

Entrez Search and Retrieval

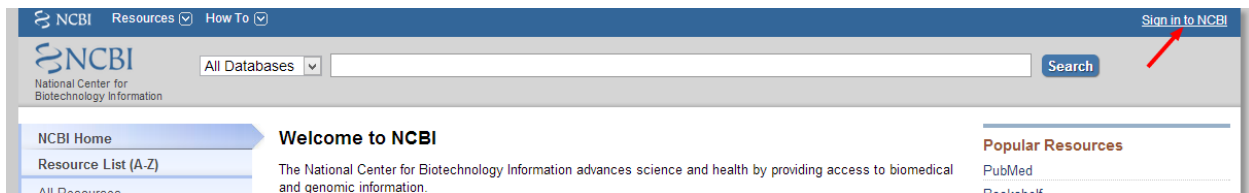
Learning Goals

- Build searches using Fields, Boolean logic, Limits, Filters, etc.
- Customize your Filters in My NCBI.
- Save your search results – My NCBI.
- Download datasets, small and large.

Guided Exercise

Create a My NCBI Account

Go to the NCBI homepage (<http://www.ncbi.nlm.nih.gov/>) and click sign in to NCBI on the top right part of the page.



Once you are on the My NCBI login page you have three options to sign in:

1. If you have already an NCBI account, you just need to enter your login and password;
2. If you don't have an account yet you can create a new one;
3. Or sign in using your google (Gmail) account.



For a better overview of the My NCBI home page let's take a look at the My NCBI YouTube video on the NCBI channel:

<http://www.youtube.com/watch?v=ks46w3mNAQE>



From Disease Name to Articles and Beyond

Our goal is to find articles, sequences, pathways and everything we can relate to Parkinson disease. Start on the NCBI home page and search All Databases with the word, "Parkinson".

We see some promising results: OMIM, dbGaP, Gene, etc., but let's look at Nucleotide and PubMed to gain some insight into the Entrez search engine.

A screenshot of the NCBI search results page. The search term "parkinson" is entered in the search box. The results are displayed in a grid format. The first row shows PubMed (56811), PubMed Central (20673), Site Search (1204), Books (1059), and OMIM (253). The second row shows Nucleotide (70270), EST (172316), GSS (615), Protein (36385), dbGaP (665), UniGene (47), CDD (36), and Clone (none).

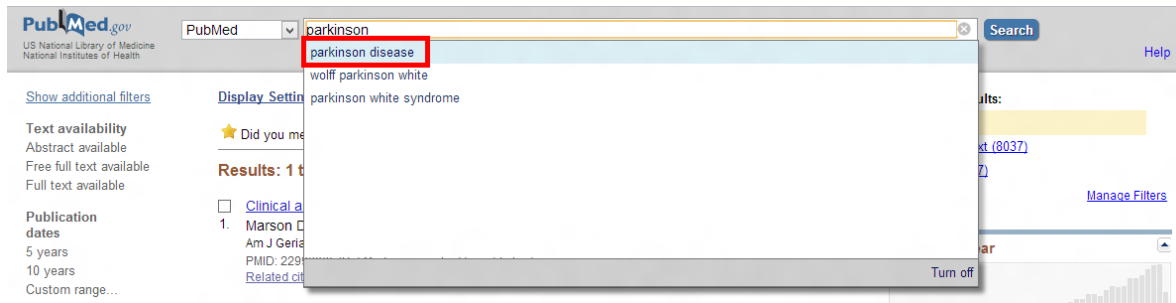
Database	Count	Description
PubMed	56811	biomedical literature citations and abstracts
PubMed Central	20673	free, full text journal articles
Site Search	1204	NCBI web and FTP sites
Books	1059	online books
OMIM	253	online Mendelian Inheritance in Man
Nucleotide	70270	Core subset of nucleotide sequence records
EST	172316	Expressed Sequence Tag records
GSS	615	Genome Survey Sequence records
Protein	36385	sequence database
dbGaP	665	genotype and phenotype
UniGene	47	gene-oriented clusters of transcript sequences
CDD	36	conserved protein domain database
Clone	none	integrated data for clone resources

View the Nucleotide results. Open one of the Dickey bacteria records to see why our query returned virus records (use the browser's find in page function). This demonstrates why using Field qualifiers in your queries is a good idea. See Search Details to see how Entrez Nucleotide translated the query term. Now add "Disease" to the cross database search; your query is now, Parkinson disease.

A screenshot of the NCBI search details page. The search term "parkinson[All Fields]" is displayed in a text box. Below the text box, there is a "Search" button and a "See more..." link.

```
Search details
parkinson[All Fields]
Search See more...
```

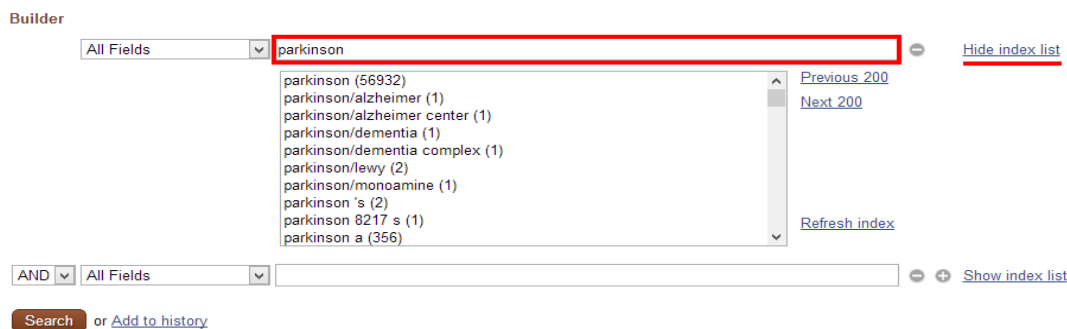
Open the PubMed set of results. This is a large set, which we will restrict later. First, delete the word 'disease' and you will see a list of suggested search terms. This demonstrates a unique feature of PubMed's search engine. Some of these terms come from MeSH, others come from language processing algorithms developed at NCBI. Select "Parkinson disease." Look at the Search Details to see how PubMed translated the query.



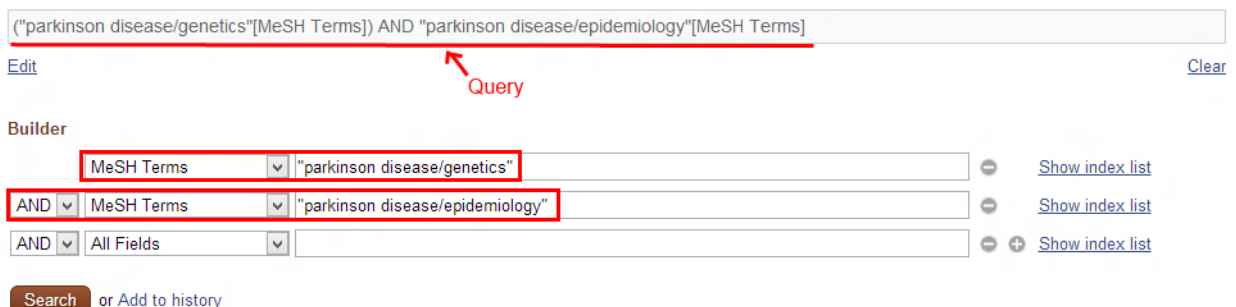
To see other Fields that might help your search, go to the Advanced Search page.



In the All Fields pull-down menu, select MeSH terms and click Show Index. Next, type parkinson in the text box and view the index again. This procedure can be used for any NCBI database in the Entrez system.



If interested in the genetics and epidemiology of Parkinson disease, follow this procedure: Select "parkinson disease/genetics" and click the Add to Search Box button. Select "parkinson disease/epidemiology" and repeat. Your query should now be "parkinson disease /epidemiology"[MeSH Terms]) AND " parkinson disease/genetics"[MeSH Terms]. Click Search.



This query returns a more manageable set of records, and Entrez suggests subsets under “Filter your results” including those with free full text or reviews.

Filter your results:

All (578)

[Free Full Text \(87\)](#)

[Review \(68\)](#)

[Manage Filters](#)

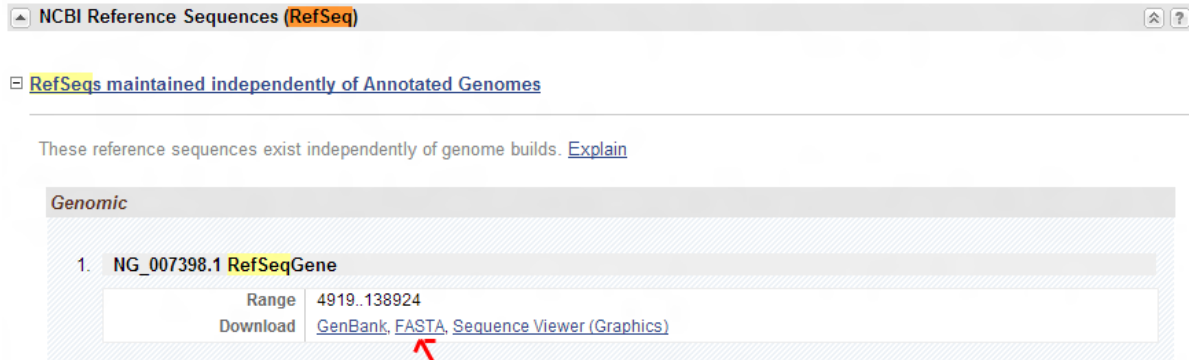
You can have Entrez display filters of your choosing. Click on Manage Filters. If already signed in to My NCBI, you are taken to your list of PubMed Filters, otherwise, sign in or register for a new account. To add a filter for OMIM, go to Search for Filters, enter OMIM, and check the box for Links to OMIM. You will now see that filter every time you run a PubMed search.

The screenshot shows the 'My NCBI — Filters' interface. At the top, it indicates 'You are managing filters for: PubMed' and allows switching to another database. Below this, there are two main sections: 'Your PubMed filter list' and 'Browse/Search for PubMed Filters'. The filter list shows three active filters: Clinical Trial, Free Full Text, and Review, all of which are 'Standard filter' type. The 'Browse/Search for PubMed Filters' section includes a 'Select category' dropdown with options: Popular (selected), LinkOut, Properties, and Links. There is also a 'Search with terms (optional):' input field and a 'Search' button. A table of available filters is shown below, with checkboxes for each: Clinical Trial (checked), English, English & Humans, Free Full Text (checked), Full Text, Humans, Items with Abstracts, Published in the last 5 years, and Review (checked). Each filter has a brief description.

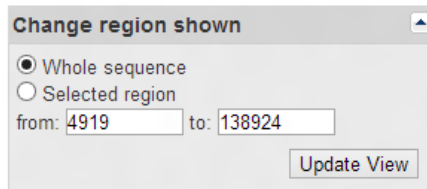
Look at Gene records associated with our original set of articles for the two MeSH terms. Note that the filter, Links to Gene, is a subset of PubMed records. If you want all Gene records associated with this set, use the “Find related data” tool. Select the Gene database, and the default Option, and click the “Find items” button. You could also have opened a single PubMed record and used the Gene link under “All links from this record.”

The screenshot shows the 'Find related data' tool. It has a title bar with a small icon. Below the title, there are two dropdown menus: 'Database:' set to 'Gene' and 'Option:' set to 'Gene (OMIM)'. Below these is a paragraph of text: 'Gene records associated with Online Mendelian Inheritance in Man (OMIM) records that cite the current articles in their reference lists.' At the bottom, there is a 'Find items' button.

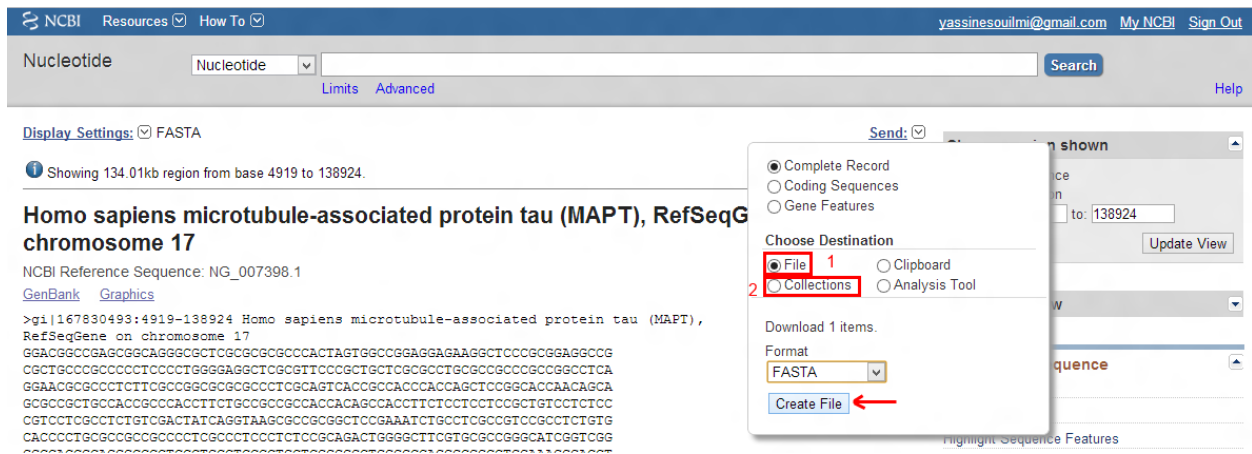
To demonstrate download and save options, open the MAPT gene record for human (Gene ID: 4137), go to Reference Sequences (NG_007398.1 RefSeqGene), and then get the RefSeq set in a local file in FASTA format.



Select the desired region using the change region shown tool.



Select the parameters and the desired file format you want and then click 'create file' (by checking File under 'choose destination' options) to download your gene RefSeq.



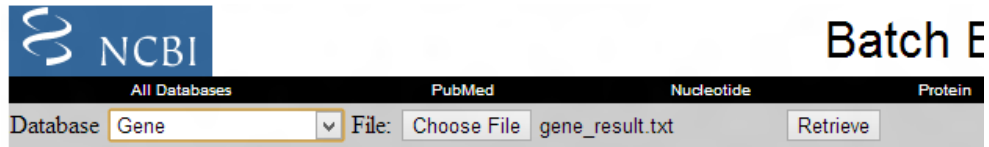
To save this set in My NCBI, where you can share the set with others, choose 'Collections' rather than 'File' (2).

If your set of records is too large to save locally, you have two basic options.

- 1) Batch Entrez
- 2) Entrez Programming Utilities (E-Utilities, or eutils)

For Batch Entrez (<http://www.ncbi.nlm.nih.gov/sites/batchentrez>), you need a text file of identifiers for the records you want. Using the set of IL10 records above, go to send to -> File -> GI List. You can then upload that file on the Batch Entrez page – be sure to select the correct database on the page. This

approach allows you to break up a very large list of gi's into smaller sets and download those smaller sets.



To automate retrieval of larger sets of records, use the E-Utilities (see <http://www.ncbi.nlm.nih.gov/books/NBK25501/>). We don't plan to demonstrate the eutils, but here is an example URL for retrieving part of a large genomic record in FASTA format:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=NC_000001.9&from=13540856&to=13546098&strand=true&retmode=text&rettype=fasta