GenBank

Resumen y búsquedas básicas

Pablo Tellería Cebrián

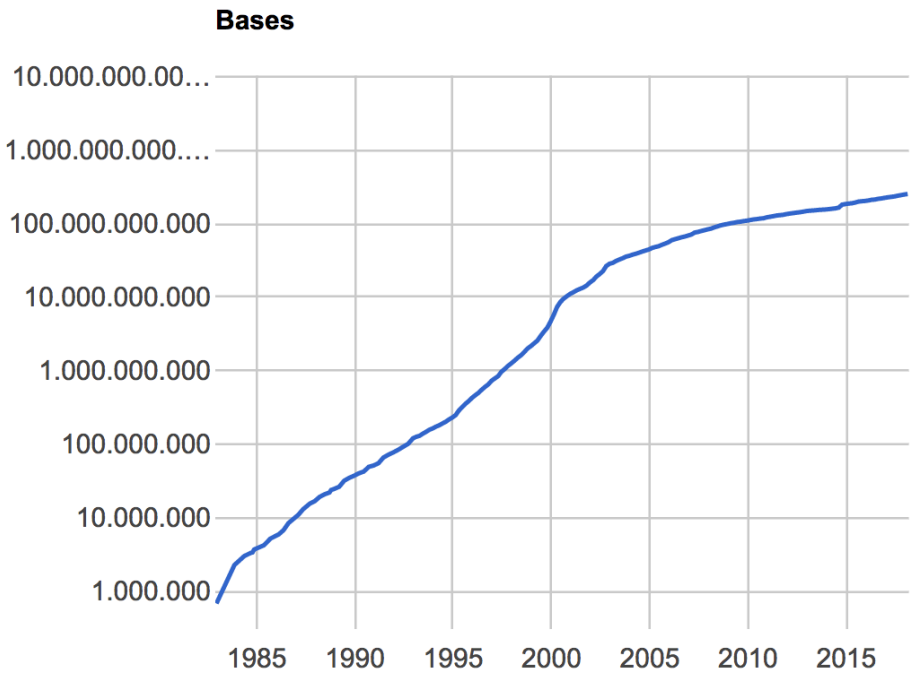
30260 — Bioinformática

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

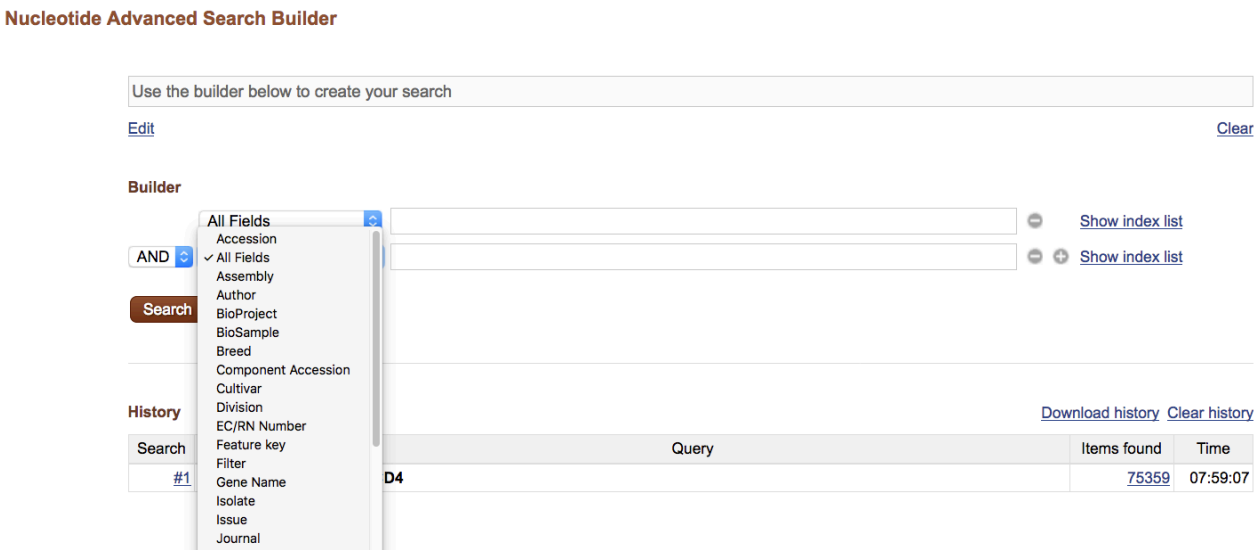
GenBank proporciona tres bases de datos sobre las que realizar consultas: *Nucleotide*, *Genome Survey Sequence* (GSS) y *Expressed Sequence Tag* (EST). Si bien todas ellas contienen secuencias de ácidos nucleicos, GSS contiene secuencias cortas sin caracterizar y EST secuencias de ADNc (ADN complementario).

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:



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.

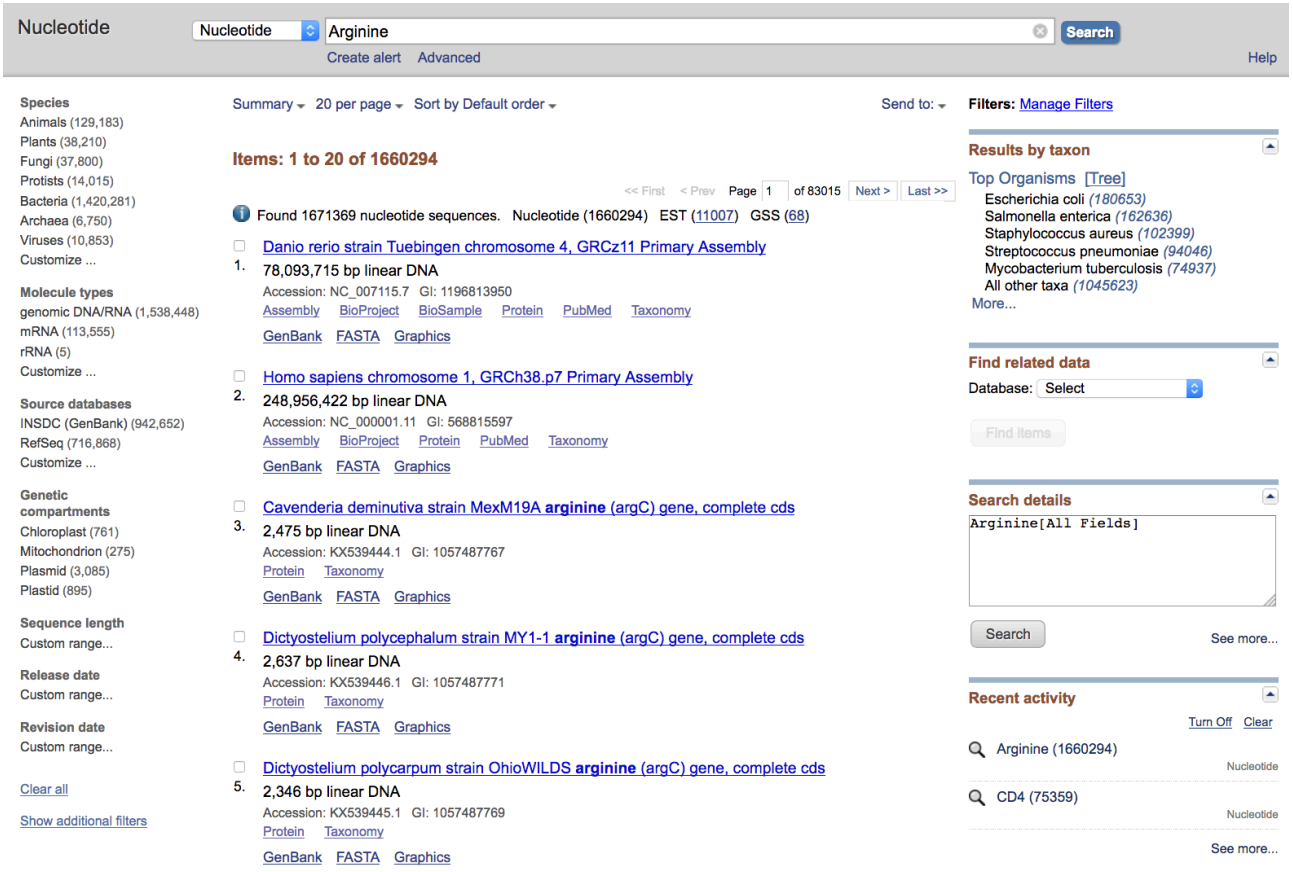
En cualquiera de los casos, la cadena de búsqueda obtenida al final será especialmente útil a la hora de referenciar o compartir una búsqueda específica.

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.

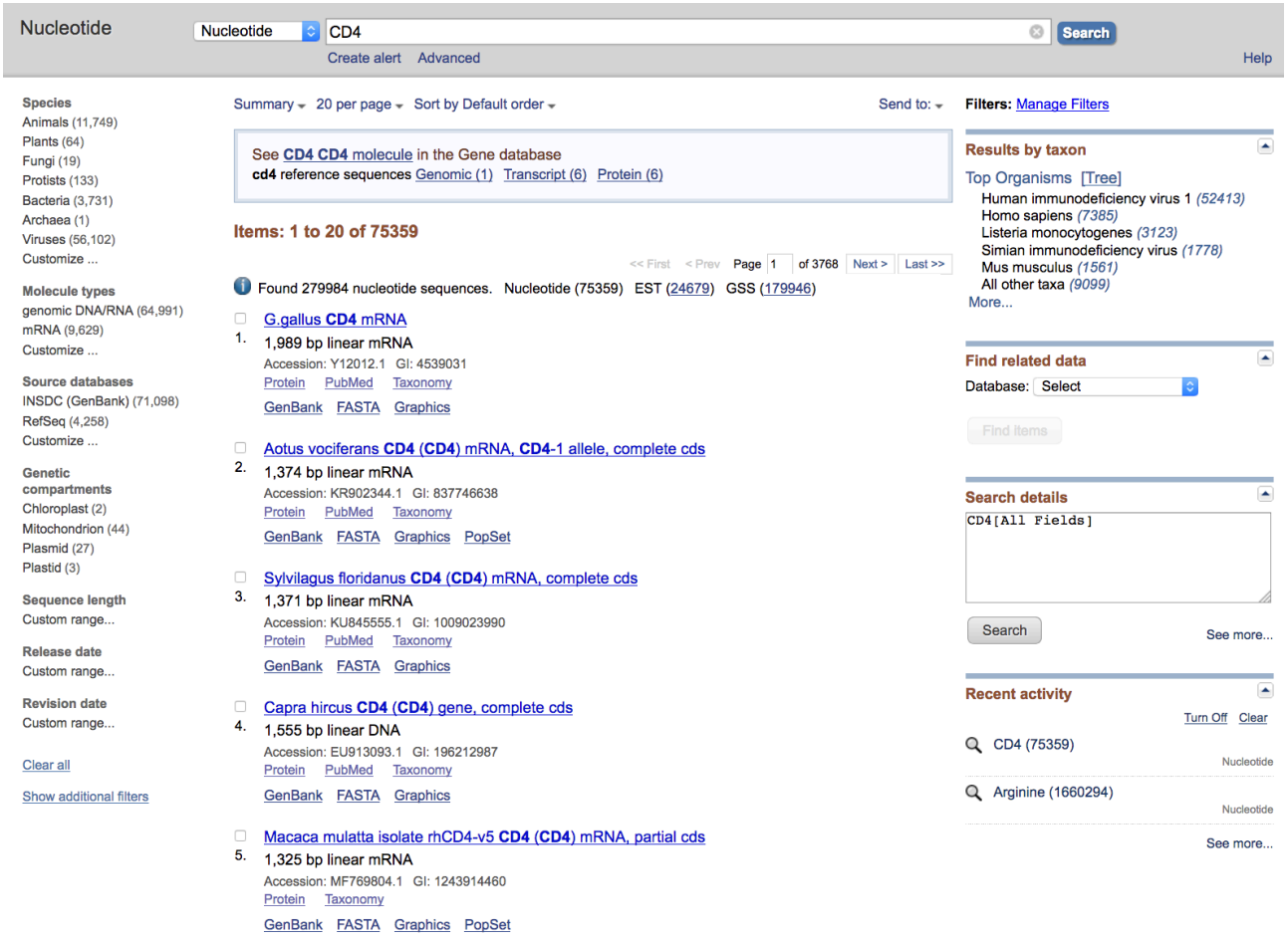


Por defecto se muestran 20 resultados por página, en formato resumen. Podemos seleccionar distintos tamaños de página (entre 5 y 200), y distintos criterios de ordenación (por accesión, fechas de lanzamiento y modificación, nombre del organismo, identificador o longitud de la secuencia). Además podemos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

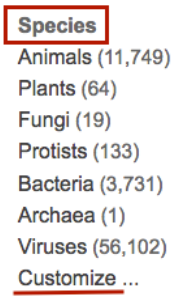
Ejemplo de búsqueda

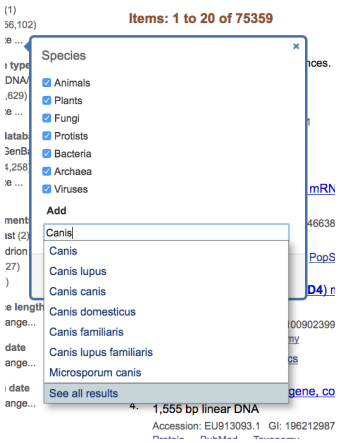
Supongamos que queremos encontrar información sobre el gen **CD4** (responsable de la expresión del antígeno CD4, fundamental para el funcionamiento del sistema inmunitario) obtenido en muestras de perros (*canis lupus familiaris*).

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:



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...*





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.

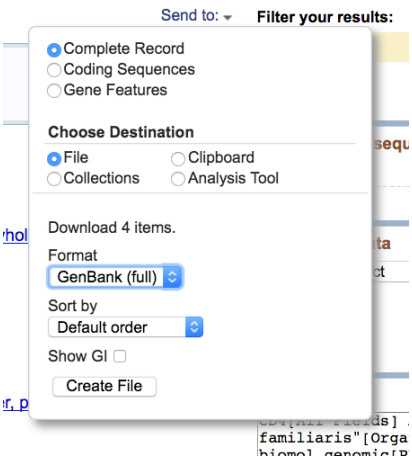
Consumo de los resultados

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:



- 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:

Bibliografía

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