

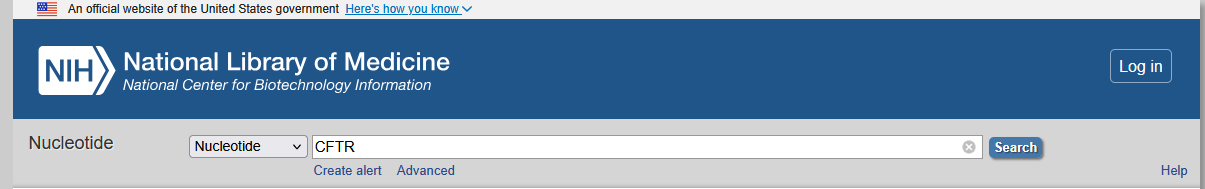
**Most important tools**

* **PubMed** – Literature search database.
* **Nucleotide**– Get nucleotide sequence for any gene, genome, protein, dna, rna etc
* **Gene**
* **Genome**
* **Protein**

NCBI Genbank is the nucleotide sequence database at NCBI.

When searching for any nucleotide sequence, Entrez queries GenBank.

# Example1: Search the nucleotide sequence of the gene CFTR





The search result gives the following:

* **Name:** CFTR mRNA, exon 5, partial cds
* **Organism Name:** Oryctolagus cuniculus
* **Sequence length:** 90 bp
* **Accession Number** **(unique identifier):** AB011005
* **Version Number:** AB011005.1
* **GI Number (obsolete identifier):** 2911136

**Note:** Given the accession number is AF086833 and the version number AF086833.2. Because the number does not contain an underscore we know that this data is in **GenBank** but is not in **RefSeq**.

**Note:** An accession number, for example, AF086833, applies tothe complete database record and remains stable even if updates/revisions are made to the record.

The version number formed by adding a dotted number such as .2 to the accession number to form AF086833.2. This number is a unique identifier for the sequence data within its record

GI numbers were integer numbers associated with a record.



### 🟥 ****Left Sidebar Filters (Highlighted in Red and Green)****

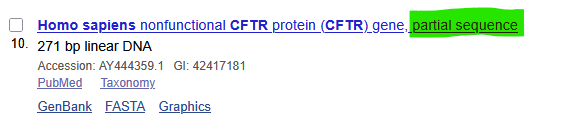
These filters help refine search results:

* Filters results by organism type
* Types of nucleotide molecules
* Classifies the type of sequences returned:

### 🟩 ****Right Sidebar (Results by Taxon)****

This panel shows **top organisms** where CFTR-related sequences are found

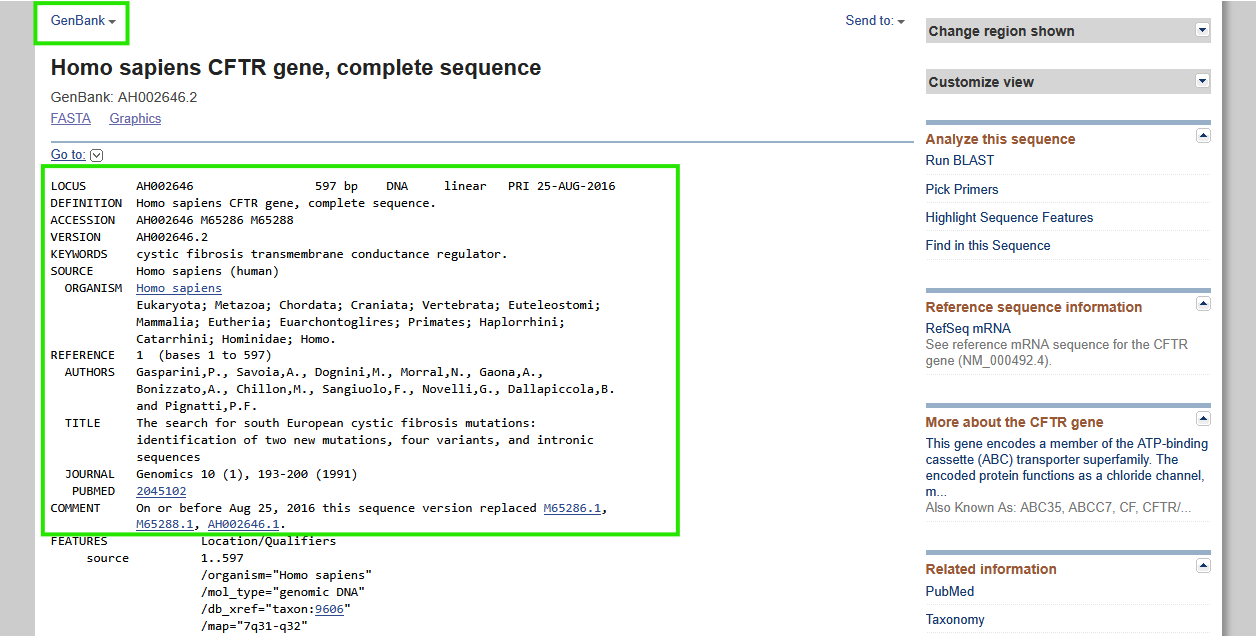




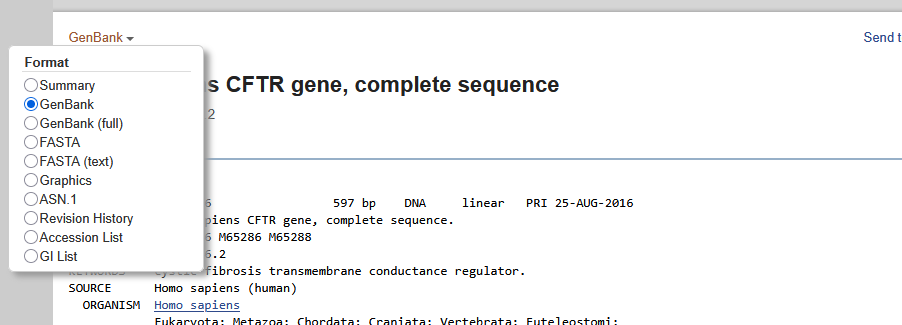
The above are examples of partial **cds (coding sequence**).

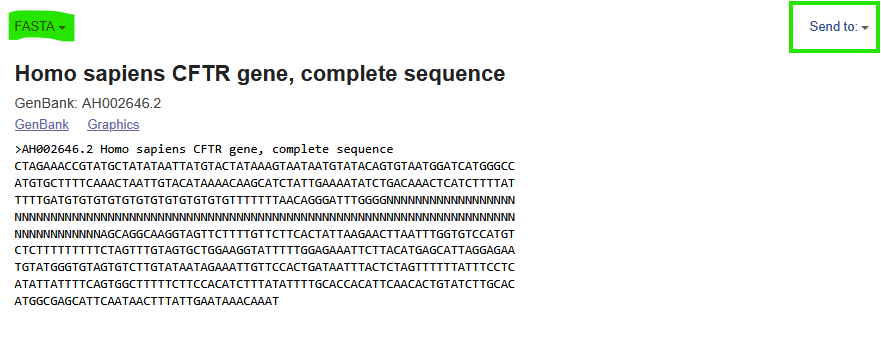


The above are examples of complete sequence.



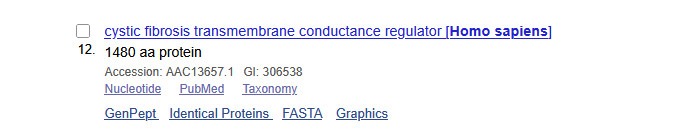
You can change output type from the highlighted.

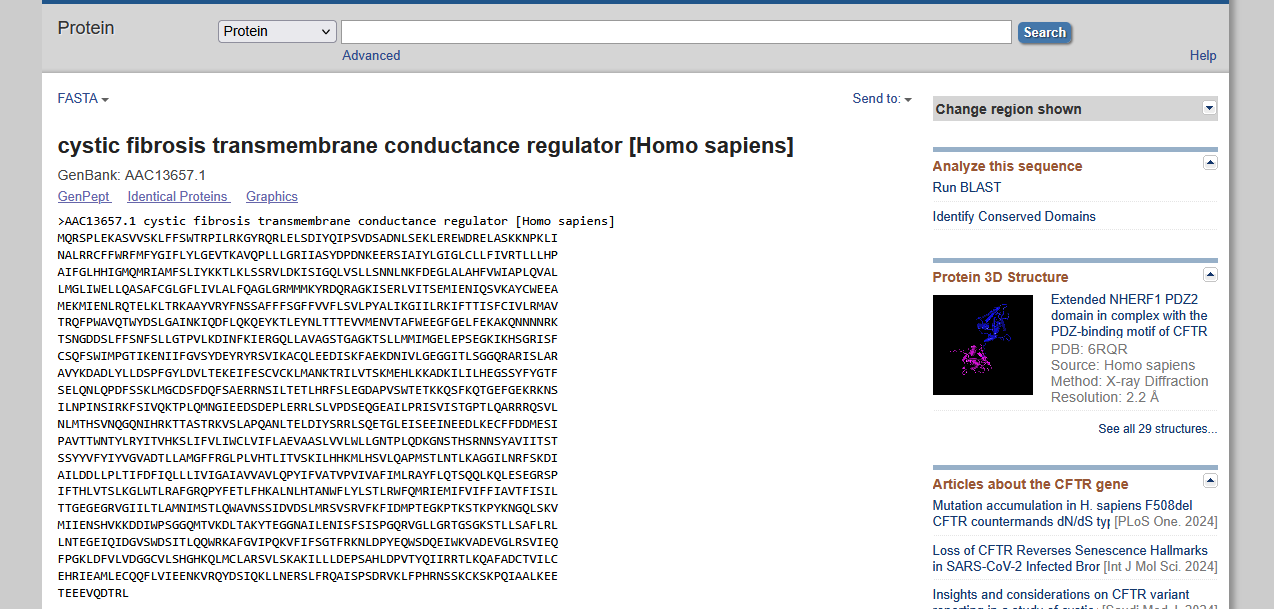


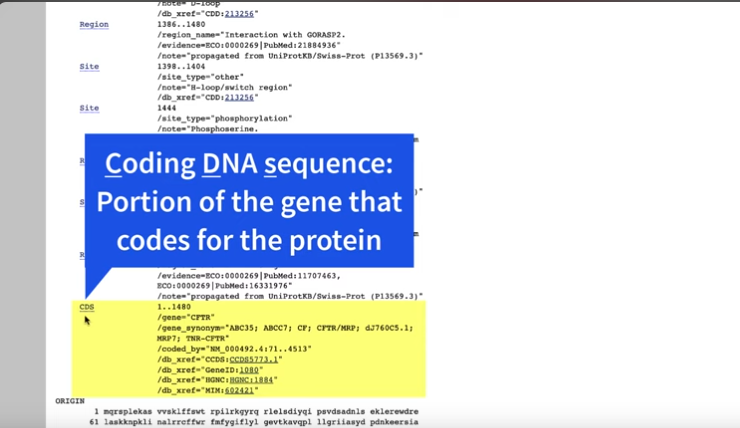


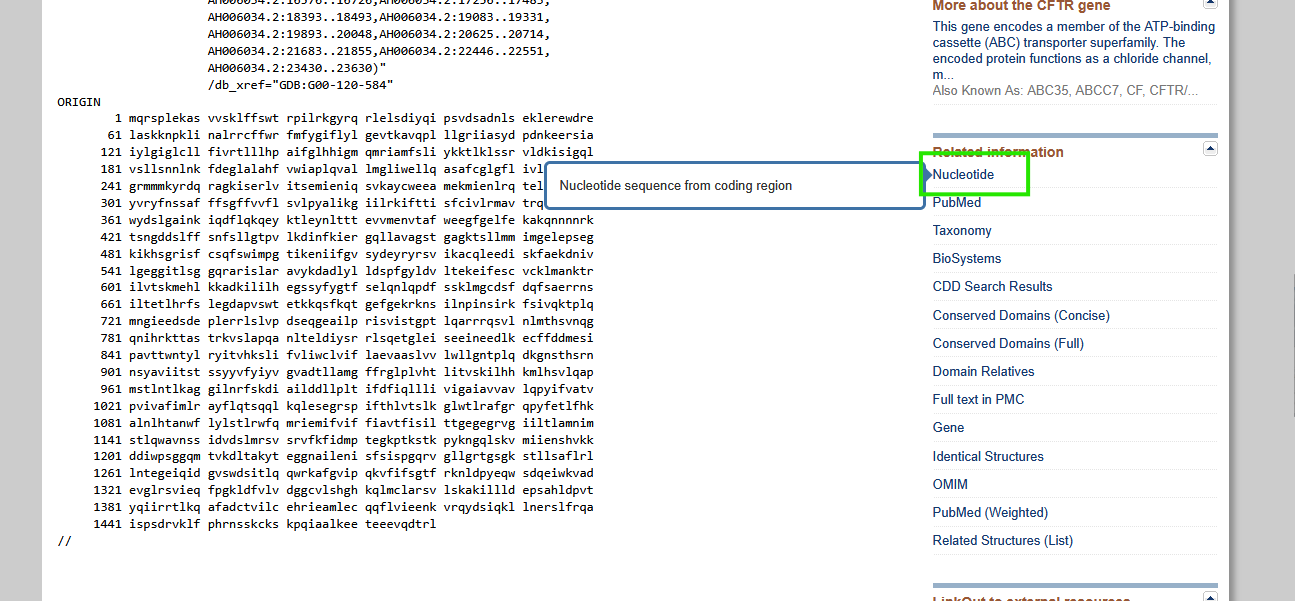
# Example2: Search the protein sequence of the gene CFTR in homo-sapiens



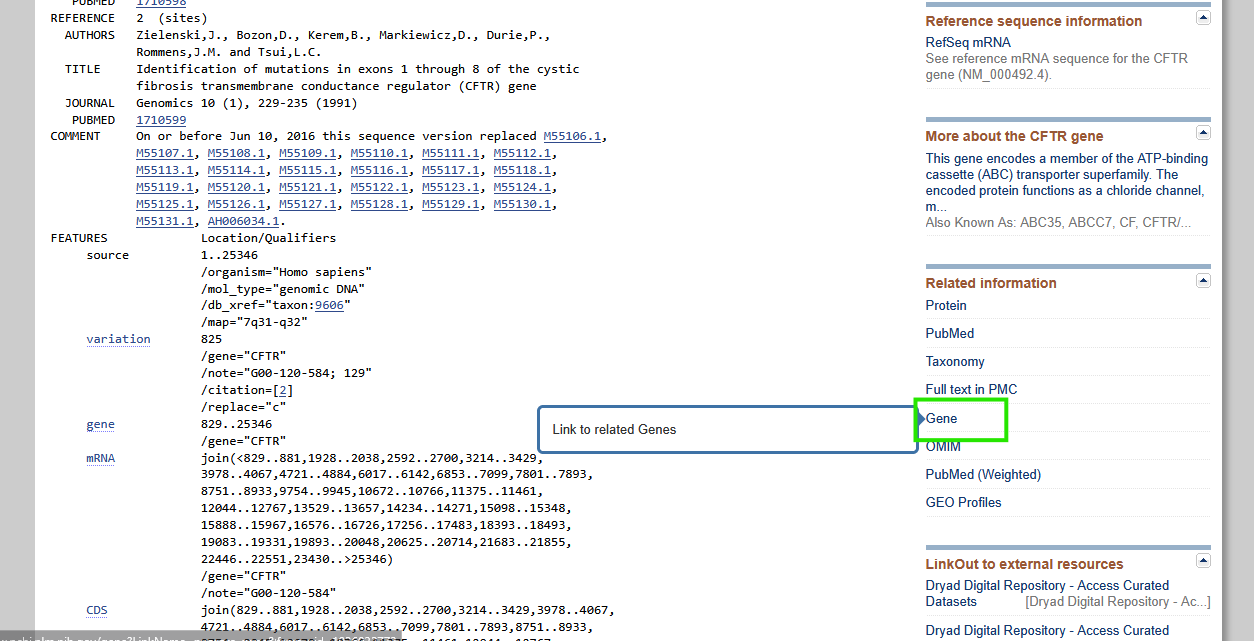








Under related information, you can get the nucleotide sequence for the corresponding protein sequence.



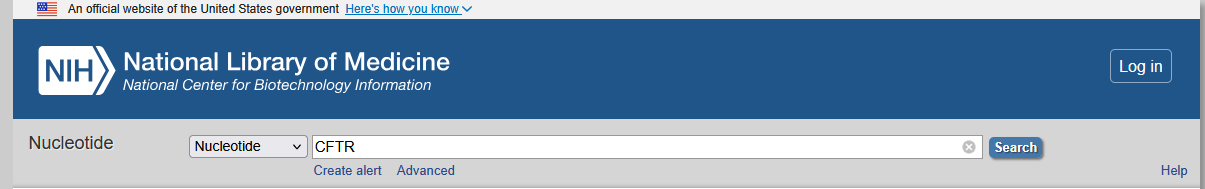
Under related information, you can get related genes



**How do I access NCBI RefSeq and GenBank?**

Entrez is NCBI’s primary text search and retrieval system that integrates the PubMed database of biomedical literature with 39 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation, and gene expression.

* NCBI web: <https://www.ncbi.nlm.nih.gov/>



* NCBI FTP: <ftp://ftp.ncbi.nlm.nih.gov/>



* NCBI’s E-utils web API



Entrez web API allows us to query NCBI data sources via a specially constructed

URL. A query URL can be of the form

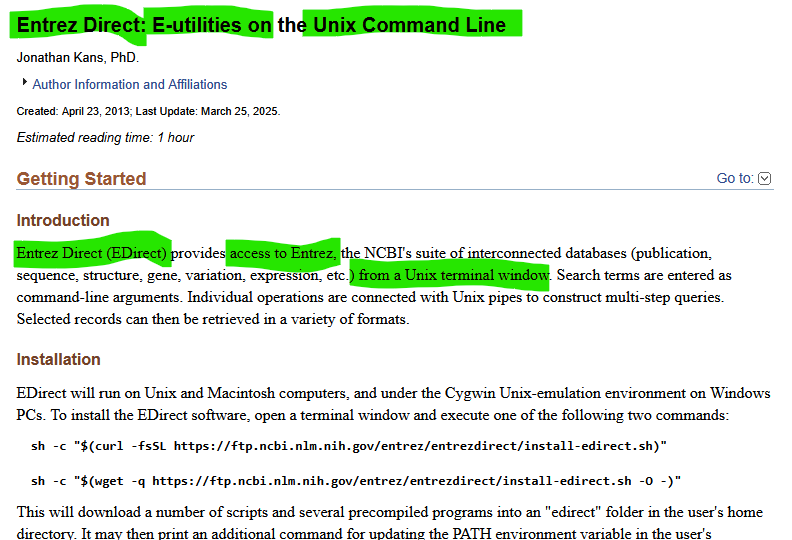
**https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch**.fcgi?param1=val1&param2=val2&

Since the & character has a special meaning for the shell, we need to either put the URL within single quotes, or we need to “escape/protect” the & character by placing a \ in front of it like so \&

### Core E-utilities:

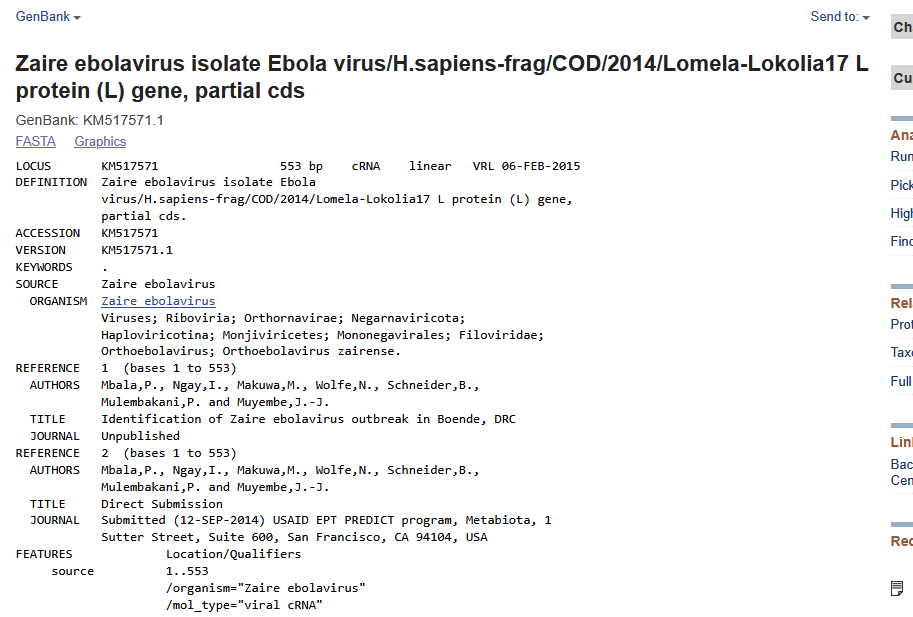
| **Utility** | **Purpose** |
| --- | --- |
| esearch | Search for IDs that match a query |
| efetch | Download full data records (e.g., sequences, abstracts) |
| esummary | Get document summaries (titles, authors, etc.) |
| elink | Find related records (e.g., gene → protein) |
| epost | Upload a list of UIDs to use in later queries |
| egquery | Check how many hits a term returns in each Entrez db |
| espell | Suggest correct spellings for queries |
| ecitmatch | Find PubMed IDs from citation info |

* Entrez-Direct (Unix-based systems)



Task: Search for the nucleotide sequence for Ebola using:

1. NCBI’s web interface

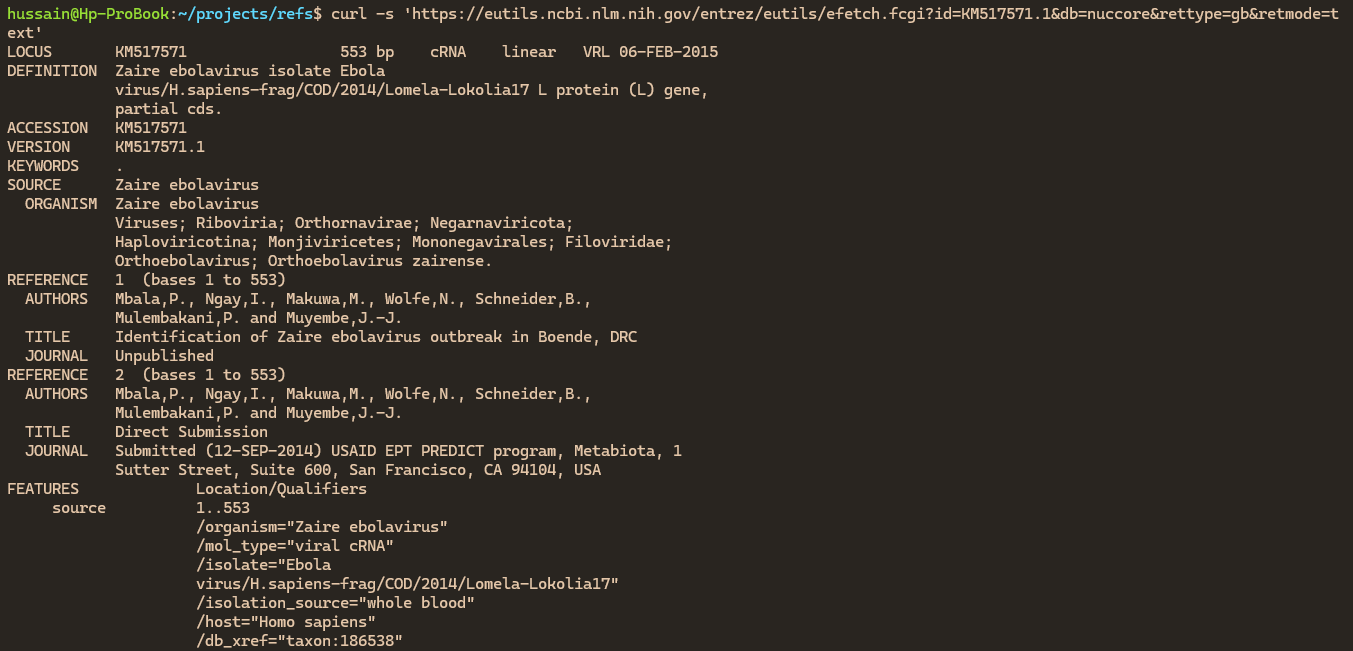


Accession ID = KM517571.1

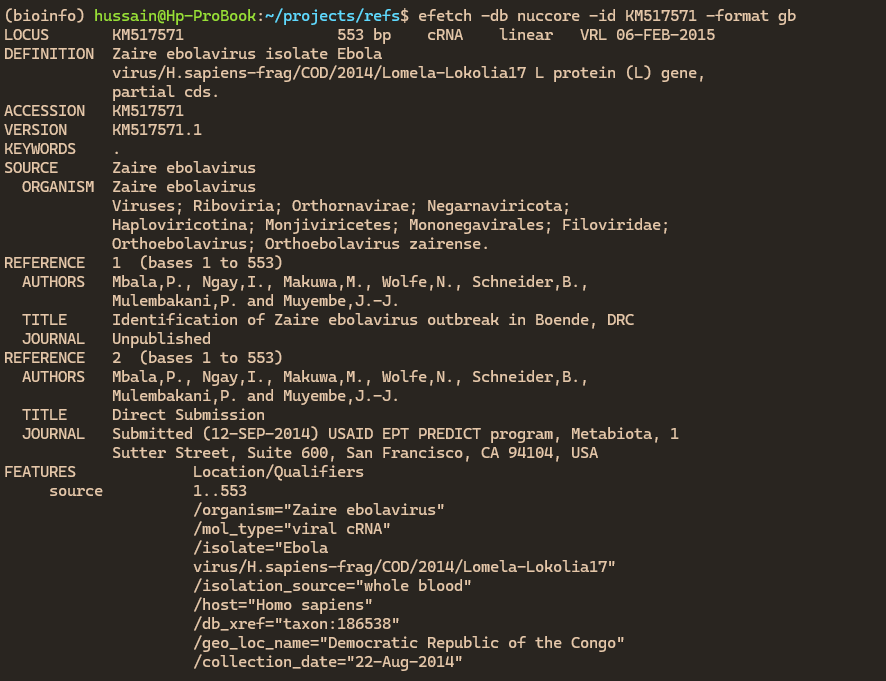
Database = Nucleotide = Nuccore

1. E-utils web API

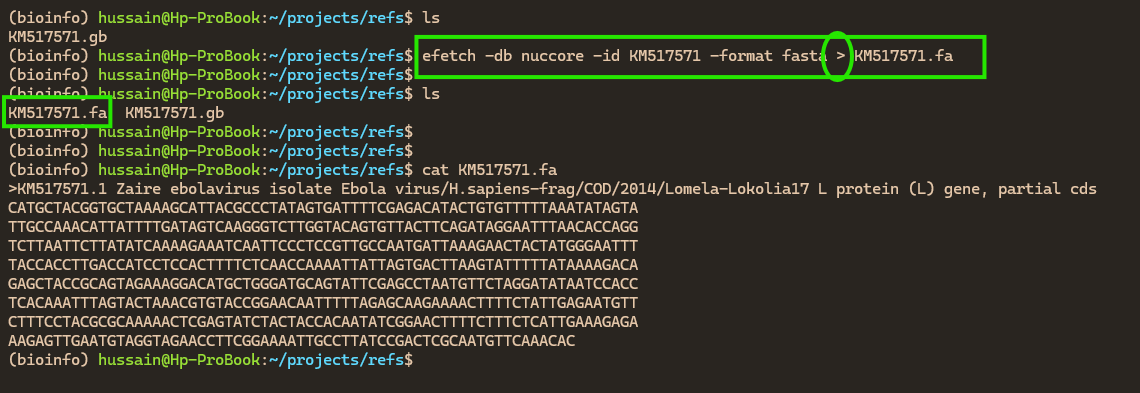
curl –s **'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch**.fcgi?id=KM517571.1&db=nuccore&rettype=gb&retmode=text'



1. Entrez-Direct



To download, just redirect it.



**Entrez Direct by example**

**How do I use efetch**

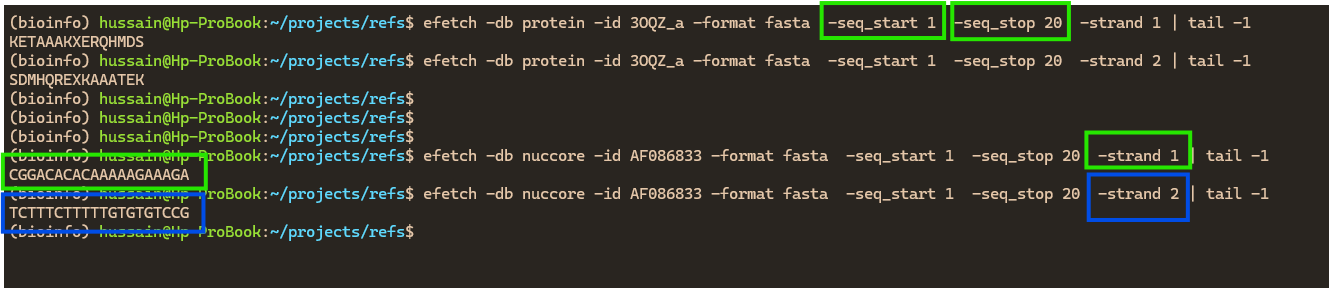
 efetch is part of the **Entrez Direct suite**.

 It pulls **full data records** using accession numbers or UIDs from Entrez.

 You specify what to retrieve using -db, -id, -format, and -mode.

Generic Syntax

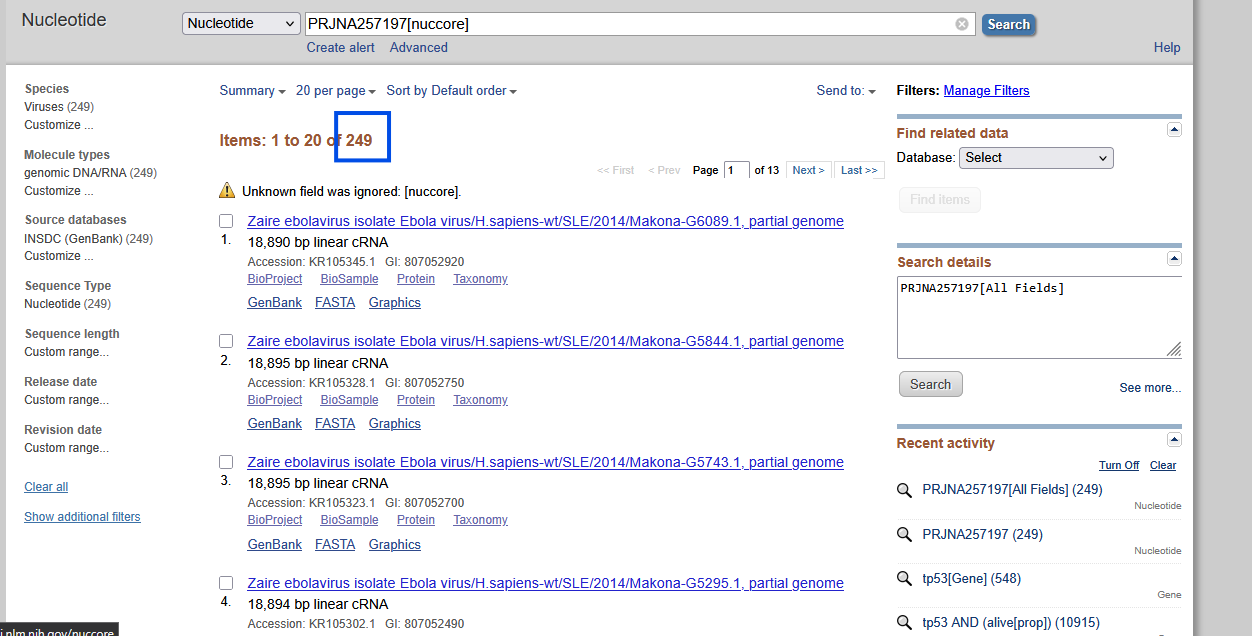
efetch -db <database> -id <accession id> -format <format> -mode <mode>

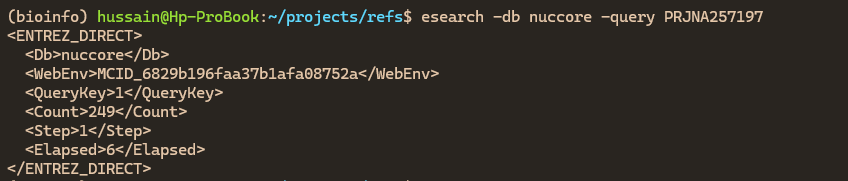


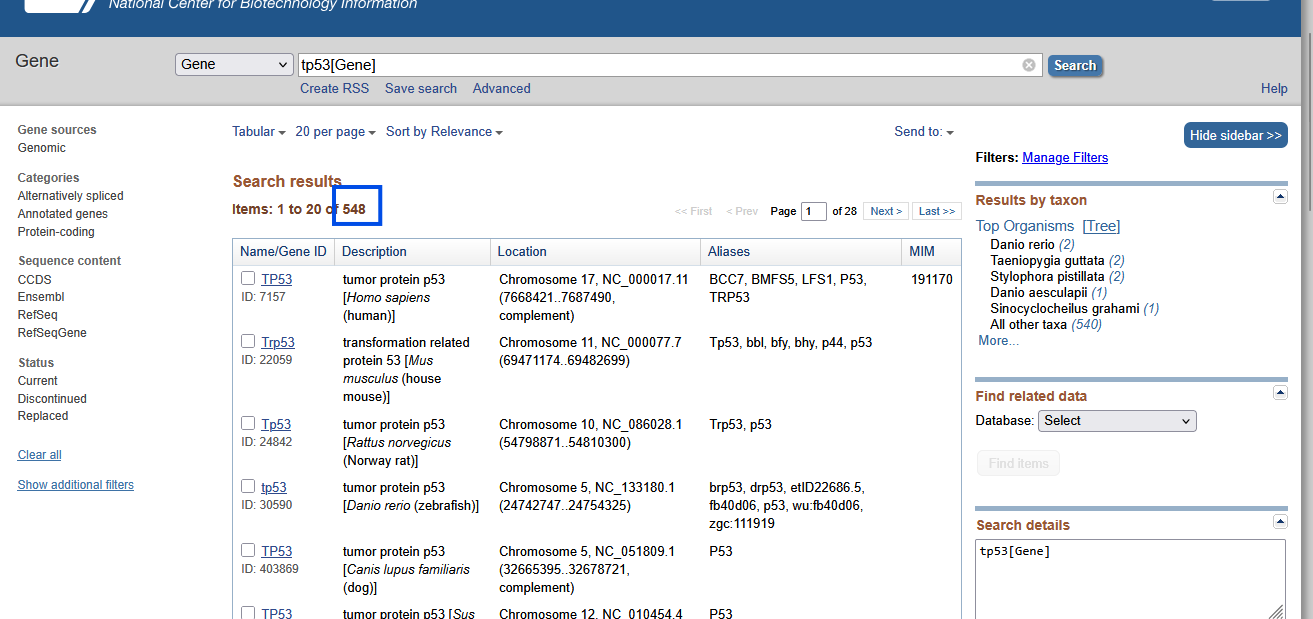
**How do I use esearch to obtain project related data?**

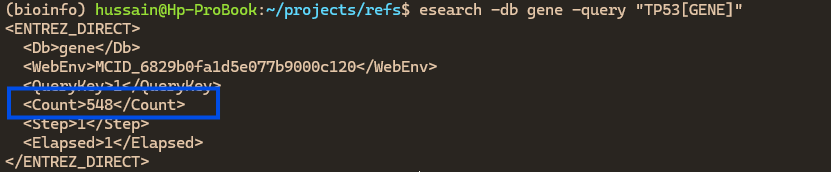
esearch is a **command-line utility in the Entrez Direct (EDirect) suite** that allows you to **search** any NCBI Entrez database using a query and get back a list of matching record IDs (UIDs).

Project accession numbers are included in published papers, or we find them in the supporting information. Once we get the project accession number, say **PRJNA257197**, we can use it to search for the data that comes with it.

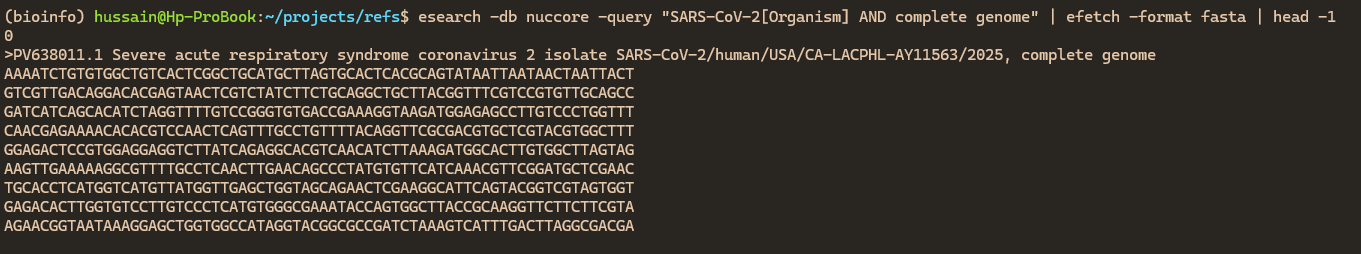








You typically use esearch in a pipeline with efetch to get actual data:



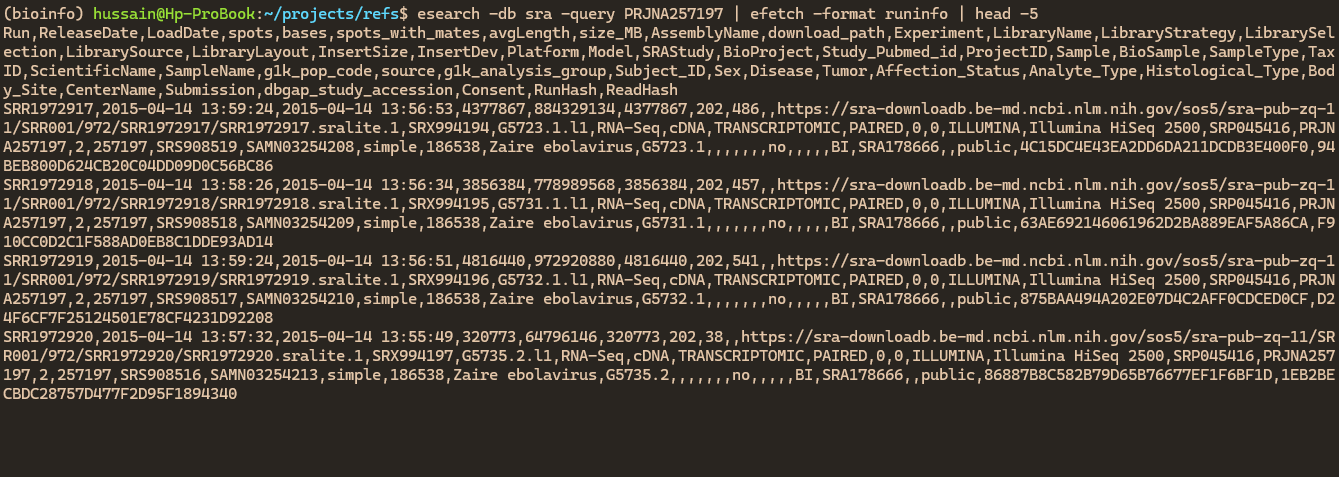
You use esearch when:

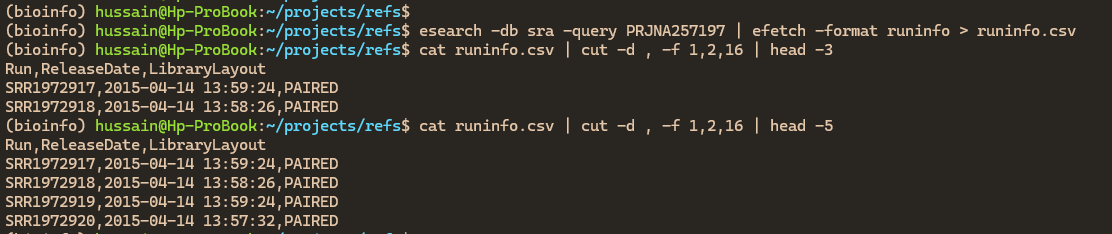
1. **You don't know the exact ID**, but you know the keywords (e.g., organism name, gene symbol, etc.).
2. You want to **search with filters** (like organism, gene name, publication year).
3. You want to retrieve **multiple records dynamically** based on a search.
4. You want to **count** the number of matching records before fetching.

**How do I get run information on a project?**

In the context of NCBI and **Entrez Direct**, especially related to the **SRA (Sequence Read Archive)** database, **Run Information** refers to **metadata about a sequencing run** — a single execution of a sequencing machine that generates raw sequence data.

A **run** represents one instance of raw sequence data generation, often as part of a larger experiment. It's a key unit in sequencing datasets.

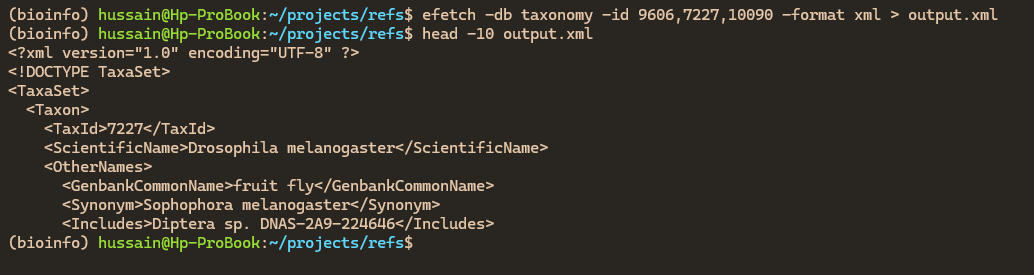
****

****

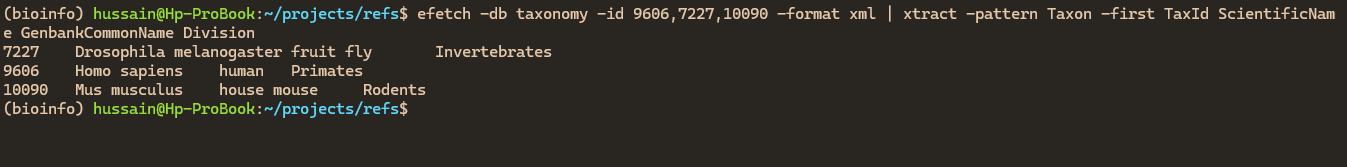
You might use Run Info to:

* Select only paired-end sequencing runs
* Identify runs with specific platforms (e.g., Illumina)
* Filter runs by read length or number of reads
* Find corresponding BioSample metadata

**How do I extract taxonomy information?**

****

It is a full XML document with lots of content in it. The xtract tool matches and extracts elements from this document



You might want to extract taxonomic info to:

* **Identify** what species or strain a sequence belongs to.
* **Classify** data in a pipeline (e.g., microbial profiling).
* **Filter** results by taxonomic rank (e.g., include only bacteria or eukaryotes).
* **Visualize** phylogenetic trees or biodiversity metrics.