Bases de datos biológicas

Karen Oróstica

28 de octubre, 2025

1

Contenido

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

2

¿Qué son las bases de datos?

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

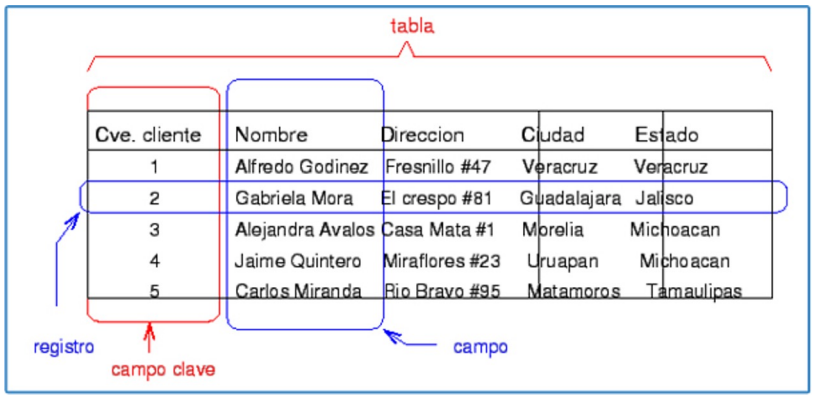
Objetivo

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.



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 Dou

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;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1562)
  AUTHORS   Kameyama,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-FIPs
  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., Nishizawa,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!!!!

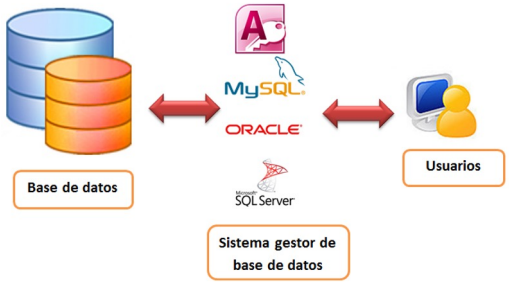
Ineficiente!!

10

Tipos de BD: flat

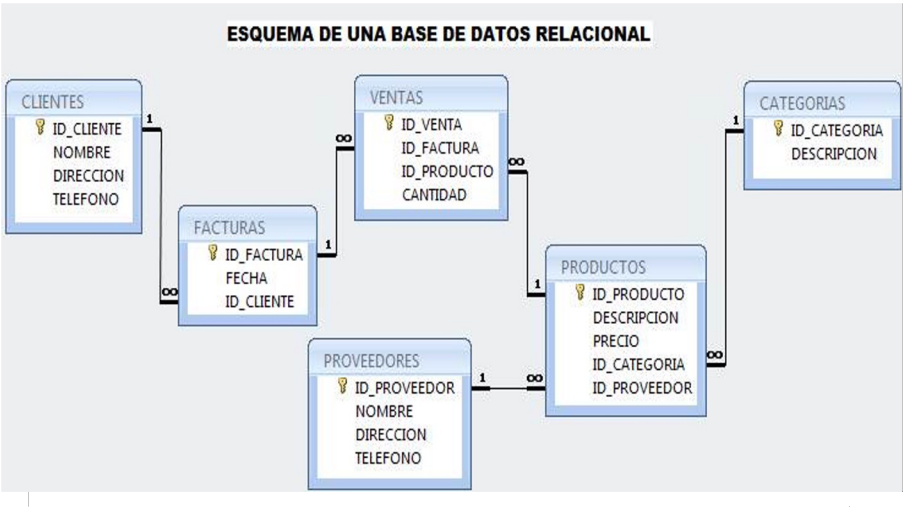
- Uso intensivo de la memoria, provoca fallos en el sistema

Sistemas de gestión de bases de datos



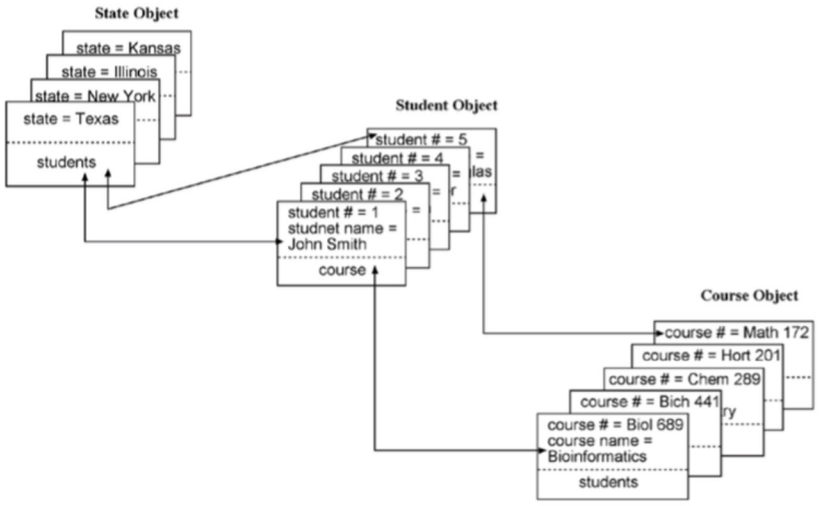
11

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.

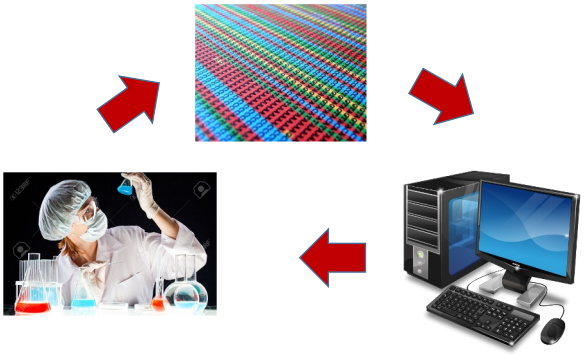
Las bases de datos son los medios para manejar grandes volúmenes de datos.

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

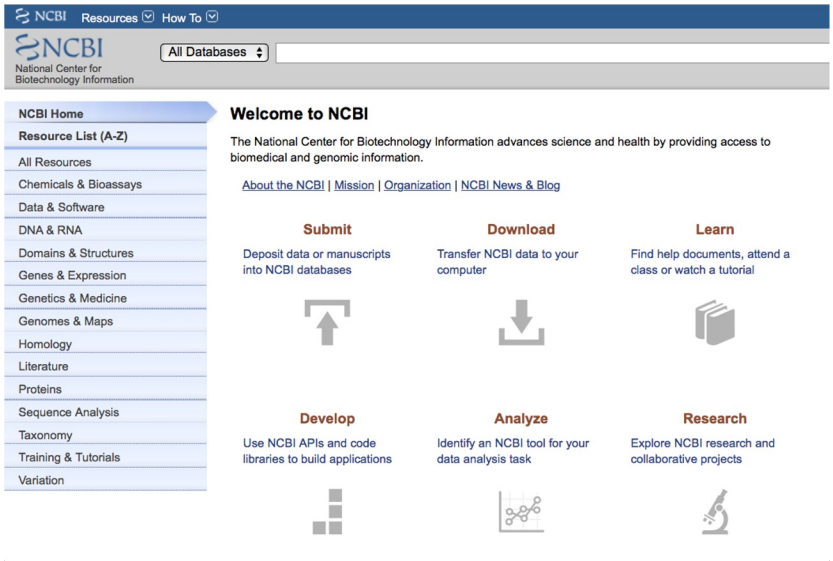
1. **Primarios**: Resultados experimentales ingresados a la BD.
   Ejemplos: *GenBank* como *PDB* (*Protein Data Bank*).
2. **Secundarios**: Análisis de resultados ingresados a la BD.
   Ejemplos: *SWISS-Prot* y *PIR* (*Protein Information Resources*).
3. **Especializadas**: dedicadas un interés de investigación 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)



20

¿Cómo acceder a los datos?

Interfaces web

Búsqueda mediante palabras claves, modificadores o identificadores.

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.

National Center for Biotechnology Information
NCBI

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

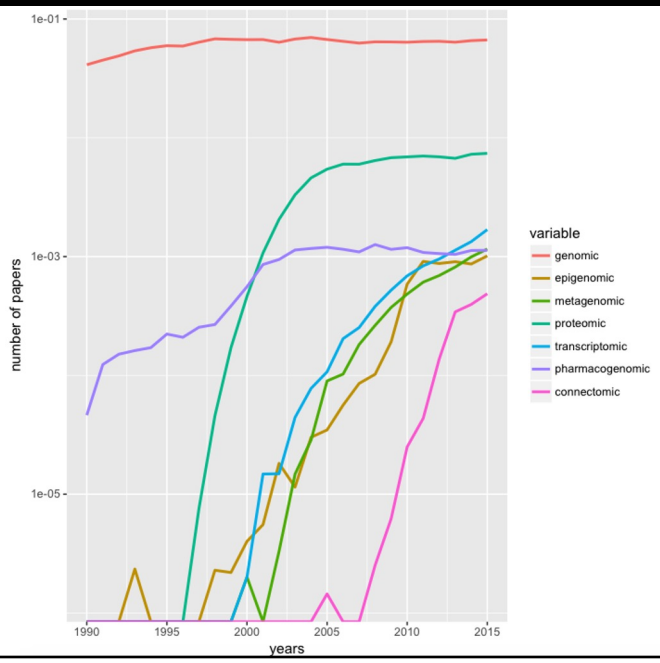
Estrategias adoptadas: rentrez

```
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”

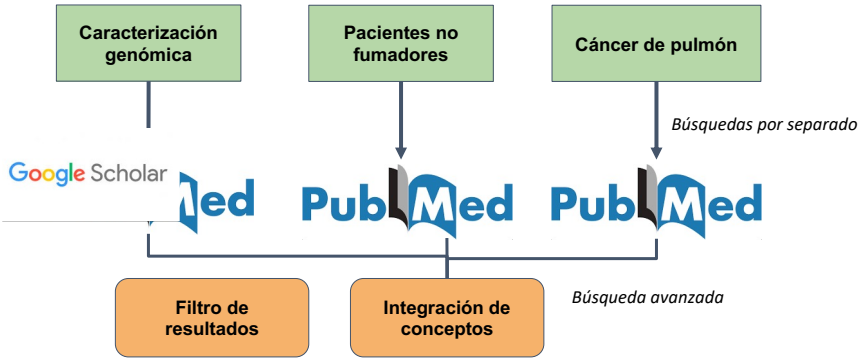
Descomponer en términos

28

Búsqueda de un tema

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

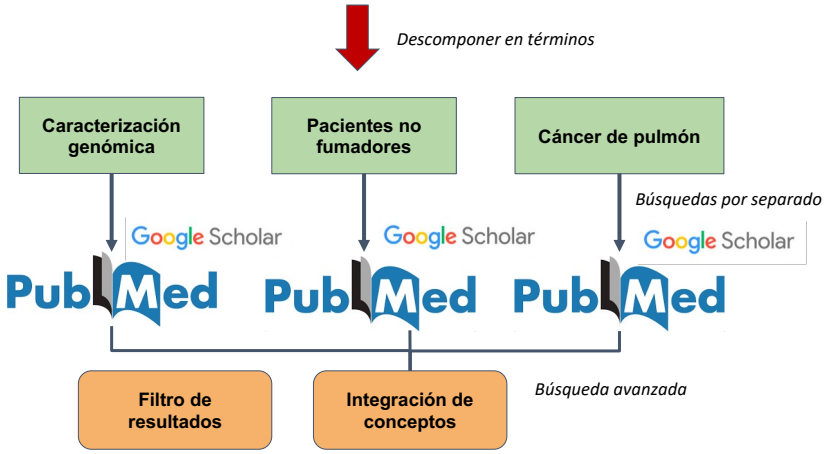
Descomponer en términos



29

Búsqueda de un tema

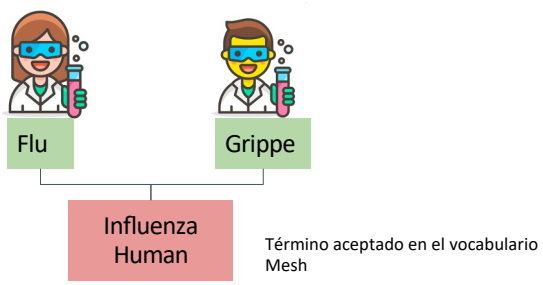
"Caracterización genómica de pacientes no fumadores con cáncer de pulmón"



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 qué analiza.



31

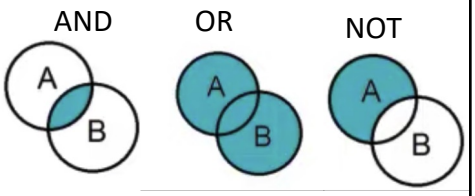
Técnicas de búsquedas

1. **Truncamiento, para encontrar variantes**
   model* = models, modeling, modelling

2. **Uso de comillas**
   "emergency medicine" **versus** emergency medicine

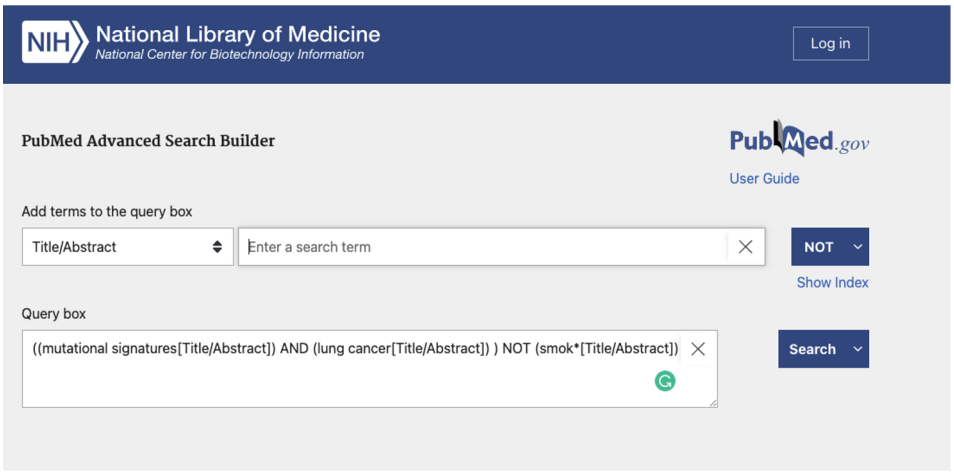
3. **Parentesis**
   Permite combinar conceptos (child* OR adolescen*)

4. **Operadores Booleanos**
   AND, OR, NOT
   child* AND "head injury"
   child* OR pedriat*
   child* NOT cancer



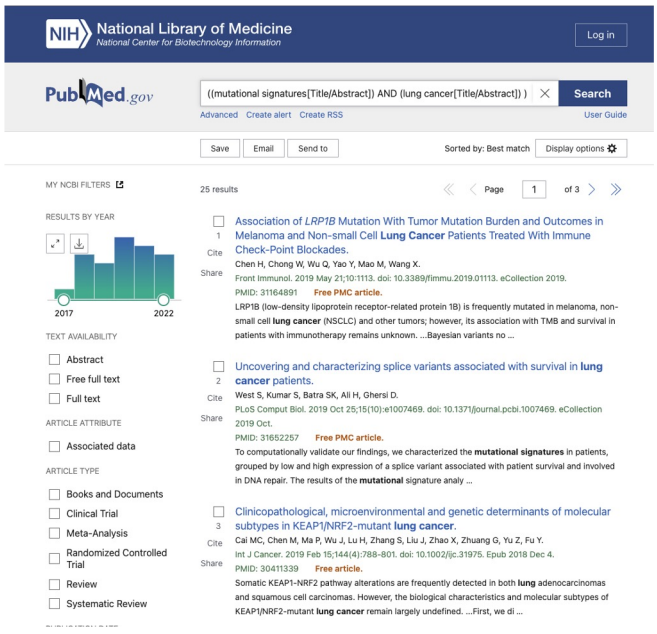
32

Busqueda avanzada



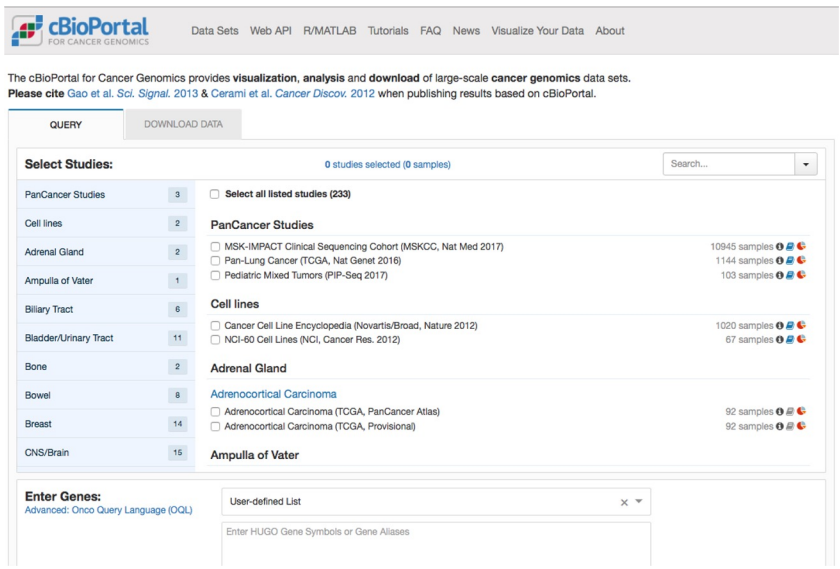
33

Búsqueda avanzada

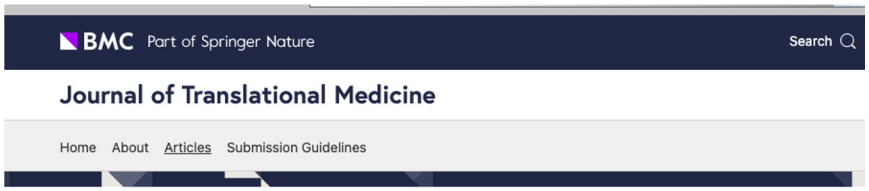


34

¿Qué es Cbiportal?



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 Armisén

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

581 Accesses | **2** Altmetric | Metrics

36