng Li<sup>2,3</sup> and Zemin Zhang<sup>1,2,4,5,\*</sup>



## 2. GEPIA

GEPIA

GoPIA Example API Help About GEPIA2 (test)

 **GEPIA**  
Gene Expression Profiling Interactive Analysis

Single Gene Analysis    Cancer Type Analysis    Multiple Gene Analysis

Enter gene name:

The indicators in search box are "symbol" or "alias (newest symbol)".

e.g. ERBB2/ENSG00000141736/2064

GoPIA!

Profile Boxplots Stage Plots Survival Analysis Similar

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



## 2. GEPIA

GEPIA

GoPIA Example API Help About GEPIA2 (t)

Click here to get the extension of tumor abbreviations.

General Differential Genes Expression DIY Survival Similar Genes Correlation PCA

The Differentially Expressed Genes On Chromosomes

Overall Survival

BRCA (n=1058, n=291) LUSC (n=486, n=338)

Description: erb-b2 receptor tyrosine kinase 1

Alias: CD340, HER-2, HER-2/neu, HER2, MLN 19, NEU, NGL, TKR1

Summary: This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases. This protein has no ligand binding domain of its own and therefore cannot bind growth factors. However, it does bind tightly to other ligand-bound EGF receptor family members to form a heterodimer, stabilizing ligand binding and enhancing kinase-mediated activation of downstream signalling pathways, such as those involving mitogen-activated protein kinase and phosphatidylinositol-3 kinase. Allelic variations at amino acid positions 654 and 655 of isoform a (positions 624 and 625 of isoform b) have been reported, with the most common allele, Ile654/Ile655, shown here. Amplification and/or overexpression of this gene has been reported in numerous cancers, including breast and ovarian tumors. Alternative splicing results in several additional transcript variants, some encoding different isoforms and others that have not been fully characterized. [provided by RefSeq, Jul 2008]

Lookup this gene in:

GeneCard NCBI Ensembl EBI OMIM COSMIC HPA DrugBank

Interactive Bodymap

The median expression of tumor and normal samples in bodymap

Log<sub>2</sub>(TPM + 1) Scale

FI20\_MEMORIA.pdf tp53.tsv

<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



The UCSC Xena Platform for cancer genomics  
data visualization and interpretation

Mary Goldman<sup>1</sup>, Brian Craft<sup>1</sup>, Akhil Kamath<sup>2</sup>, Angela Brooks<sup>1</sup>, Jing Zhu<sup>1</sup>, and David Haussler<sup>1</sup>

<sup>1</sup>UCSC Genomics Institute, UC Santa Cruz

<sup>2</sup>BITS Pilani KK Birla Goa Campus, Pilani, India



## 2. UCSC Xena



## 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

The screenshot shows the NCBI GEO DataSets homepage. At the top, there's a navigation bar with links for "NCBI", "Resources", "How To", "Sign in to NCBI", and "Help". Below the navigation is a search bar with dropdown menus for "GEO DataSets" and "Advanced". The main content area features a large banner with a collage of gene expression terms like "expression", "cells", "genes", "microarray", etc., followed by a detailed description of the database. Below the banner, there are three columns: "Getting Started" (with links to Documentation, FAQ, About, Query, and Options), "GEO Tools" (with links to Submit, Advanced Search, Browser, Programmatic Access, and GEO2R), and "More Resources" (with links to Home, Profiles, Epigenomics, and SRA). At the bottom, there's a section for "Example Searches".



## 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

The screenshot shows the GEO2R web application interface. At the top, there are links for NCBI, GEO Publications, FAQ, HOME, and Email GEO. A user account link for 'laranonell' is also present. The main header reads '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.' Below this, there's a 'Full Instructions' link and a 'Top 250' button. A 'Set' button is located next to a 'GEO accession' input field. Below the input field, there are tabs for 'GEO2R', 'Value distribution', 'Options', 'Profile graph', and 'R script'. A 'Quick start' section provides instructions for using the tool, mentioning steps like specifying a GEO Series accession and platform, defining groups, assigning samples, and clicking 'Top 250'. A 'How to use' section is also present. At the bottom, there are buttons for 'Top 250' and 'Save all results'.



## 2. GEO2R. Data

---

# SCIENTIFIC REPORTS



OPEN

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

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

Received: 31 May 2018

Accepted: 12 October 2018

Published online: 26 October 2018

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

NCBI

GEO Gene Expression Omnibus

HOME SEARCH SITE MAP NCBI > GEO > Accession Display GSE74706 Contact: laranell | My submissions | Sign Out

Scope: Self Format: HTML Amount: Quick GEO accession: GSE74706

**Series GSE74706** Query DataSets for GSE74706

Status Public on May 18, 2016  
Title Transcriptome of human NSCLC tissues  
Organism Homo sapiens  
Experiment type Expression profiling by array  
Summary RNA from patient samples was isolated to examine the TGF $\beta$  pathway expression between matching pairs of tumor-free lung and NSCLC specimen  
Overall design RNA from patient samples was isolated to examine the TGF $\beta$  pathway expression between matching pairs of tumor-free lung and NSCLC specimen  
Contributor(s) Marwitz S, Ammerpohl O, Reck M, Klingmueller U, Goldmann T  
Citation(s) Marwitz S, Depner S, Dvornikov D, Merkle R et al. Downregulation of the TGF $\beta$  Pseudoreceptor BAMBI in Non-Small Cell Lung Cancer Enhances TGF $\beta$  Signaling and Invasion. *Cancer Res* 2015 Jul 1;76(13):3785-801. PMID: 27197161  
Submission date Nov 05, 2015  
Last update date Jan 09, 2018  
Contact name Sebastian Marwitz  
Organization name Research Center Borstel - Leibniz Lung Center  
Department Pathology  
Street address Parkallee 3a  
City Borstel  
ZIP/Postal code 23845  
Country Germany

Platforms (1) GPL13497 Agilent-026652 Whole Human Genome Microarray 4x44K v2 (Probe Name version)  
Samples (36) GSM1930516\_17962\_08\_Lung  
at More... GSM1930517\_17962\_08\_Tumor  
GSM1930518\_21577\_08\_Lung

**Relations**  
BioProject PRJNA301214

Analyze with GEO2R

Download family Format  
SOFT formatted family file(s) SOFT  
MINIML formatted family file(s) MINIML  
Series Matrix File(s) TXT

[GSE74706](#)

Supplementary file	Size	Download	File type/resource
GSE74706_RAW.tar	226.0 Mb	(http://custom)	TAR (of TXT)

Raw data provided as supplementary file  
Processed data included within Sample table

# 2. GEO2R. Analysis

GEO accession | GSE74706 | Set | Transcriptome of human NSCLC tissues

**Samples**

Group	Accession	Title	Source name	Individual id	Tissue	Tissue type
Lung	GSM1	17982_08 Lung	Lung	17982	Tumor-free lung	
Tumor	GSM1	17982_08 Tumor	Lung	17982	NSCLC	Adenocarcinoma
Lung	GSM1	21577_08 Lung	Lung	21577	Tumor-free lung	
Tumor	GSM1	21577_08 Tumor	Lung	21577	NSCLC	Adenocarcinoma
Lung	GSM1930620	6495_08 Lung	Lung	6495	Tumor-free lung	
Tumor	GSM1930621	6495_08 Tumor	Lung	6495	NSCLC	Squamous Cell Carcinoma
Lung	GSM1930622	12097_07 Lung	Lung	12097	Tumor-free lung	
Tumor	GSM1930623	12097_07 Tumor	Lung	12097	NSCLC	Squamous Cell Carcinoma
Lung	GSM1930624	15982_07 Lung	Lung	15982	Tumor-free lung	
Tumor	GSM1930625	15982_07 Tumor	Lung	15982	NSCLC	Squamous Cell Carcinoma
Lung	GSM1930626	12799_12 Lung	Lung	12799	Tumor-free lung	
Tumor	GSM1930627	12799_12 Tumor	Lung	12799	NSCLC	Squamous Cell Carcinoma
Lung	GSM1930628	1888_12 Lung	Lung	1888	Tumor-free lung	
Tumor	GSM1930629	1888_12 Tumor	Lung	1888	NSCLC	Squamous Cell Carcinoma
Lung	GSM1930630	11580_12 Lung	Lung	11580	Tumor-free lung	
Tumor	GSM1930631	11580_12 Tumor	Lung	11580	NSCLC	Adenocarcinoma
Lung	GSM1930632	3302_12 Lung	Lung	3302	Tumor-free lung	
Tumor	GSM1930633	3302_12 Tumor	Lung	3302	NSCLC	Adenocarcinoma
Lung	GSM1930634	12893_12 Lung	Lung	12893	Tumor-free lung	
Tumor	GSM1930635	12893_12 Tumor	Lung	12893	NSCLC	Adenocarcinoma

**GEO2R**

**Value distribution** (highlighted)

Options | Profile graph | R script

**GEO2R start**

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare.
- Assign Samples to each group. Highlight Sample rows then click the group name.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The top 250 genes are shown.
- You may change settings in Options tab.

**How to use:**

Top 250 | Save all results (highlighted)

**GEO2R**

Value distribution | Options | Profile graph | R script

Calculate the distribution of value data for the Samples you have selected. Distributions may be viewed graphically as a box plot or exported as a number summary table. The plot is useful for determining if value data are median-centered across Samples, and thus suitable for cross-comparison. More...

**GSE74706/GPL13497, selected samples**

Legend: □ Lung □ Tumor

Sample identifiers listed on x-axis: GSM1930616, GSM1930618, GSM1930620, GSM1930622, GSM1930624, GSM1930626, GSM1930628, GSM1930630, GSM1930632, GSM1930634, GSM1930636, GSM1930638, GSM1930640, GSM1930642, GSM1930644, GSM1930646, GSM1930648, GSM1930650, GSM1930652, GSM1930654, GSM1930656, GSM1930658, GSM1930660, GSM1930662, GSM1930664, GSM1930666, GSM1930668, GSM1930670, GSM1930672, GSM1930674, GSM1930676, GSM1930678, GSM1930680, GSM1930682, GSM1930684, GSM1930686, GSM1930688, GSM1930690, GSM1930692, GSM1930694, GSM1930696, GSM1930698, GSM1930700, GSM1930702, GSM1930704, GSM1930706, GSM1930708, GSM1930710, GSM1930712, GSM1930714, GSM1930716, GSM1930718, GSM1930720, GSM1930722, GSM1930724, GSM1930726, GSM1930728, GSM1930730, GSM1930732, GSM1930734, GSM1930736, GSM1930738, GSM1930740, GSM1930742, GSM1930744, GSM1930746, GSM1930748, GSM1930750, GSM1930752, GSM1930754, GSM1930756, GSM1930758, GSM1930760, GSM1930762, GSM1930764, GSM1930766, GSM1930768, GSM1930770, GSM1930772, GSM1930774, GSM1930776, GSM1930778, GSM1930780, GSM1930782, GSM1930784, GSM1930786, GSM1930788, GSM1930790, GSM1930792, GSM1930794, GSM1930796, GSM1930798, GSM1930800, GSM1930802, GSM1930804, GSM1930806, GSM1930808, GSM1930810, GSM1930812, GSM1930814, GSM1930816, GSM1930818, GSM1930820, GSM1930822, GSM1930824, GSM1930826, GSM1930828, GSM1930830, GSM1930832, GSM1930834, GSM1930836, GSM1930838, GSM1930840, GSM1930842, GSM1930844, GSM1930846, GSM1930848, GSM1930850, GSM1930852, GSM1930854, GSM1930856, GSM1930858, GSM1930860, GSM1930862, GSM1930864, GSM1930866, GSM1930868, GSM1930870, GSM1930872, GSM1930874, GSM1930876, GSM1930878, GSM1930880, GSM1930882, GSM1930884, GSM1930886, GSM1930888, GSM1930890, GSM1930892, GSM1930894, GSM1930896, GSM1930898, GSM1930900, GSM1930902, GSM1930904, GSM1930906, GSM1930908, GSM1930910, GSM1930912, GSM1930914, GSM1930916, GSM1930918, GSM1930920, GSM1930922, GSM1930924, GSM1930926, GSM1930928, GSM1930930, GSM1930932, GSM1930934, GSM1930936, GSM1930938, GSM1930940, GSM1930942, GSM1930944, GSM1930946, GSM1930948, GSM1930950, GSM1930952, GSM1930954, GSM1930956, GSM1930958, GSM1930960, GSM1930962, GSM1930964, GSM1930966, GSM1930968, GSM1930970, GSM1930972, GSM1930974, GSM1930976, GSM1930978, GSM1930980, GSM1930982, GSM1930984, GSM1930986, GSM1930988, GSM1930990, GSM1930992, GSM1930994, GSM1930996, GSM1930998, GSM1930999, GSM1931000.



# 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](#) [YouTube](#)

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-18	4.81e-21	-19.5	37.6	-6.63	A_23_P23783	NM_000261	4653	MYOC	myocilin, trabecular...	ATGCATTACTACAG...
A_23_P211207	9.99e-18	7.68e-20	-18	35	-3.05	A_23_P211207	NM_001112	104	ADARB1	adenosine deaminas...	GGAACCGATGGGC...
A_23_P93360	9.09e-18	8.77e-20	-17.9	34.9	-5.65	A_23_P93360	NM_001138	177	AGER	advanced glycosylati...	CCCTTGAACGTTC...
A_33_P3317580	2.24e-15	2.62e-19	-17.3	33.8	-2.36	A_33_P3317580	NM_001495	2875	GFRA2	GDNF family recepto...	GTGGGGGCTCTG...
A_23_P426305	1.04e-14	1.97e-18	-16.3	31.9	-3.98	A_23_P426305	NM_003734	8839	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	1038	CDO1	cysteine dioxygenase...	AGGAACTTAGGC...
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_P01334	1.04e-14	2.44e-18	-16.2	31.6	-3.25	A_23_P01334	NM_052070	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...	AACCTCACCATG...
A_33_P3376806	1.15e-14	3.38e-18	-16	31.3	-3.39	A_33_P3376806	NM_153714	256815	C10orf67	chromosome 10 ope...	GGAGTTATTGCGCTG...
A_23_P84880	1.25e-14	4.01e-18	-15.9	31.2	-5.18	A_23_P84880	NM_007177	11170	FAM107A	family with sequence...	TCTCTTTGCAGCT...
A_33_P3381338	1.92e-14	7.06e-18	-15.7	30.6	-4.43	A_33_P3381338	NM_019105	7148	TNXB	tenascin XB	AGTTCTCGGTG...
A_33_P3209491	1.92e-14	7.29e-18	-15.6	30.6	-2.98	A_33_P3209491	NM_022848	7145	TNS1	tensin 1	TTGAAACTCTCTG...
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 (Dros...	AAGCAGCAGGGCT...
A_33_P3378514	2.45e-14	1.13e-17	-15.4	30.2	-3.1	A_33_P3378514	NM_001083	8854	PDE5A	phosphodiesterase 5...	CTGTCAGTGGTT...
A_33_P3293918	2.45e-14	1.17e-17	-15.4	30.1	-2.68	A_33_P3293918	NM_170800	10044	SH2D3C	SH2 domain containi...	CCATCTCACTGTAA...
A_23_P67661	2.45e-14	1.38e-17	-15.3	30	-2.55	A_23_P67661	NM_001884	1348	COX7A1	cytochrome c oxidase...	CATCCCGTTGTAC...
A_23_P418774	2.45e-14	1.40e-17	-15.3	29.9	-4.46	A_23_P418774	NM_018029	53405	CLIC5	chloride intracellular...	TGCCCTGATTCACT...
A_23_P252471	2.45e-14	1.40e-17	-15.3	29.9	-2.64	A_23_P252471	NM_000442	6175	PECAM1	platelet/endothelial c...	CCCTCTGTGAAATAC...
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...	TCTGCTGTACATAC...
A_23_P151805	2.96e-14	1.90e-17	-15.2	29.6	-3.38	A_23_P151805	NM_006329	10516	FBLN5	fibrulin 5	GGGAACCTGGGA...
A_33_P3202785	2.96e-14	1.99e-17	-15.2	29.6	-3	A_33_P3202785	210348	PLAC0	placenta-specific 0	CATGTTGAAATAAAA	



# 2. GEO2R. Results

► 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_SYMBOL	SEQUENCE
A_23_P23783	1.64e-16	4.81e-21	-19.5	37.6	-6.63	A_23_P23783	NM_000261	MYOC	ATGCATTTACTACAG...

GSE74706/A\_23\_P23783

Sample values

Lung Tumor

expression value

A_23_P211207	9.99e-16	7.68e-20	-18	35	-3.05	A_23_P211207	NM_001112	ADARB1	GGAACCGATGGGC...
A_23_P93360	9.99e-16	8.77e-20	-17.9	34.9	-5.65	A_23_P93360	NM_001136	AGFR	CCCTTGAACGTGTC...



# 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](#) [YouTube](#)

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													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-18	4.81e-21	-19.5	37.6	-6.63	A_23_P23783	NM_000261	4653	MYOC	myocilin, trabecular ...	ATGCATTACTACA...						
A_23_P211207	9.99e-18	7.68e-20	-18	35	-3.05	A_23_P211207	NM_001112	104	ADARB1	adenosine deaminas...	GGAACCGATGGGC...						
A_23_P93360	9.09e-18	8.77e-20	-17.9	34.9	-5.65	A_23_P93360	NM_001138	177	AGER	advanced glycosylati...	CCCTTGAACGTTC...						
A_33_P3317580	2.24e-15	2.62e-19	-17.3	33.8	-2.36	A_33_P3317580	NM_001495	2875	GFRA2	GDNF family recepto...	GTGGGGGCTCTG...						
A_23_P426305	1.04e-14	1.97e-18	-16.3	31.9	-3.98	A_23_P426305	NM_003734	8839	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	1038	CDO1	cysteine dioxygenase...	AGGAACTTAGCCG...						
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_P01334	1.04e-14	2.44e-18	-16.2	31.6	-3.25	A_23_P01334	NM_052070	116835	HSPA1B2	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 ...	AACCTCACCATG...						
A_33_P3376806	1.15e-14	3.38e-18	-16	31.3	-3.39	A_33_P3376806	NM_153714	256815	C10orf67	chromosome 10 ope...	GGAGTTATTGCGCTG...						
A_23_P84880	1.25e-14	4.01e-18	-15.9	31.2	-5.18	A_23_P84880	NM_007177	11170	FAM107A	family with sequence...	TCTCTTTGCAGCT...						
A_33_P3381338	1.92e-14	7.06e-18	-15.7	30.6	-4.43	A_33_P3381338	NM_019105	7148	TNXB	tenascin XB	AGTTCTCGGTGCCC...						
A_33_P3209491	1.92e-14	7.29e-18	-15.6	30.6	-2.98	A_33_P3209491	NM_022848	7145	TNS1	tensin 1	TTGAAACTCTCTG...						
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 (Dros...	AAGCAGCAGGGCT...						
A_33_P3378514	2.45e-14	1.13e-17	-15.4	30.2	-3.1	A_33_P3378514	NM_001083	8854	PDE5A	phosphodiesterase 5...	CTGTCAGTGGTT...						
A_33_P3293918	2.45e-14	1.17e-17	-15.4	30.1	-2.68	A_33_P3293918	NM_170800	10044	SH2D3C	SH2 domain containi...	CCATCTCACTGTAA...						
A_23_P67661	2.45e-14	1.38e-17	-15.3	30	-2.55	A_23_P67661	NM_001884	1348	COX7A1	cytochrome c oxidase...	CATCCCGTTGTAC...						
A_23_P418774	2.45e-14	1.40e-17	-15.3	29.9	-4.46	A_23_P418774	NM_018029	53405	CLIC5	chloride intracellular...	TGCCCTGATTCAACT...						
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...	CCCTCTGTGAAATAC...						
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_153386	79987	SVEP1	sushi, von Willebrand...	TCTGCTGTACATAC...						
A_23_P151805	2.96e-14	1.90e-17	-15.2	29.6	-3.38	A_23_P151805	NM_006329	10516	FBLN5	fibrulin 5	GGGAACCTGGGA...						
A_33_P3292785	2.96e-14	1.99e-17	-15.2	29.6	-3	A_33_P3292785		219348	PLAC0	placenta-specific 0	CATATGTAATAAAAAA...						



# 2. GEO2R. Results

Autoguardado □ Buscar □ Lara Nonell

Inicio Insertar Disposición de página Fórmulas Datos Revisar Vista Ayuda Foxit PDF Compartir Comentarios

Calibri 11 A A General Formato condicional □ Insertar □ Sumatoria □ Ordenar y filtrar □ Ideas

N K S □ Dar formato como tabla □ Eliminar □ Buscar y seleccionar □ Ideas

Portapapeles □ Estilos de celda □ Formato □ Ideas

Fuente Alineación Número Estilos Celdas Edición Confidencialidad

N123

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
2	A_23_P2378	1.64e-16	4.81e-21	-19.529.454	37.625.655	-6.63	A_23_P2378	NM_000261	MYOC	ATGCATTACTACAGTTGGCTCTAATGCTTCAGATAGAACATACAGTTGGCTCACATAA					
3	A_23_P2112	9.99e-16	7.68e-20	-17.978.458	34.979.518	-3.05	A_23_P2112	NM_001112	ADARB1	GGAACCGATGGGCATTAACATGAACCTGAAACGGTAAAGCAGCTATGGAACGCTA					
4	A_23_P9336	9.99e-16	8.77e-20	-17.906.747	3.485.214	-5.65	A_23_P9336	NM_001136	AGER	CCCTTGAACGTGTTCTGGCCTCAGACCAACTCTCTCTGTATAATCTCTCTGTATAAC					
5	A_33_P3317	2.24e-15	2.62e-19	-17.323.259	33.798.303	-2.36	A_33_P3317	NM_001495	GFRA2	GTGGGGCTCTGATCCGATCCAAGCTAACCAAGGCTCAATAAACGTGCTAGGAAGC					
6	A_23_P4263	1.04e-14	1.97e-18	-1.628.866	31.850.378	-3.98	A_23_P4263	NM_003734	AOC3	AAGGCAAGGCCAGAAAATGTGTATAGCGCACTTCCCATTGTGTTTAGAAAGGAGTAGA					
7	A_23_P3029	1.04e-14	2.25e-18	-16.222.511	31.722.245	-4.34	A_23_P3029	NM_001801	CDO1	AGGAACCTACGGCACTTGCCTATTGTTATGAAAGGAAGTCAGAGGAGTTAGAAAAGT					
8	A_23_P5339	1.04e-14	2.40e-18	-16.189.176	31.657.506	-3.32	A_23_P5339	NM_002837	PTPRB	ATGGTCCAGACTGAGTGTCACTGTCTACCATCAGTGTGTAAGAGATGTCTCAGA					
9	A_23_P9133	1.04e-14	2.44e-18	-16.180.222	31.640.098	-3.25	A_23_P9133	NM_052970	HSPA12B	ATCTCTAATGTGGAGGTGGGAACATTATTGTTGGAGGCAATTATGAGGGTAGCATTC					
10	A_23_P3444	1.05e-14	2.76e-18	-16.120.576	31.523.925	-3.38	A_23_P3444	NM_019055	ROBO4	AACCTCACCATGGAAAGAAAATAATTGAATGCCACTGAGGCACTGAGGCCCTACCTCA					
11	A_33_P3376	1.15e-14	3.38e-18	-16.019.764	31.326.748	-3.39	A_33_P3376	NM_153714	C10orf67	GGAGTTATCGCTGGGGCTTCTCTCAGCTAACAGCTAACAGATATCAGGATGACCT					
12	A_23_P8486	1.25e-14	4.01e-18	-15.935.286	31.160.713	-5.18	A_23_P8486	NM_007177	FAM107A	TCTCTTGAGCTTAGGATGGCTATGTGATCAGGTGTTGCCAATGAAATTGAAGAGGAA					
13	A_33_P3381	1.92e-14	7.06e-18	-15.657.744	30.610.016	-4.43	A_33_P3381	NM_019105	TNXB	AGTTCTCGGTGCCCTCACGGAAATGAAGCTGAGACCAAGAACCTTCGCTCCCCAGCGG					
14	A_33_P3209	1.92e-14	7.29e-18	-15.642.224	30.578.984	-2.98	A_33_P3209	NM_022648	TNS1	TTGAAACTCTCTGTGTTGTTGTTGCGTGTGAGAGCACATCAGTGT					
15	A_23_P1443	2.45e-14	1.01e-17	-15.482.525	30.258.181	-3.15	A_23_P1443	NM_004787	SLT2	AAGCAGCAGGGCTATGCTGCTTGCCAAACAACCAAGAAGGTGTCGGATTAGAGTGCAGA					
16	A_33_P3378	2.45e-14	1.13e-17	-15.428.981	30.150.014	-3.10	A_33_P3378	NM_001083	PDE5A	CTGTCAGTGGTTCTCTATTGCGATGTTGAAAAATAAAAAGAATGATCAAGTAGG					
17	A_33_P3293	2.45e-14	1.17e-17	-15.414.878	30.121.473	-2.58	A_33_P3293	NM_170600	SH2D3C	CCATCTCACTGAAATAAGCTCTGTTGAAATAGATGTCAGAACGATGTTATTC					
18	A_23_P6766	2.45e-14	1.38e-17	-15.334.439	29.958.277	-2.55	A_23_P6766	NM_001864	COX7A1	CATCCCGTTGACTCTGAAAGGGCGGCATGTTGACAAACATCTGTACCGAGTGACAATGAC					
19	A_23_P4167	2.45e-14	1.40e-17	-15.329.226	29.947.677	-4.46	A_23_P4167	NM_016929	CLIC5	TGCCTGATTCAACTGAGGCTGAAATGGTAAAGCCACATTAGGAGGTGGCTGATCA					
20	A_23_P2524	2.45e-14	1.40e-17	-15.326.698	29.942.536	-2.64	A_23_P2524	NM_000442	PECAM1	CGCCTGTGAAATACCAACCTGAAAGACACGGTTCATTAGGCAACGCCACAAACAGAAAAT					
21	A_33_P3248	2.45e-14	1.43e-17	-15.316.614	29.922.019	-4.98	A_33_P3248	NM_001199	INMT	TCTGTCACCAATGCTGCCAACATGGGGCTGCTTATTGTGGCTCGCAAGAAGCCTGGG					



## 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

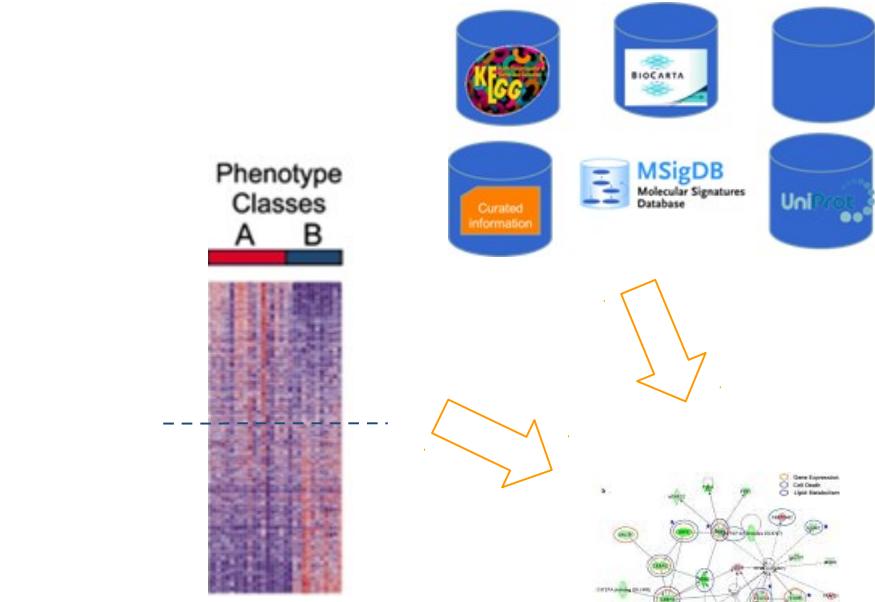
Several databases:

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

Several statistical approaches

Several tools:

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

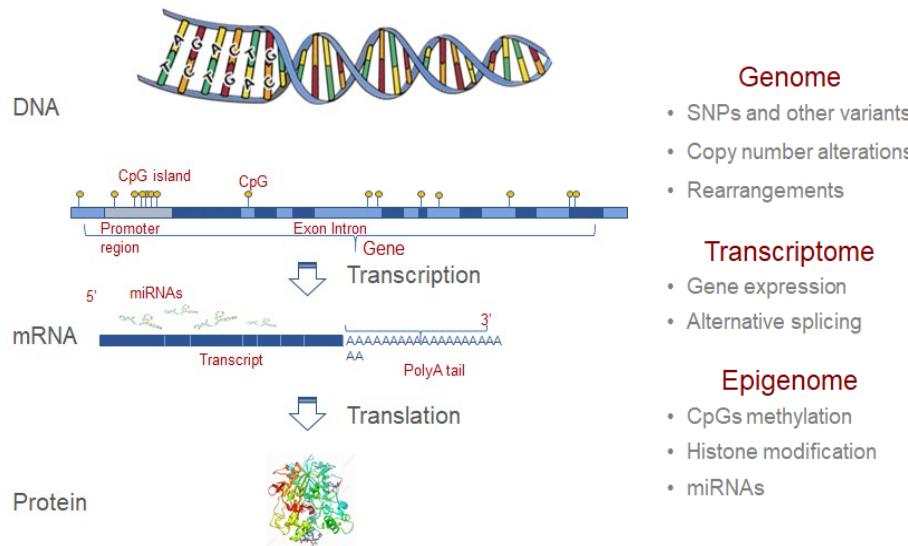


Differentially expressed genes

# 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

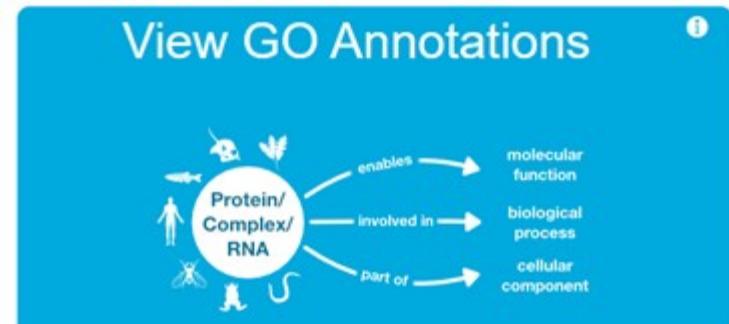
Biological process in which gene is involved

Ex: mitosis, apoptosis

### - Cellular Component

Cell location of gen depending on functions

Ex: nucleous, telomeres



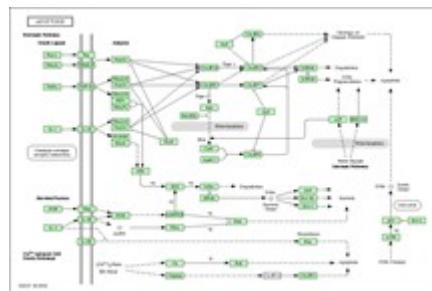
- GO:0003673 : Gene Ontology (65883) ●
  - ⓘ GO:0008150 : biological process (44405) ●
    - ⊕ ⓘ GO:0007610 : behavior (357)
      - ⓘ GO:0000004 : biological process unknown (7877)
    - ⊕ ⓘ GO:0009987 : cellular process (32672) ●
      - ⊕ ⓘ GO:0007154 : cell communication (5384)
      - ⊕ ⓘ GO:0008219 : 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)



### 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

**UC San Diego**  
**BROAD INSTITUTE**

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

**MSigDB**  
Molecular Signatures Database

Molecular Signatures Database v7.0

**Overview**

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this website, you can

- ▶ [Search](#) for gene sets by keyword.
- ▶ [Browse](#) gene sets by name or collection.
- ▶ [Examine](#) a gene set and its annotations. See, for example, the [GO\\_NOTCH\\_SIGNALING\\_PATHWAY](#) gene set page.
- ▶ [Download](#) gene sets.
- ▶ [Investigate](#) gene sets:
  - ▶ [Compute overlaps](#) between your gene set and gene sets in MSigDB.
  - ▶ [Categorize](#) members of a gene set by gene families.
  - ▶ [View the expression profile](#) of a gene set in a provided public expression compendia.

**Collections**

The MSigDB gene sets are divided into 8 major collections:

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

**License Terms**

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

**Current Version**

MSigDB database v7.0 updated August 2019. Release notes.

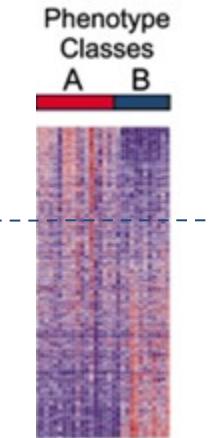
**Gene Sets from Community Contributors**



# 3. Statistical approaches

---

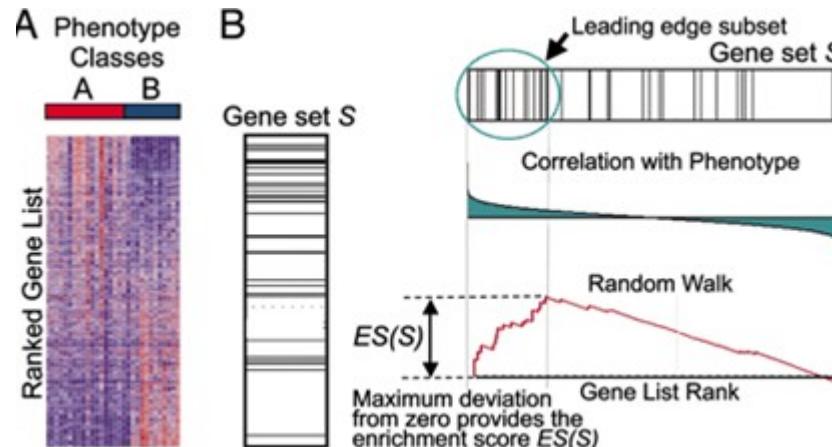
- Specific terms:
  - Gene set
  - (gene) Universe
- How to detect enrichment
  - Mere overlapping with a gene set
  - Statistical overrepresentation of a gene list in a gene set
    - Fisher's exact test (or hypergeometric)
      - List of differentially expressed genes (DEG)
      - Human genome (background/universe)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

**DAVID Bioinformatics Resources 6.8**  
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

[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](#). \*\*\*

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

**Welcome to DAVID 6.8**

2003 - 2020

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 [comprises a full Knowledgebase update to the sixth version](#) of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

**What's Important in DAVID?**

- [Cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

**Statistics of DAVID**

DAVID Citations (2003-2019)

Year	Citations
03	~100
04	~200
05	~300
06	~400
07	~500
08	~600
09	~700
10	~800
11	~900
12	~1000
13	~1100
14	~1200
15	~1300
16	~1400
17	~1500
18	~1600
19	~1700

- > 38,000 Citations
- Average Daily Usage: ~2,700 gene lists/sublists from ~900 unique researchers.
- Average Annual Usage: ~1,000,000 gene lists/sublists from >100 countries

# 3. DAVID

The screenshot shows the DAVID Bioinformatics Resources 6.8 Analysis Wizard interface. The left panel is the submission form, and the right panel is the results summary.

**Left Panel: Upload Gene List**

- Step 1: Enter Gene List**
  - A: Paste a list
    - MYOC
    - ADARB1
    - AGFR
    - GFRA2
  - 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
- Step 4: Submit List**
  - Submit List

**Right Panel: Analysis Wizard**

\*\*\* Welcome to DAVID 6.8 \*\*\*  
\*\*\* If you are looking for DAVID 6.7, please visit our [development site](#). \*\*\*

Step 1. Successfully submitted  
Current Gene List: List\_1  
Current Background: Homo sapiens

Step 2. Analyze above gene list  
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  
194 DAVID IDs  
Check Defaults  Clear All  
Help and Tool Manual

\*\*\* 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