

# Bases de datos biológicas

Karen Oróstica  
28 de octubre, 2025

1

## Contenido

- Introducción
- Bases de datos
- Características y clasificación
- Pubmed
- cbiportal

2

## ¿Qué son las bases de datos?

- Es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior utilización.



Objetivo

Organizar los datos en un conjunto de *registros* estructurados que permitan recuperar fácilmente la información .

3

## ¿Qué son las bases de datos?

- Cada **registro** está compuesto por un número determinado de **campos** que contienen datos específicos.

| Cve. cliente | Nombre           | Direccion      | Ciudad      | Estado     |
|--------------|------------------|----------------|-------------|------------|
| 1            | Alfredo Godinez  | Fresnillo #47  | Veracruz    | Veracruz   |
| 2            | Gabriela Mora    | El creso #81   | Guadalajara | Jalisco    |
| 3            | Alejandra Avalos | Casa Mata #1   | Morelia     | Michoacan  |
| 4            | Jaime Quintero   | Miraflores #23 | Uruapan     | Michoacan  |
| 5            | Carlos Miranda   | Rio Bravo #95  | Matamoros   | Tamaulipas |

4

## ¿Qué son las bases de datos?

- Para recuperar un registro particular de la base de datos, un usuario puede especificar una pieza de información, llamada **valor**, que será encontrada en un campo en especial.
- La computadora entonces recuperará el registro completo.
- Este proceso es llamado **consulta**.
- La recuperación de información es el principal objetivo de todas las BD.

5

## ¿Qué son las bases de datos?

- Las BD biológicas a menudo tienen un requerimiento de más alto nivel, conocido como **descubrimiento de conocimiento**.



**Identificación de conexiones entre piezas de información que no eran conocidas cuando la información fue introducida.**

6

## ¿Qué son las bases de datos?

- BD que contienen información cruda (sin procesar) de secuencias de ADN se pueden realizar tareas extras para identificar **homología de secuencias o motivos conservados**.



**Facilitar el descubrimiento de nuevos conocimientos biológicos**

7

## Tipos de BD

- Registros de datos crudos e instrucciones operacionales (estructuras de datos) ayudan a **identificar las conexiones ocultas** entre los registros.
- El propósito de establecer una estructura de datos es para facilitar la ejecución de las consultas y para combinar diferentes registros con el fin de formar informes de consultas.
- Dependiendo de los tipos de estructuras de datos, estos sistemas se pueden clasificar en diferentes tipos: **jerárquicos, de red, relacionales, orientados a objetos**.

8

## Tipos de BD: flat

### Tipos de BD: flat

- Bases de datos de texto pl que contiene varias entrad
- No contienen instruccione específica o para la creació
- Ejemplos: GenBank y UniP

Name, States, Course number,  
Bioinformatics|Jane Doe, Kans  
Illinois, Chem 289, Organic Che  
Horticulture|Howard Doug

[Display Settings](#): ☐ GenBank

Homo sapiens tumor susceptibility gene 101 (TSG101), mRNA

NCBI Reference Sequence: NM\_006292.3

[FASTA](#) [Graphics](#)

[Go to:](#) ☐

LOCUS NM\_006292 1562 bp mRNA linear PRI 25-NOV-2012  
DEFINITION Homo sapiens tumor susceptibility gene 101 (TSG101), mRNA.  
ACCESSION NM\_006292  
VERSION NM\_006292.3 GI:332000018  
KEYWORDS .  
SOURCE Homo sapiens (human)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
Cathartini; Hominoidea; Homo.  
REFERENCE 1 (bases 1 to 1562)  
AUTHORS Kaneyama,T., Suzuki,H. and Mayeda,A.  
TITLE Re-splicing of mature mRNA in cancer cells promotes activation of  
distant weak alternative splice sites  
JOURNAL Nucleic Acids Res. 40 (16), 7896-7906 (2012)  
PUBMED 22675076  
REMARK GeneRIF: The results provide evidence for a two-step splicing  
pathway of the TSG101 mRNA in which the initial constitutive  
splicing removes all 14 authentic splice sites, thereby bringing  
the weak alternative splice sites into close proximity.  
REFERENCE 2 (bases 1 to 1562)  
AUTHORS Gu,R.J., Wang,S.C., Sun,G., Zhuang,B.W. and Liu,D.L.  
TITLE [Expression and significance of tumor susceptibility gene 101 in  
hepatocellular carcinoma tissues]  
JOURNAL Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 28 (7), 738-740 (2012)  
PUBMED 22768867  
REMARK GeneRIF: The expression of TSG101 in HCC is higher than that in  
corresponding non-cancer tissues and the expression level is  
closely correlated with TNM stage and metastasis of HCC.  
REFERENCE 3 (bases 1 to 1562)  
AUTHORS Horgan,C.P., Hanscom,S.R., Kelly,E.E. and McCaffrey,M.W.  
TITLE Tumor susceptibility gene 101 (TSG101) is a novel binding-partner  
for the class II Rab11-FTS  
JOURNAL PLoS ONE 7 (2), E32030 (2012)  
PUBMED 22348143  
REMARK GeneRIF: Identified TSG101 as a novel FIP4-binding protein, which  
can also bind FIP3. alpha-helical coiled-coil regions of both  
TSG101 and FIP4 mediate the interaction with the cognate protein  
REFERENCE 4 (bases 1 to 1562)  
AUTHORS Nagashima,S., Takahashi,M., Jirintai,S., Tanaka,T., Mishizawa,T.,

9

## Tipos de BD: flat

- El archivo de texto puede ser considerado como una tabla única.



Para buscar en un archivo plano  
se debe leer el archivo completo!!!!



**Ineficiente!!**

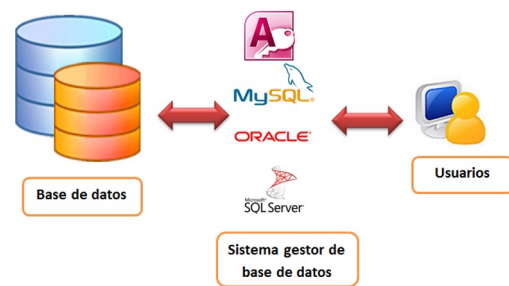
10

## Tipos de BD: flat

- Uso intensivo de la memoria, provoca fallos en el sistema

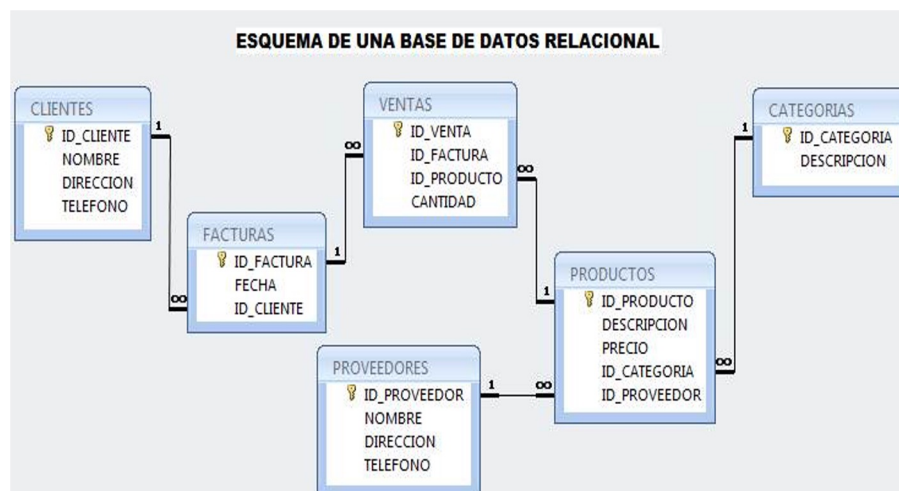


### Sistemas de gestión de bases de datos



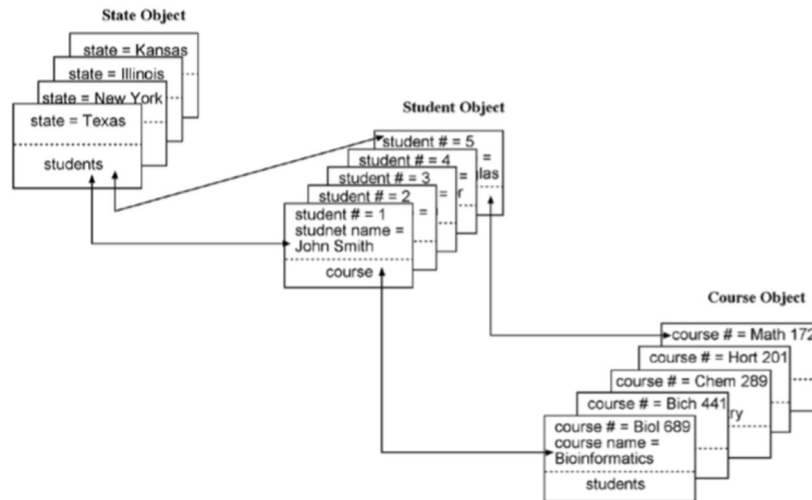
11

## Tipos de BD: BD relacional



12

## Tipos de BD: Orientadas a objetos



13

## Bases de datos biológicas

- Tres principales tipos de sistemas de gestión de bases de datos utilizados para almacenar datos biológicos:
  1. Archivos planos
  2. BD relacionales
  3. BD orientadas a objetos
- Muchas BD biológicas siguen utilizando el formato flat. (no necesita conocimientos de BD)

14

## ¿Por qué son necesarias bases de datos?

- **Diluvio de datos biológicos**

Experimentos High-throughput, genómica, metagenómica, proteómica, metabolómica etc..

- **Necesidad de almacenar y comunicar grandes conjuntos de datos.**

Métodos para almacenamiento adecuado, la búsqueda y recuperación de datos.



Las bases de datos son los medios para manejar grandes volúmenes de datos.

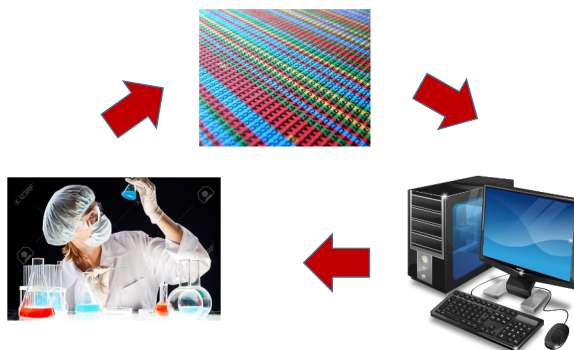


15

## Capacidades de las bases de datos

- **Disponibilidad de datos biológicos**

1. Análisis computacional.
2. Manejar y compartir grandes volúmenes de datos.
3. Interfaz para sistemas (algoritmos, interfaces Web).
4. Almacenamiento automatizado de datos (definición de formatos).
5. Recuperación de datos experimentales.



16



## Clasificación

### Tipos de datos:

1. Secuencias de proteínas y nucleótidos
2. Anotación de secuencias
3. Patrones y motivos macromoleculares
4. Estructuras 3D
5. Datos de expresión génicos
6. Rutas metabólicas
7. Mutaciones en genes asociados a cáncer

17

## Clasificación de DB: Basado en su contenido

1. **Primarios:** Resultados experimentales ingresados a la BD.  
*Ejemplos: GenBank como PDB (Protein Data Bank).*
2. **Secundarios:** Análisis de resultados ingresados a la BD.  
*Ejemplos: SWISS-Prot y PIR (Protein Information Resources).*
3. **Especializadas:** dedicadas un interés de investigación particular.
4. *Ejemplos: Flybase, HIV sequence database, y Ribosomal Database Project*

### Diseño técnico

1. Flat-files
2. Bases de datos relacionales(SQL)
3. Bases de datos orientado a objeto
4. Tecnologías intercambio / replicación (FTP, HTML, COBRA, XML, SOAP)

19

## The National Center for Biotechnology Information (NCBI)

The screenshot shows the NCBI homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar with 'All Databases' selected. A left sidebar lists various resources like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area is titled 'Welcome to NCBI' and includes a brief description of the center's mission. Below this, there are six interactive tiles: 'Submit' (Deposit data or manuscripts into NCBI databases), 'Download' (Transfer NCBI data to your computer), 'Learn' (Find help documents, attend a class or watch a tutorial), 'Develop' (Use NCBI APIs and code libraries to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research and collaborative projects). Each tile has a corresponding icon.

20

## ¿Cómo acceder a los datos?

### Interfaces web

Búsqueda mediante palabras claves, modificadores o identificadores.

The screenshot shows the UniProtKB search interface. It has a header with 'UniProt' and 'UniProtKB'. Below the header are tabs for 'Search', 'Blast', 'Align', 'Retrieve', and 'ID Mapping \*'. The 'Search' tab is active. Under 'Search in', there's a dropdown menu set to 'Protein Knowledgebase (UniProtKB)'. The 'Query' field contains the text 'glucokinase homo sapiens'. A 'Search' button is located to the right of the query field.

**Web service** (SOAP, CORBA)

**Flat files** (script based, large scale)

**Database dump** (script based, large scale)

21

## Estrategias adoptadas: rentrez

1. **El NCBI comparte muchos datos.** La base de datos de nucleótidos NCBI (que incluye GenBank) tiene datos para 256.7 millones de secuencias diferentes, y dbSNP describe 1070.2 millones de variantes genéticas diferentes. Todos estos registros pueden ser referenciados con los 1.33 millones de especies en la taxonomía NCBI o 25.7 mil registros asociados con enfermedades en OMIM.
1. El NCBI hace que estos datos estén disponibles a través de una interfaz web, un servidor FTP y una API REST llamada Entrez Utilities (Eutils para abreviar).
1. **rentrez** proporciona funciones para usar una API, permitiendo a los usuarios recopilar y combinar datos de múltiples bases de datos NCBI en la comodidad de una sesión R o script.



22

## Estrategias adoptadas: rentrez

```
entrez_dbs()
```

```
## [1] "pubmed"      "protein"     "nuccore"
## [4] "ipg"         "nucleotide" "nucgss"
## [7] "nucest"     "structure"   "sparcle"
## [10] "genome"     "annotinfo"   "assembly"
## [13] "bioproject" "biosample"   "blastdbinfo"
## [16] "books"      "cdd"         "clinvar"
## [19] "clone"      "gap"         "gapplus"
## [22] "grasp"      "dbvar"       "gene"
## [25] "gds"        "geoprofiles" "homologene"
## [28] "medgen"     "mesh"        "ncbisearch"
## [31] "nlmcatalog" "omim"        "orgtrack"
## [34] "pmc"        "popset"      "probe"
## [37] "proteinclusters" "pcassay"    "biosystems"
## [40] "pccompound" "pcsubstance" "pubmedhealth"
## [43] "seqannot"   "snp"         "sra"
## [46] "taxonomy"   "biocollections" "unigene"
## [49] "gencoll"    "gtr"
```

23

## Estrategias adoptadas: rentrez

```
library(rentrez)

vivax_search <- entrez_search(db = "taxonomy",
                             term = "Streptomyces leeuwenhoekii[ORGN]")

multi_summs <- entrez_summary(db="taxonomy", id=vivax_search$ids)
uid          <- extract_from_esummary(multi_summs, "uid")
status       <- extract_from_esummary(multi_summs, "status")
division     <- extract_from_esummary(multi_summs, "division")
genus        <- extract_from_esummary(multi_summs, "genus")
species      <- extract_from_esummary(multi_summs, "species")
scientificname <- extract_from_esummary(multi_summs, "scientificname")

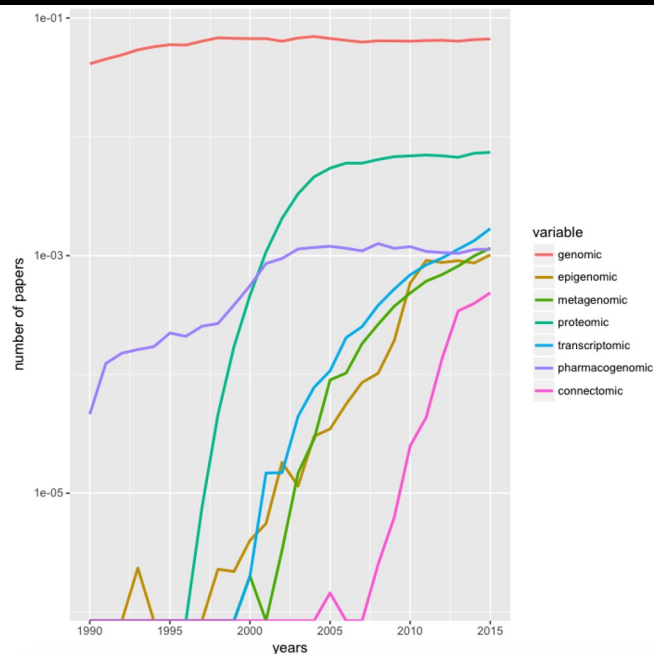
taxonomy <- data.frame(uid, scientificname, status, division, genus, species)
```



| uid     | scientificname             | status | division      | genus        | species       |
|---------|----------------------------|--------|---------------|--------------|---------------|
| 1437453 | Streptomyces leeuwenhoekii | active | high GC Gram+ | Streptomyces | leeuwenhoekii |

24

## Estrategias adoptadas: rentrez



"Term" y par de años

25

## **Revisión bibliográfica sistematizada**

- 1. Proceso de búsqueda de un tema o tópico en la literatura**
- 2. Entendiendo MeSH term**
- 3. Técnicas de búsqueda en bases de datos**
- 4. Búsqueda avanzada en Pubmed**

26

## **Búsqueda de un tema**

**“Caracterización genómica de pacientes no fumadores con cáncer de pulmón”**

27

## Búsqueda de un tema

**“Caracterización genómica de pacientes no fumadores con cáncer de pulmón”**



*Descomponer en términos*

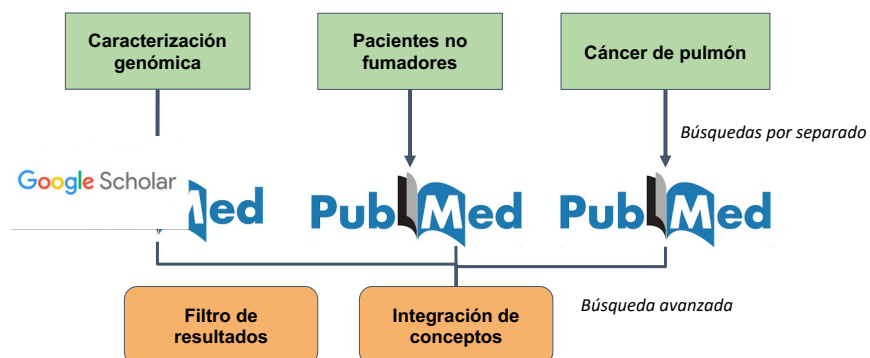
28

## Búsqueda de un tema

**“Caracterización genómica de pacientes no fumadores con cáncer de pulmón”**



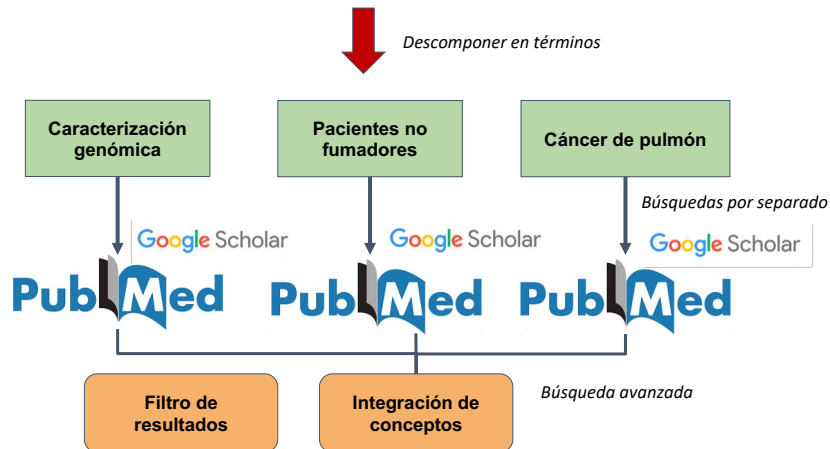
*Descomponer en términos*



29

## Búsqueda de un tema

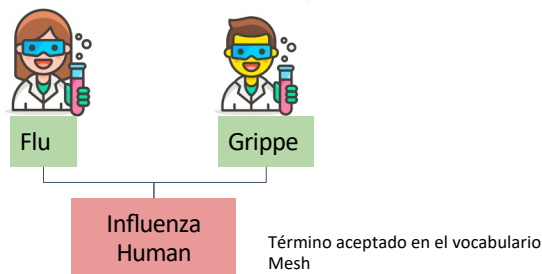
**“Caracterización genómica de pacientes no fumadores con cáncer de pulmón”**



30

## MeSH term

1. El MeSH (Medical Subject Headings) es un vocabulario controlado que contiene los descriptores utilizados en la base de datos.
2. Cada registro de PubMed tiene asignados unos términos (descriptores) que definen de manera exacta el tema que analiza.



31

## Técnicas de búsquedas

### 1. Truncamiento, para encontrar variantes

model\* = models, modeling, modelling

### 2. Uso de comillas

“emergency medicine” **versus** emergency medicine

### 3. Parentesis

Permite combinar conceptos (child\* OR adolescen\*)

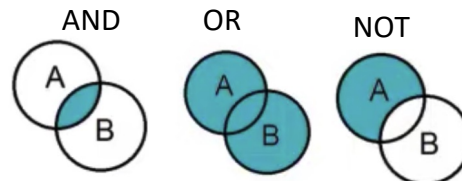
### 4. Operadores Booleanos

AND, OR, NOT

child\* AND “head injury”

child\* OR pedriat\*

child\* NOT cancer



32

## Busqueda avanzada

NIH National Library of Medicine  
National Center for Biotechnology Information

Log in

PubMed Advanced Search Builder

PubMed.gov  
User Guide

Add terms to the query box

Title/Abstract Enter a search term NOT Show Index

Query box

((mutational signatures[Title/Abstract]) AND (lung cancer[Title/Abstract])) NOT (smok\*[Title/Abstract]) Search

33



## Búsqueda avanzada

NIH National Library of Medicine  
National Center for Biotechnology Information

PubMed.gov

Search: ((mutational signatures[Title/Abstract]) AND (lung cancer[Title/Abstract]))

Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Best match Display options

MY NCBI FILTERS

RESULTS BY YEAR

2017 2018 2019 2020 2021 2022

TEXT AVAILABILITY

☐ Abstract  
☐ Free full text  
☐ Full text

ARTICLE ATTRIBUTE

☐ Associated data

ARTICLE TYPE

☐ Books and Documents  
☐ Clinical Trial  
☐ Meta-Analysis  
☐ Randomized Controlled Trial  
☐ Review  
☐ Systematic Review

25 results

1 Association of *LRP1B* Mutation With Tumor Mutation Burden and Outcomes in Melanoma and Non-small Cell Lung Cancer Patients Treated With Immune Check-Point Blockades.  
Chen H, Chong W, Wu Q, Yao Y, Mao M, Wang X.  
Front Immunol. 2019 May 21;10:1113. doi: 10.3389/fimmu.2019.01113. eCollection 2019.  
PMID: 31164891 Free PMC article.  
LRP1B (low-density lipoprotein receptor-related protein 1B) is frequently mutated in melanoma, non-small cell lung cancer (NSCLC) and other tumors; however, its association with TMB and survival in patients with immunotherapy remains unknown. ...Bayesian variants no ...

2 Uncovering and characterizing splice variants associated with survival in lung cancer patients.  
West S, Kumar S, Batra SK, Ali H, Ghersi D.  
PLoS Comput Biol. 2019 Oct 25;15(10):e1007469. doi: 10.1371/journal.pcbi.1007469. eCollection 2019 Oct.  
PMID: 31652257 Free PMC article.  
To computationally validate our findings, we characterized the **mutational signatures** in patients, grouped by low and high expression of a splice variant associated with patient survival and involved in DNA repair. The results of the **mutational signature** analy ...

3 Clinicopathological, microenvironmental and genetic determinants of molecular subtypes in KEAP1/NRF2-mutant lung cancer.  
Cai MC, Chen M, Ma P, Wu J, Lu H, Zhang S, Liu J, Zhao X, Zhuang G, Yu Z, Fu Y.  
Int J Cancer. 2019 Feb 15;144(4):788-801. doi: 10.1002/ijc.31975. Epub 2018 Dec 4.  
PMID: 30411339 Free article.  
Somatic KEAP1-NRF2 pathway alterations are frequently detected in both lung adenocarcinomas and squamous cell carcinomas. However, the biological characteristics and molecular subtypes of KEAP1/NRF2-mutant lung cancer remain largely undefined. ...First, we di ...

34

## ¿Qué es Cbiportal?

cBioPortal FOR CANCER GENOMICS

Data Sets Web API R/MATLAB Tutorials FAQ News Visualize Your Data About

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)

PanCancer Studies 3 ☐ Select all listed studies (233)

Cell lines 2

Adrenal Gland 2

Ampulla of Vater 1

Biliary Tract 6

Bladder/Urinary Tract 11

Bone 2

Bowel 8

Breast 14

CNS/Brain 15

PanCancer Studies

☐ MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017) 10945 samples

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

☐ Pediatric Mixed Tumors (PIP-Seq 2017) 103 samples

Cell lines

☐ Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012) 1020 samples

☐ NCI-60 Cell Lines (NCI, Cancer Res. 2012) 67 samples

Adrenal Gland

Adrenocortical Carcinoma

☐ Adrenocortical Carcinoma (TCGA, PanCancer Atlas) 92 samples

☐ Adrenocortical Carcinoma (TCGA, Provisional) 92 samples


Ampulla of Vater


Enter Genes: Advanced: Onco Query Language (OQL)

User-defined List

Enter HUGO Gene Symbols or Gene Aliases

35

 **BMC** Part of Springer Nature



Search 

**Journal of Translational Medicine**

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#)

Research | [Open Access](#) | [Published: 18 August 2022](#)

**Total mutational load and clinical features as predictors of the metastatic status in lung adenocarcinoma and squamous cell carcinoma patients**

[Karen Y. Oróstica](#), [Juan Saez-Hidalgo](#), [Pamela R. de Santiago](#), [Solange Rivas](#), [Sebastian Contreras](#), [Gonzalo Navarro](#), [Juan A. Asenjo](#), [Álvaro Olivera-Nappa](#)  & [Ricardo Armisen](#) 

[Journal of Translational Medicine](#) **20**, Article number: 373 (2022) | [Cite this article](#)

581 Accesses | 2 Altmetric | [Metrics](#)