# **GenBank** Resumen y búsquedas básicas

Pablo Tellería Cebrián 30260 — Bioinformática Febrero de 2018

# Qué es GenBank

GenBank es una base de datos de secuencias genéticas de acceso abierto administrada por el **NCBI** (Centro Nacional para la Información sobre Biotecnologías), a su vez parte del **NHI** (Instituto Nacional de Salud) de los Estados Unidos de América. Como tal, proporciona acceso a la colección anotada de todas las secuencias de ADN de dominio público. GenBank es una de las tres integrantes del **INSDC** (Colaboración Internacional de Bases de Datos de Secuencias de Nucleótidos), junto con el **DDBJ** (DNA DataBank de Japón) y el **ENA** (Archivo Europeo de Nucleótidos). Estas tres organizaciones intercambian datos de forma diaria.

En funcionamiento desde 1982, en la actualidad GenBank recibe secuencias obtenidas en laboratorios de todo el mundo, procedentes de más de 100.000 organismos vivos distintos. En los últimos años ha tenido un crecimiento exponencial, doblando su tamaño aproximadamente cada 18 meses (nótese la escala logarítmica del tamaño de la base de datos):



GenBank hace público su contenido a través de lanzamientos (*releases*), que tienen lugar cada dos meses. A la fecha de la redacción de este resumen, el lanzamiento más reciente es el número **224**, con un tamaño de **256,63 Gbp**.

# Búsquedas en GenBank

GenBank proporciona tres bases de datos sobre las que realizar consultas: *Nucleotide*, *Genome Survey Sequence* (GSS) y *Expressed Sequence Tag* (EST). Si bien todas ellas contienen secuencias de ácidos nucleicos, GSS contiene secuencias cortas sin caracterizar y EST secuencias de ADNc (ADN complementario).

Una búsqueda en cualquiera de las tres bases de datos proporciona enlaces a los resultados en las otras dos. Salvo que se sepa de antemano que se buscan secuencias típicas de GSS o EST, los mejores resultados se obtendrán en *Nucleotide*, a partir de consultas de texto y el uso de los filtros.

Para las búsquedas de texto, se pueden emplear los nombres y símbolos de genes y proteínas directamente. Además, se puede buscar el nombre de autor o persona que envió las muestras o los identificadores propios de la base de datos, como los números gi (*gi numbers*) o de accesión (*accession*).

Si se desean consultas más complejas están disponibles operadores típicos en buscadores de texto, como son las comillas dobles para buscar una frase exacta, o los operadores AND, OR y NOT.

Para facilitar la construcción de estas búsquedas complejas se dispone de un constructor avanzado de consultas, en la que se pueden añadir de manera visual distintos criterios (autor, organismo, nombre del gen, fecha, especie, raza... así hasta 32) y relacionarlos entre sí mediante los operadores AND, OR y NOT:

Edit				Cle
Builder				
	All Fields	0	Show index list	st
	Accession		A Show index lit	
				<u>st</u>
_	Author			
Search	BioProject			
	BioSample			
	Breed Component Accession			
	Cultivar			
History	Division		Download history C	lear hist
	EC/RN Number			
Search	Feature key		Query Items found	Time
<u>#1</u>	Gene Name	D4	75359	07:59:
	Isolate			
	Issue			
	Journal			

#### Nucleotide Advanced Search Builder

Cabe destacar que cualquiera de los tres métodos de búsqueda que empleemos (cadena de búsqueda, constructor avanzado de consultas o filtrado de resultados) **quedará finalmente reducido a una cadena de búsqueda**. Para el usuario avanzado, con un conocimiento profundo de los operadores y los dominios de búsqueda, es posible que la forma más eficaz de realizar consultas sea directamente mediante dichas cadenas. Para un usuario menos experimentado, sin duda los mejores resultados se obtendrán a partir del filtrado y refinamiento sucesivo.

En cualquiera de los casos, la cadena de búsqueda obtenida al final será especialmente útil a la hora de referenciar o compartir una búsqueda específica.

#### Filtrado y ordenación de los resultados

Una vez obtenidos los resultados de búsqueda, podemos filtrarlos conforme a varios criterios para afinar mejor nuestra consulta. Los listaremos a continuación, para ilustrar su uso mediante un sencillo ejemplo más adelante:

- Por especie (species): GenBank nos muestra por defecto los reinos biológicos (animales, plantas, hongos, protistas, arqueas, bacterias y virus) sobre los que existen resultados para nuestra consulta. Además de esta selección rápida, podemos personalizar estos enlaces para mostrar cualquier otra familia, género o especie que nos interese. De manera adicional, en la parte derecha de la pantalla tenemos un listado de los principales organismos con resultados para la búsqueda actual (*Top Organisms*), que podemos emplear también como filtro.
- Por **tipo de molécula** (*molecule type*): nos permite seleccionar entre ADN/ARN genómico y otras variantes de RNA no genómico (mensajero, complementario, no codificante, ribosómico, de transferencia o transcrito)
- Por **tipo de enzima** (*enzyme type*): cuando proceda, podremos seleccionar entre hidrolasas, isomerasas, ligasas...
- Por **base de datos de origen** (*source databases*): podemos elegir entre varios orígenes de datos, como el propio GenBank, el DDBJ de Japón, etc.
- Por compartimento genético (genetic compartments): muy útil cuando estemos buscando secuencias de origen extracromosómico (halladas en cloroplastos, mitocondrias, plásmidos o plastos)

- Por **rango de longitud** (*sequence length*) de las secuencias resultado, expresado en pares de bases (bp).
- Por fecha de lanzamiento (release date) o de revisión (revision date)
- Por **campos de búsqueda** (*search fields*): podemos elegir sobre qué campos queremos que se efectúe la búsqueda textual que hemos introducido.

Nucleotide	Nucleotide 🗘 Arginine	Search			
	Create alert Advanced	Hel			
Species Animals (129,183)	Summary - 20 per page - Sort by Default order - Send to: - Filters: Manage Filters				
Plants (38,210) Fungi (37,800) Protists (14,015) Bacteria (1,420,281) Archaea (6,750) Viruses (10,853) Customize Molecule types genomic DNA/RNA (1,538,444	Items: 1 to 20 of 1660294         << First < Prev Page 1 of 83015 Next > Last >>         Image: Second Secon	Results by taxon Top Organisms [Tree] Escherichia coli (180653) Salmonella enterica (162636) Staphylococcus aureus (102399) Streptococcus pneumoniae (94046) Mycobacterium tuberculosis (74937) All other taxa (1045623) More			
mRNA (113,555) rRNA (5) Customize Source databases INSDC (GenBank) (942,652) RefSeq (716,868) Customize	GenBank       FASTA       Graphics         Homo sapiens chromosome 1, GRCh38.p7 Primary Assembly         2.       248,956,422 bp linear DNA         Accession: NC_00001.11 GI: 568815597         Assembly       BioProject         Protein       PubMed         Taxonomy         GenBank       FASTA         Graphics	Find related data Database: Select			
Genetic compartments Chloroplast (761) Mitochondrion (275) Plasmid (3,085) Plastid (895)	Cavenderia deminutiva strain MexM19A arginine (argC) gene, complete cds 3. 2,475 bp linear DNA Accession: KX539444.1 Gi: 1057487767 Protein Taxonomy GenBank FASTA Graphics	Search details			
Sequence length Custom range Release date Custom range	<ul> <li>Dictyostelium polycephalum strain MY1-1 arginine (argC) gene, complete cds</li> <li>2,637 bp linear DNA Accession: KX539446.1 Gl: 1057487771 Protein Taxonomy</li> </ul>	Search See more. Recent activity			
Revision date Custom range <u>Clear all</u> Show additional filters	GenBank       FASTA       Graphics         Dictyostelium polycarpum strain OhioWILDS arginine (argC) gene, complete cds         5.       2,346 bp linear DNA         Accession: KX539445.1       Gl: 1057487769         Protein       Taxonomy         GenBank       FASTA         Constant       Constant	CD4 (75359)			
Custom range Revision date Custom range <u>Clear all</u> Show additional filters	Protein Taxonomy GenBank FASTA Graphics Dictyostelium polycarpum strain OhioWILDS arginine (argC) gene, complete cds 5. 2,346 bp linear DNA Accession: KX539445.1 GI: 1057487769 Protein Taxonomy GenBank FASTA Graphics	Recent activity         Turn Off         Cle           Q         Arginine (1660294)         Nuclear           Q         CD4 (75359)         Nuclear           See mode         See mode			

Por defecto se muestran 20 resultados por página, en formato resumen. Podemos seleccionar distintos tamaños de página (entre 5 y 200), y distintos criterios de ordenación (por accesión, fechas de lanzamiento y modificación, nombre del organismo, identificador o longitud de la secuencia). Además podemos elegir distintos formatos para el resumen mostrado en el listado de resultados, de acuerdo con varios estándares (GenPept, FASTA, ASN...).

# Ejemplo de búsqueda

Supongamos que queremos encontrar información sobre el gen **CD4** (responsable de la expresión del antígeno CD4, fundamental para el funcionamiento del sistema inmunitario) obtenido en muestras de perros (*canis lupus familiaris*).

Podemos comenzar con una búsqueda del texto 'CD4' en la base de datos por defecto (*Nucleotide*), lo que nos devuelve un listado con más de 75000 resultados:

Nucleotide	Nucleotide CD4 Create alert Advanced	Search	Help
Species Animals (11,749)	Summary - 20 per page - Sort by Default order - Send to: -	Filters: Manage Filters	
Plants (64) Fungi (19) Protists (133)	See <u>CD4 CD4 molecule</u> in the Gene database cd4 reference sequences <u>Genomic (1)</u> <u>Transcript (6)</u> <u>Protein (6)</u>	Results by taxon Top Organisms [Tree] Human immunodeficiency virus 1 (52413) Homo sapiens (7385) Listeria moncoytogenes (3123) Simian immunodeficiency virus (1778) Mus musculus (1561)	
Archaea (1) Viruses (56,102) Customize	Items: 1 to 20 of 75359 << First < Prev Page 1 of 3768 Next > Last >>		
Molecule types genomic DNA/RNA (64,991) mRNA (9,629) Customize	<ul> <li>Found 279984 nucleotide sequences. Nucleotide (75359) EST (<u>24679</u>) GSS (<u>179946</u>)</li> <li><u>G.gallus CD4 mRNA</u></li> <li>1,989 bp linear mRNA</li> </ul>	All other taxa (9099) More	
Source databases INSDC (GenBank) (71,098) RefSec (4 258)	Accession: Y12012.1 GI: 4539031 Protein PubMed Taxonomy GenBank FASTA Graphics	Find related data Database: Select	•
Customize Genetic	Actus vociferans CD4 (CD4) mRNA, CD4-1 allele, complete cds     1,374 bp linear mRNA	Find Items	
Chloroplast (2) Mitochondrion (44) Plasmid (27)	Accession: NK40/2344.1 Taxonomy Protein PubMed Taxonomy GenBank FASTA Graphics PopSet	Search details CD4[All Fields]	
Sequence length Custom range	<ul> <li>Sylvilagus floridanus CD4 (CD4) mRNA, complete cds</li> <li>1,371 bp linear mRNA Accession: KU845555.1 Gl: 1009023990</li> </ul>	Search	/
Release date Custom range	Protein PubMed Taxonomy GenBank FASTA Graphics	Recent activity	See more
Clear all	Capra hircus CD4 (CD4) gene, complete cds  1,555 bp linear DNA Accession: EU913093.1 GI: 196212987 Protein_Publed_Taxonomy	Q CD4 (75359)	Turn Off Clear Nucleotide
Show additional filters	GenBank FASTA Graphics	Q Arginine (1660294)	Nucleotide
	Macaca mulatta isolate rhCD4-v5 CD4 (CD4) mRNA, partial cds     1,325 bp linear mRNA     Accession: MF769804.1 GI: 1243914460     Protein Taxonomy		See more

Como solo estamos interesados en una especie en concreto (perros), emplearemos el filtrado por especie. Entre las opciones por defecto no se encuentra algo tan específico, así que dentro de la sección *Species* de la columna izquierda haremos click en enlace *Customize...* 

GenBank FASTA Graphics PopSet

Species Animals (11,749) Plants (64) Fungi (19) Protists (133) Bacteria (3,731) Archaea (1) Viruses (56,102) Customize ...



Buscaremos 'Canis lupus familiaris' y lo añadiremos a nuestros enlaces de búsqueda. Una vez hecho esto lo seleccionaremos, lo que reducirá nuestros resultados a tan solo 15.

Estamos interesados en un gen, de modo que podemos afinar aún más nuestra búsqueda filtrando por tipo de molécula y seleccionando 'ADN/ARN genómico'. Hecho esto, habremos reducido los resultados a 4.

Como hemos comentado antes, todos los métodos de búsqueda acaban reducidos a una cadena y diversos operadores. Como puede verse en la caja de texto de la columna derecha, el mismo resultado podría haberse obtenido buscando directamente la siguiente cadena:

```
CD4[All Fields] AND ("Canis lupus familiaris"[Organism] AND
biomol_genomic[PROP])
```

Esta cadena puede introducirse directamente en el buscador de texto, o construirla con el interfaz avanzado. Examinar estas cadenas según se hace uso de los filtros es una buena manera de familiarizarse con los operadores y dominios de búsqueda.

### Consumo de los resultados

Además de consultar por pantalla los detalles de cada resultado de búsqueda (lo que no tiene mucho sentido para secuencias del orden de millones de pares de bases), desde el listado de resultados podemos seleccionar aquellos que nos interesen y enviarlos a un destino de datos:

- En primer lugar, podemos elegir qué información queremos seleccionar: el registro completo (*Complete Record*), solo las secuencias codificantes (*Coding Sequences*) o las características de los genes (*Gene Features*).
- Una vez hecho esto, tenemos cuatro destinos posibles:
  - Fichero (File), que se descagará nuestro navegador
  - Copiar en el **portapapeles** (*Clipboard*), para pegar posteriormente en alguna otra aplicación

- Añadir a nuestras **colecciones** (*Collections*). Se requerirá que estemos registrados en la plataforma para emplear esta característica.
- Herramienta de análisis (Analysis Tools), donde podremos enviar las secuencias seleccionadas a BLAST (*Basic Local Alignment Search Tool*), la herramienta de alineamiento de secuencias del NCBI.

Siguiendo nuestro ejemplo anterior, si quisiéramos almacenar en disco las secuencias completas obtenidas al final de nuestra búsqueda, haríamos lo siguiente:

	Send to: - Filter your results:
	Complete Record     Coding Sequences     Gene Features
_	Choose Destination
	File     Clipboard     Collections     Analysis Tool
<u>hol</u>	Download 4 items.
	Format ct
	Sort by Default order
	Show GI
р	Create File
	familiaris"[Orga biomol_genomic[P

- Seleccionar aquellos resultados que queramos exportar. Si no seleccionamos ninguno, se entenderá que queremos exportarlos todos.
- Hacer click en Send To:, elegir Complete Record y File.
- Elegir el formato GenBank (full).
- Hacer click en Create File.

Como resultado, obtendremos un fichero de algo más de 100 Mb con las secuencias completas en el formato seleccionado:

				sequence.	gb			
<> seque	nce.gb >							
209079	/691821	tatttatatt	tttcacaaag	gagttaacgc	cgtgattcgg	cttttccatt	tttttttttt	
209080	7691881	ccttccttct	ccgccttccc	ttcccttcct	tccttccttt	cccttccttc	ccttcccttc	
209081	7691941	ccttcctctc	cttcctttcc	tcccttccct	tcttcccctt	cccttccttc	cttcccttcc	
209082	7692001	ctttttccct	tccttccttt	ctttcccttc	ccttccttcc	ttccctcctt	cctttcctcc	
209083	7692061	ttccctccct	tctttcttcc	gttcctttcc	ttccttcctt	ccttcctccc	ttccctccct	
209084	7692121	tccttccttc	tgtccctcct	tttcctcctt	ccttcccttc	cttccttcct	tccttccttc	CONST.
209085	7692181	cttccttcct	tccttccttc	cttccttcct	tccttccttc	tgtccttcct	ttccttcttt	
209086	7692241	ccttccttcc	ctccttcctt	tctttccctc	ctccccatcc	ttccttccct	ttctttcctt	
209087	7692301	ccttcctccc	cttccttcct	tctctatcca	tcgaaagtcc	tattgatcct	acctgacgct	
209088	7692361	tcagggtcac	ttttcctccc	cacccacccg	ggtgtttgtc	ctgtcatttt	gccttgcgtc	
209089	7692421	atccgcgcct	tttacaccgc	atactcctca	ggacaaggac	tttgctcttt	gttttggcat	
209090	7692481	aattctataa	ccccgggtac	actcctggcg	ctataaaaat	gttaaagaag	aaaaagaaaa	
209091	7692541	aaagagggga	aaaaaaaaaa	aaaaaggagg	caatccccca	ggaggccagc	ccggcgggcg	115555
209092	7692601	gcgggtgctg	tccagcagca	tctccatctc	cggccgcgcc	tggcggctcg	gggcccgagt	
209093	7692661	cctgctctgc	ggccgcgcgc	ggggacctgc	ggggacctgc	ggggaccccc	ggggacagcc	
209094	7692721	cgcgagctcc	gaggctggcg	gcttctgctg	ggacttcgcc	agaggaacca	gtgagtgagt	193669
209095	7692781	gagacgtggc	aaggacctgg	aggaggagga	ggggtgaggc	agggcagccg	cggggctcta	
209096	7692841	agaggggttt	cctgcgtgcg	cctgtggccc	ccgggccgtg	gcctcgggct	ctggatcgtc	
209097	7692901	cccaaagggc	aggcgcgtgg	gcacgcgggc	gggatgcgcc	tgaggcccta	tgcgccagcc	
209098	7692961	gatctggacc	tactaaccgg	agagcatttg	tgtttccagc	ctggctttgg	gagcgtctgc	
209099	7693021	aaagccatca	aatgggttgc	agccctgcag	acctggagcg	attttatcat	aggcccgaga	
209100	7693081	acacatcatc	cctgtgttat	tccttttttg	ggttcttgag	gcaatggtgt	agaaacagca	
209101	7693141	cacttaaaaa	tatgacgctt	tgttggcatg	tgtcacggtg	ctgattttaa	ttatacttgc	
209102	7693201	agtacagagg	agtccgtgct	atccgtggtg	gacttacata	tatttaaagt	ttttatttg	
209103	7693261	aagtgattgt	acagtcacca	gaagttgtaa	aggtggcaca	gagaggcccc	cccccctct	
209104	7693321	gtctttcacc	cagtttcctc	cagtggttac	attgggcatg	attttcatat	aagatcaaaa	
209105	7693381	ctcagacttt	gatgttggta	ctatatgtgt	gtacaaaacg	tatgtatcgt	ttccacgtca	
209106	7693441	ttcgatcaca	cgcgaaggtt	tgtgcagcca	ccagcacaac	caaggtacgg	aacttgcatc	
209107	7693501	acttgaaaga	tcccctgct	gggattcctt	tacagccaca	cttgcctcct	tcccccttcc	
209108	7693561	cctgacaaca	gatgtttccc	tgctctatat	tttaattgtt	tttgcaaata	ttatgtaagc	
209109	7693621	agaatcacac	tgtgatctct	ttgagactgg	titttgtctc	titttgactg	gttatgctga	
209110	7693681	atgcataaca	ttctagaggt	ctagagatac	gttgctgtat	ccaaagttgg	gtccttttta	
209111	7693741	tcactgaata	gtattccagt	tgcttcacca	catttagtgg	tagcgttttt	cttcttcccc	
209112	7693801	ttccccttct	cettettete	ttettettet	ttettette	ttettettet	tettettett	
209113	7693861	cttcttcttc	ttettettet	tettettett	cttcttcttc	ttettettet	ttttcttctt	
209114	7693921	cttcttcttc	ttettettet	tettettett	cttttcttct	tettettett	cttcttcttc	
209115	7693981	ttettttett	cttcttcttc	ttettettet	tttcttcttc	ttettettet	tettttette	

# Bibliografía

- Sitio web de GenBank (<u>https://www.ncbi.nlm.nih.gov/genbank/</u>)
- Elaboración propia