



GenBank



Práctica 1 - Bioinformática

Julia Guerrero Viu
Sergio Izquierdo Barranco
Ubaldo Ramón Júlvez

- 1. Presentación GenBank**
- 2. Consultas**
- 3. Descarga datos**

— ¿Qué es **GenBank**?

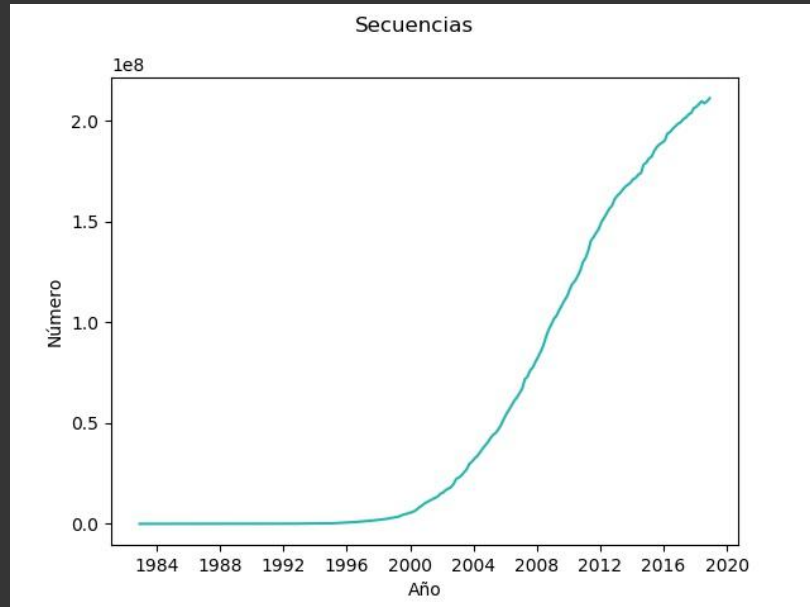
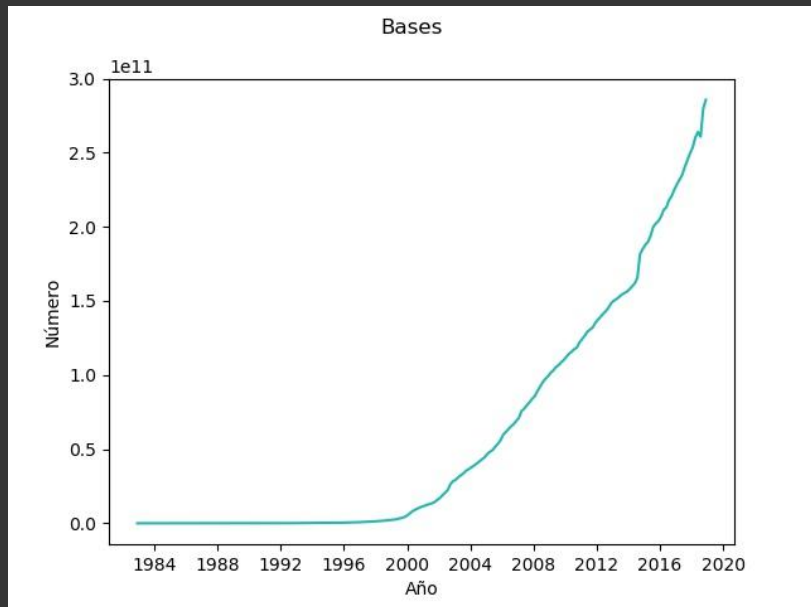
**Colección totalmente pública de todas las
secuencias de nucleótidos y su traducción
a proteínas**

Historia GenBank

- 1979 Walter Goad in Los Álamos National Laboratory
- 1982 GenBank (NIH)
- 1992 NCBI
- INSDC (International Nucleotide Sequence Database Collaboration)



—
285.688.542.186 bases **iy**
sigue creciendo!



Algunos datos

Cada **2 meses** sale una nueva versión de GenBank

Cada **18 meses** se duplica el número de secuencias de GenBank

Recibe secuencias diariamente de más de **100.000** organismos internacionales distintos

ACCEDER a GenBank

- Directamente BD (herramienta web)
- BLAST (herramienta de alineamiento de secuencias)
- Programas del NCBI
 - Servidor FTP (base de datos completa: <ftp://ftp.ncbi.nih.gov/genbank>)

<http://www.ncbi.nlm.nih.gov/genbank/>

The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar with a dropdown menu set to 'Nucleotide' and a 'Search' button. A secondary navigation bar includes links for 'GenBank', 'Submit', 'Genomes', 'WGS', 'Metagenomes', 'TPA', 'TSA', 'INSDC', and 'Other'. The main content area is titled 'GenBank Overview' and contains several sections: 'What is GenBank?' which explains the database's origin and purpose; 'Access to GenBank' which lists various search and retrieval methods; and 'GenBank Data Usage' which provides information on the database's design and access policies. On the right side, there's a 'GenBank Resources' section with links to 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'.

FUENTES (BBDD):

- Entrez Protein
- Entrez Nucleotide
 - Nucleotide (core)
 - EST and GSS
 - > se van a fusionar en 2019

<https://ncbiinsights.ncbi.nlm.nih.gov/2018/07/30/upcoming-changes-est-gss-databases/>

How To 

Protein 

Create alert Advanced

How To 

Nucleotide 

Create alert Advanced

Items: 1 to 20 of 191351407

<< First < Prev Page 1 of 9567571 Next > Last >>

 Found 202026066 nucleotide sequences. Nucleotide (191351407) EST (3) GSS (10674656)

☐ [Blumeria graminis f. sp. tritici genome assembly, chromosome: Bgt -09](#)

1. 19,721,624 bp linear DNA
Accession: LR026992.1 GI: 1564972449
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Blumeria graminis f. sp. tritici genome assembly, chromosome: Bgt -02](#)

2. 15,587,074 bp linear DNA
Accession: LR026985.1 GI: 1564948149
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Blumeria graminis f. sp. tritici genome assembly, chromosome: Bgt -07](#)

3. 16,229,997 bp linear DNA
Accession: LR026990.1 GI: 1564961704
[BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

Divisiones de GenBank

Submissions phases

Organismal Divisions

Database	Division	BLAST	Example
BCT	Bacterial sequences	nr, month	Human Phase 3
PRI	Primate sequences	nr, month	
ROD	Rodent sequences	nr, month	
MAM	Other mammalian sequences	nr, month	
VRT	Other vertebrate sequences	nr, month	
INV	Invertebrate sequences	nr, month	Drosophila, C. elegans Phase 3
PLN	Plant and Fungal sequences	nr, month	Arabidopsis Phase 3
VRL	Viral sequences	nr, month	
PHG	Phage sequences	nr, month	
RNA	Structural RNA sequences	nr, month	
SYN	Synthetic and chimeric sequences	nr, month	
UNA	Unannotated sequences	nr, month	

Functional Divisions

Database	Division	BLAST	Example
EST	Expressed Sequence Tags	dbest, month	All Organisms: Phase 0, 1, and 2
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	

- **Phase 0** sequences are single-pass reads of a single clone (not contigs usually).
- **Phase 1** sequences are unfinished, unordered, and contain gaps.
- **Phase 2** sequences are unfinished, ordered, and can contain one or more gaps.
- **Phase 3** sequences are high quality finished sequences that do not contain gaps.

Consultas en GenBank



- Consultas de frase
 1. Proteínas
 2. Genes
 3. Autores

kinase

Smith JR

Felis catus

"duchenne muscular dystrophy"

- Operadores Binarios
 - OR, AND, NOT
- Filtros
 - Especie
 - Tipo de molécula (DNA, RNA)
 - Fecha
 - etc.

contactin AND neurofascin

"Felis catus"[Organism]

biomol_genomic[PROP]

all[filter]

- Más filtros

Species
Animals
Archaea
Bacteria
Fungi
Plants
Protists
Viruses
Customize ...

Molecule types
genomic DNA/RNA (337)
mRNA (6)
Customize ...

Source databases
INSDC (GenBank) (49)
RefSeq (292)
Customize ...

Sequence Type
Nucleotide (343)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

[Clear all](#)

[Show additional filters](#)

Summary ▾ 20 per page ▾ S

See [LOC113060289 \(CO](#)
[contactin](#) reference seque

Items: 1 to 20 of 343

- ☐ [Columba livia breed Da](#)
1. [4,078,870 bp linear DN](#)
Accession: AKCR02000055
[BioProject](#) [BioSample](#)
[GenBank](#) [FASTA](#) [Gra](#)
- ☐ [Crassostrea gigas stra](#)
2. [whole genome shotgun](#)
450,376 bp linear DNA
Accession: NW_011937980
[Assembly](#) [BioProject](#)
[GenBank](#) [FASTA](#) [Gra](#)
- ☐ [Amazona aestiva isola](#)
3. [3,954,725 bp linear DN](#)
Accession: LMAW0100044
[Assembly](#) [BioProject](#)
[GenBank](#) [FASTA](#) [Gra](#)
- ☐ [Tupaia chinensis unpla](#)
4. [8,920,270 bp linear DN](#)

- Filtros personalizados

Species

Animals (30,005,804)

Plants (12,585,115)

Fungi (16,082,609)

Protists (4,486,205)

Bacteria (559,009,588)

Archaea (5,194,244)

Viruses (5,226,910)

Customize ...

Source database

PDB (418,770)

RefSeq (133,3

UniProtKB / S

Prot (559,189

Customize ...

Genetic compartment

Chloroplast (1,

Mitochondrion

Plasmid (2,094

Plastid (1,626,

Sequence length

Custom range

Molecular weight

Custom range

Release date

Summary ▾ 20 per page ▾ Sort by D

See the [results of this search](#) (2137

Items: 1 to 20 of 645724899

Species

☒ Animals

☒ Plants

☒ Fungi

☒ Protists

☒ Bacteria

☒ Archaea

☒ Viruses

☒ Felis catus

Add

Show

4. 227 aa protein

Accession: NP_001255561.1 Cl: 4

- **Ordenación de resultados**
 - Fecha
 - Organismo
 - Longitud secuencia
 - Taxonomía

Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 22

Found 229183 nucleotide sequences

☐ [Synthetic DNA entries](#)

1. [AIDS-antibody](#)

246 bp linear DNA

Accession: E01172.1 GI: 2169431

[Taxonomy](#)

<< First < Prev

26526) GSS (2657

[antibody against](#)

Sort by

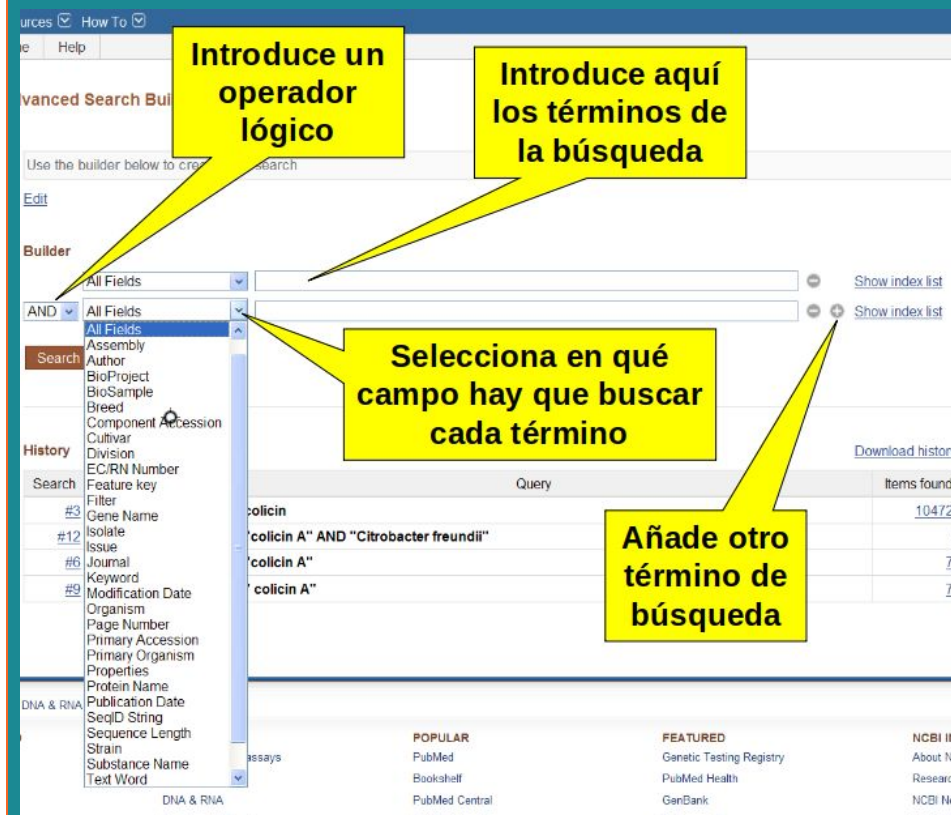
- ☒ Default order
- ☐ Accession
- ☐ Date Modified
- ☐ Date Released
- ☐ Organism Name
- ☐ Taxonomy ID
- ☐ Sequence Length

- Búsqueda AVANZADA



Nucleotide

- Campos
- Operadores lógicos
- Historial



Introduce un operador lógico

Introduce aquí los términos de la búsqueda

Selecciona en qué campo hay que buscar cada término

Añade otro término de búsqueda

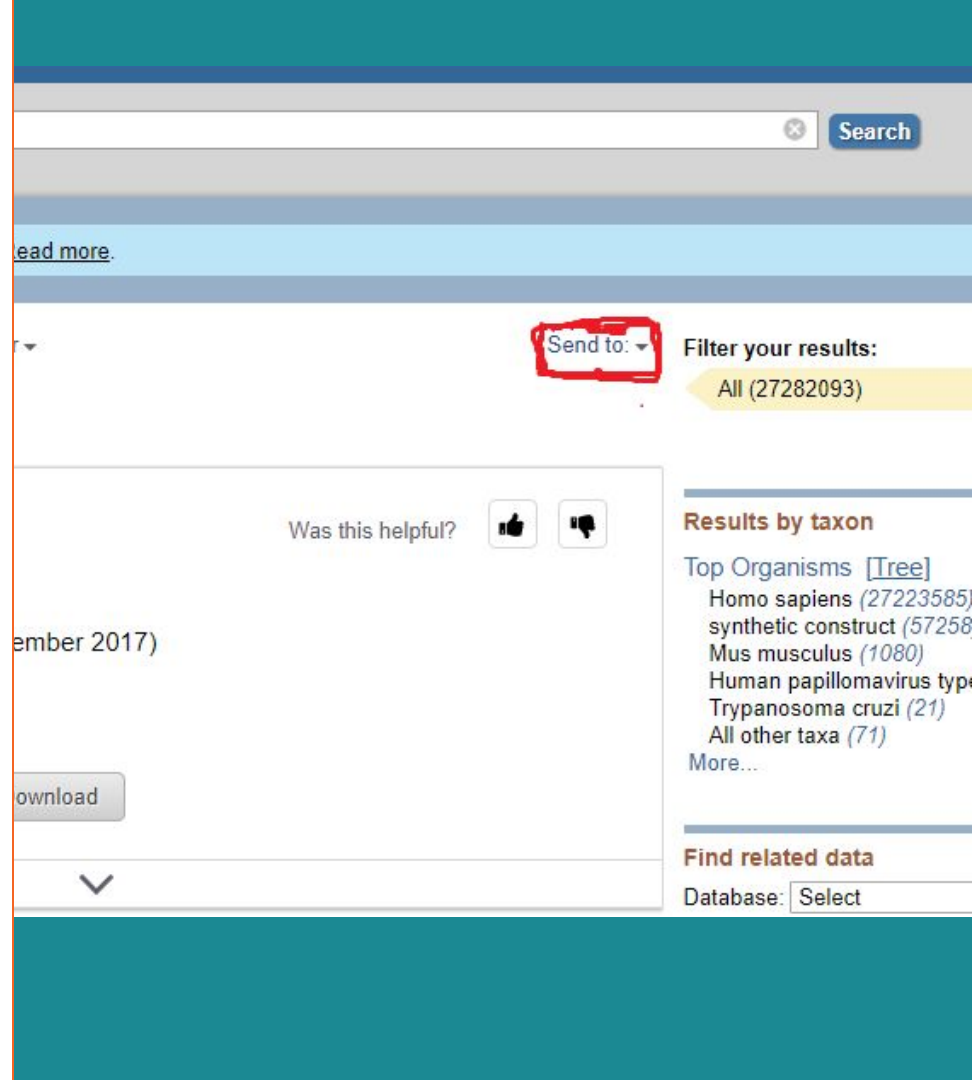
Query: colicin "colicin A" AND "Citrobacter freundii" "colicin A" "colicin A"

Search	Add to builder	Query	Items found	Time
#4	Add	Search genome	191351407	11:25:04
#1	Add	Search aids	226526	11:24:45

[Download history](#) [Clear history](#)

Consulta y modifica consultas anteriores

- Descarga de datos



Formato GenBank:

- Cabecera
- Referencias
- Features
- Secuencia

```
LOCUS      AP018036                327 bp    DNA        linear    BCT 10-JAN-2019
DEFINITION Mycobacterium tuberculosis DNA, complete genome, strain: HN-506.
ACCESSION  AP018036 REGION: 1646..1972
VERSION    AP018036.1
DBLINK     BioProject: PRJDB5607
           BioSample: SAMD00076258
KEYWORDS   .
SOURCE     Mycobacterium tuberculosis
ORGANISM   Mycobacterium tuberculosis
           Bacteria; Actinobacteria; Corynebacteriales; Mycobacteriaceae;
           Mycobacterium; Mycobacterium tuberculosis complex.
REFERENCE  1
AUTHORS    Wada,T., Hijikata,M., Maeda,S., Hang,N.T.L., Thuong,P.H.,
           Hoang,N.P., Hung,N.V. and Keicho,N.
TITLE      Complete Genome Sequences of Three Representative Mycobacterium
           tuberculosis Beijing Family Strains Belonging to Distinct Genotype
           Clusters in Hanoi, Vietnam, during 2007 to 2009
JOURNAL    Genome Announc 5 (27), e00510-17 (2017)
PUBMED     28684565
REMARK     DOI:10.1128/genomeA.00510-17
           Publication Status: Online-Only
REFERENCE  2 (bases 1 to 327)
AUTHORS    Wada,T., Hijikata,M., Maeda,S., Hang,N.T., Thuong,P.H., Hoang,N.P.,
           Hung,N.V. and Keicho,N.
TITLE      Direct Submission
JOURNAL    Submitted (06-APR-2017) Contact:Takayuki Wada Institute of Tropical
           Medicine, Department of International Health; Sakamoto 1-12-4,
           Nagasaki, Nagasaki 852-8523, Japan
COMMENT    This project was organized by Department of Pathophysiology and
           Host Defense, The Research Institute of Tuberculosis, Japan
           Anti-Tuberculosis Association, Tokyo, Japan, collaborating with
           Hanoi Lung Hospital, Hanoi, Vietnam.
           This project was supported by J-GRID, granted by Japan Agency for
           Medical Research and Development (AMED) and Ministry of Education,
           Culture, Sports, Science and Technology (MEXT).
           Annotated using prokka 1.11 from http://www.vicbioinformatics.com.
           Annotated at D-FAST https://dfast.nig.ac.jp
```

Formato GenBank:

- Cabecera
- Referencias
- Features
- Secuencia

```
FEATURES             Location/Qualifiers
     source            1..327
                        /organism="Mycobacterium tuberculosis"
                        /mol_type="genomic DNA"
                        /strain="HN-506"
                        /host="Homo sapiens"
                        /db_xref="taxon:1773"
                        /country="Viet Nam:Hanoi"
                        /collection_date="2009-03-24"
     gene              1..327
                        /locus_tag="HN506_00002"
     CDS               1..327
                        /locus_tag="HN506_00002"
                        /inference="ab initio prediction:Prodigal:2.6"
                        /inference="similar to AA sequence:UniProtKB:P59801"
                        /codon_start=1
                        /transl_table=11
                        /product="insertion element IS6110 uncharacterized 12.0
                        kDa protein"
                        /protein_id="BAX47168.1"
                        /translation="MSGGSSRRYPPELRERAVRMVAEIRGQHDSEWAAISEVARLLGV
                        GCAETVRKWVRQAQVDAGARPGTTTEESAE LKRLRRD NAELRRANAILKTASAFFAAE
                        LDRPAR"

ORIGIN
      1 atgtcagggtg gttcatcgag gaggtaccgc cgggagctgc gtgagcgggc ggtgcggatg
     61 gtcgcagaga tccgcggtca gcacgattcg gagtgggcag cgatcagtga ggtcgcccg
    121 ctacttggtg ttgctgcgc ggagacggtg cgtaagtggg tgcgccaggc gcaggtcgat
    181 gccggcgcac ggcccgggac caccgaccgaa gaatccgctg agctgaagcg cttgcggcgg
    241 gacaacgcgc aattgcgaag ggccaacgcg attttaaaga ccgcgtcggc tttcttcgcg
    301 gccgagctcg accggccagc acgctaa

//
```

Formato Fasta:

- Cabecera
- Secuencia

```
>AP018036.1 Mycobacterium tuberculosis DNA, complete genome, strain: HN-506
TTGACCGATGACCCCGGTT CAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACCTAACGGCGACC
CTAAGGTTGACGACGGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTG
GCTCAATCTCGTCCAGCCATTGACCATCGTCGAGGGGTTTGTCTGTTATCCGTCGCCAGCAGCTTTGTC
CAAAACGAAATCGAGCGCCATCTGCGGGCCCCGATTACCGACGCTCTCAGCCGCCGACTCGGACATCAGA
TCCAACTCGGGGTCGCGATCGCTCCGCCGGCGACCGAGAAGCCGACGACATACCGTGCCGCTTCCGA
AAATCCTGCTACCATCGCCAGACACCAACCGACAACGACGAGATTGATGACAGCGCTGCGGCACGG
GGCGATAACCAGCACAGTTGGCCAAGTTACTTCACCGAGCGCCGACAATACCGATTCCGCTACCGCTG
GCGTAACCAAGCCTTAACCGTCGCTACACCTTTGATACGTTCTGTTATCGGCGCCTCCAACCGGTTGCGCA
CGCCGCCGCTTGGCGATCGCAGAAGCACCGCCCGCGCTTACAACCCCTGTTTCATCTGGGGCGAGTCC
GGTCTCGGCAAGACACCTGCTACACGCGGCAGGCAACTATGCCAACGGTTGTTCCGGGAATGCGGG
TCAAATATGTCTCCACCGAGGAATTCACCAACGACTTCATTAACCTGCTCCGCGATGACCGCAAGGTCGC
ATTCAAACGCAGCTACCGCGACGTAGACGTGCTGTTGGTCGACGACATCCAATTCATTGAAGGCAAAGAG
GGTATTCAAGAGGAGTTCTTCCACACCTTCAACACCTTGACAAATGCCAACAAGCAAATCGTCATCTCAT
CTGACCGCCACCCAAGCAGCTCGCCACCTCGAGGACCGGCTGAGAACCCGCTTTGAGTGGGGGTGAT
CACTGACGTACAACACCCGAGCTGGAGACCCGATCGCCATCTTGCGCAAGAAAGCACAGATGGAACGG
CTCGGGTCCCCGACGATGTCCTCGAACTCATCGCCAGCAGTATCGAACGCAATATCCGTGAACTCGAGG
GCGCGCTGATCCGGGTACCGCGTTTCGCTCATTGAACAAAACACCAATCGACAAAGCGCTGGCCGAGAT
TGTGCTTCGCGATCTGATCGCCGACGCCAACACCATGCAAAATCAGCGCGCGCAGATCATGGCTGCCACC
GCCGAATACTTCGACACTACCGTCGAAGAGCTTCGCGGGCCCGGAAGACCCGAGCACTGGCCAGTCAC
GACGATTGCGATGTACCTGTGTCGTGAGCTACCGATCTTTGTTGCCAAAATCGGCCAAGCGTTCCG
CCGTGATCACACAACCGTCATGTACGCCAACGCAAGATCCTGTCCGAGATGGCCGAGCGCCGTGAGGTC
TTTGATCAGCTCAAAGAACTCACCCTCGCATCCGTAGCGCTCCAAGCGCTAGCACGGCGTGTCTTCC
GACAACTTCTTAAAAAACTTCTCTCTCCAGGTCACACAGTCACAGAGATTTGAACCGCCCCGGCAT
GTCCGGAGACTCCAGTTCTTGAAAGGATGGGGTCATGTACGGTGGTTTCATCGAGGAGGTACCCGCCGGA
GCTGCGTGAGCGGGCGGTGCGGATGGTCGAGAGATCCGCGGTGAGCAGGATTCGGAGTGGGCGAGCGATC
AGTGAGGTGCGCCGTCTACTTGTGTTGGTCGCGGAGACGGTGCGTAAGTGGGTGCGCCAGGCGCAGG
```

Limitar el número de resultados a descargar.

Items: 1 to 20 of 27282093

Selected: 3

<< First < Prev Page 1

 Filters activated: Homo sapiens. [Clear all](#)

 Found 28475230 nucleotide sequences. Nucleotide (27282093) GSS ([119313](#))

☒ [Homo sapiens](#) ATP synthase F1 subunit alpha (ATP5F1A), RefSeqGene

1. 35,986 bp linear DNA

Accession: NG_041769.2 GI: 1573294352

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

☒ [Homo sapiens](#) aspartoacylase (ASPA), RefSeqGene on chromosome 1

2. 36,760 bp linear DNA

Accession: NG_008399.2 GI: 1573293807

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

☒ [Homo sapiens](#) aldehyde dehydrogenase 2 family member (ALDH2), RefSeqGene

3. [12](#)

57,646 bp linear DNA

Accession: NG_012250.2 GI: 1573289586

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

☐ [Homo sapiens](#) adducin 2 (ADD2), RefSeqGene on chromosome 2

4. 118,460 bp linear DNA

Accession: NG_029481.2 GI: 1573279893

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