

Retrieve transcript sequences from protein IDs on NCBI

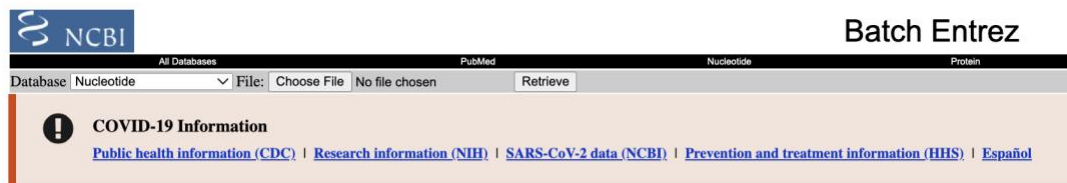
- 1) Generate a plain text file that has one gene protein ID per line.

Example plain text file:

```
CAE1329857.1
CAE1289477.1
CAE1153723.1
CAE1315817.1
CAE1296720.1
CAE1158496.1
CAE1140664.1
CAE1252791.1
CAE1328585.1
CAE1167180.1
CAE1323142.1
CAE1323437.1
```

- 2) Navigate to Batch Entrez on a web browser. <https://www.ncbi.nlm.nih.gov/sites/batchentrez>

- 3) You should see a site that looks like the following:



#### Batch Entrez

Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

#### Instructions

1. Start with a local file containing a list of accession numbers or identifiers
2. Select the database corresponding to the type of accession numbers or identifiers in your input file
3. Use the **Browse** or **Choose File...** button to select the input file
4. Press the **Retrieve** button to see a list of document summaries
5. Select a format in which to display the data for viewing, and/or saving
6. Select 'Send to file' to save the file.

#### Tips

- To download entire genome records, check the NCBI FTP site, instead of using Batch Entrez.
- Some lists of record identifiers can be tens of thousands of lines long, so Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems.
- When loading large numbers of genome records, put several thousand record identifiers per file, one per line, left-adjusted.
- Please note that Batch Entrez will check for duplicate identifiers when reporting results from a list that you have imported.
- When retrieving a list of Nucleotide accessions, you must select the specific component database from which the accessions or GIs were saved. For Nucleotide, choose either the CoreNucleotide, the EST or the GSS selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.
- In all cases, be certain to select the database that corresponds to the identifiers you are uploading. For example, if you have saved a list of protein accession numbers, be sure to select the Protein database.

- 4) In the upper left-hand corner of the site, you will need to change the database from Nucleotide to protein if you have protein IDs. You will also need to supply the plain text file you created in step 1. When you are all done click "Retrieve".

All Databases PubMed  
 Database: Nucleotide File: Choose File No file chosen Retrieve  
 All Databases PubMed Nucleotide Protein  
 Database: Protein File: Choose File s\_phar\_genes.txt Retrieve

- 5) Batch Entrez should retrieve the records for each line in your plain text file and return your fetch results that should look like the following:

Received lines: 1129  
 Rejected lines: 0  
 Removed duplicates: 0  
 Passed to Entrez: 1129  
[Retrieve records for 1129 UID\(s\)](#)

- 6) Click on Retrieve records for ## UID(s).

Received lines: 1129  
 Rejected lines: 0  
 Removed duplicates: 0  
 Passed to Entrez: 1129  
[Retrieve records for 1129 UID\(s\)](#)

- 7) This should bring you to a site with your individual records that should look like the following:

NCBI Resources How To Sign in to NCBI

Protein Protein Search Help

Advanced

**COVID-19 Information**  
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Species: Animals (1,129) Customize ...  
 Source databases: Customize ...  
 Sequence length: Custom range ...  
 Molecular weight: Custom range ...  
 Release date: Custom range ...  
 Revision date: Custom range ...  
[Clear all](#)  
[Show additional filters](#)

Summary 20 per page Sort by Default order

Items: 1 to 20 of 1129

1. [IDH3 \[Sepia pharaonis\]](#)  
 713 aa protein  
 Accession: CAE1332632.1 GI: 1969825128  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

2. [ARL3 \[Sepia pharaonis\]](#)  
 181 aa protein  
 Accession: CAE1332634.1 GI: 1969825129  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

3. [PMPCA \[Sepia pharaonis\]](#)  
 497 aa protein  
 Accession: CAE1332636.1 GI: 1969825131  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

4. [PDGD6IP \[Sepia pharaonis\]](#)  
 853 aa protein  
 Accession: CAE1332537.1 GI: 1969825187  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

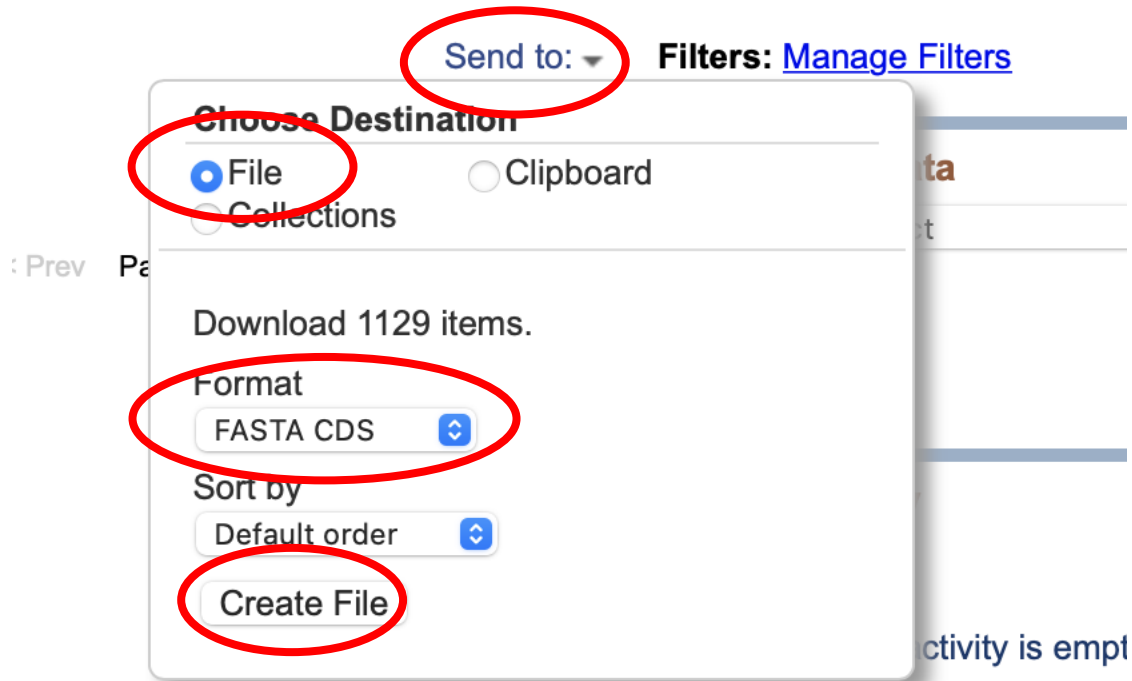
5. [UBX domain-containing protein 1-UBX domain-containing protein 1-B-UBX domain-containing protein 1-A \[Sepia pharaonis\]](#)  
 543 aa protein  
 Accession: CAE1332402.1 GI: 1969825247  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Send to: Filters: [Manage Filters](#)

Find related data  
 Database: Select Find items

Recent activity  
 Turn Off Clear  
 Your browsing activity is empty.

- 8) In the upper middle of the page, click on the drop-down box that says "Send to:". Within the drop-down box, select "File". This will populate a new field. When the new field shows, click on the drop-down box under "Format" and select "FASTA CDS". Finally, click "Create File"
- 9)



- 10) A download will initiate that has your transcript gene information.