The Ultimate Guide to Navigating GEO Effectively

The Gene Expression Omnibus (GEO) is a public database/ repository. GEO provides a flexible and open design that facilitates the submission, storage, and retrieval of heterogeneous data sets from high-throughput gene expression and genomic hybridization experiments.

This guide is your one-stop resource to use GEO effectively. It gives you a comprehensive set of answers which can help you with all the why(s), how(s) and what(s) of using GEO.

TABLE OF CONTENTS

WHY GEO

GEO is one of the largest open-source repositories for **high-throughput data on gene expression studies** as well other data applications, including those that examine **genome methylation, chromatin structure,** and **genome–protein interactions.** It serves as a good playground for beginners to start their research in these areas as well as a platform for researchers/ scientists working in this domain to find relevant data.

THE FOCUS AREAS OF GEO

- **Providing a robust, versatile database:**
  To store high-throughput functional genomic data efficiently.
- **Offering simple submission procedures:**
  To support complete and well-annotated data deposits.
- **Providing user-friendly tools:**
  To query, review, and download studies and gene expression profiles of interest.

DATA THAT YOU CAN FIND ON GEO

GEO provides access to data from most of the high-throughput and parallel molecular abundance-measuring technologies in use today. These include data generated from:

- **Gene expression profiling**
- **Non-coding RNA profiling**
- **Chromatin immunoprecipitation (ChIP) profiling**
- **High-throughput RT-PCR**
- **Genome variation profiling**
- **SNP arrays**
- **Serial Analysis of Gene Expression (SAGE)**
- **Protein arrays**

HOW IS DATA ORGANIZED ON GEO

GEO collects the following data from a submitter:

- General information about the study: The why(s) and how(s)of the experiment
  - These form a **'Series'** record.

- Specifics about the sample and data for each sample
  - These forms a **'Sample'** record.

- Information about the array
  - These forms a **'Platform'** record.

The data collected is then organized into a standard format which is explained in the figure.

- A **Platform record** is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template. Each Platform record is assigned a unique and stable GEO accession number (GPLxxx).

- A **Sample record** describes the conditions under which an individual Sample was handled, the -manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx).

- A **Series record** links together a group of related Samples and provides a focal point and -description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx).

FIGURE SHOWS HOW DATA IS ORGANIZED ON GEO.

A- Description of array,

B- Table of the array template, protocols used

C- Description of the biological sample and the

D- Table of processed hybridization result

E- Original raw/ processed sequence data file,

F- Description of the overall experiment,

G- Tar archive of original raw/ processed sequence data files

Few selected data undergo a higher level of structuring into **Datasets** and gene **Profile** records. These data (from the GEO Series record) are reassembled by GEO staff into GEO Dataset records (GDSxxx). Gene expression profiles derived from curated GEO Datasets are stored in the GEO Profiles database.

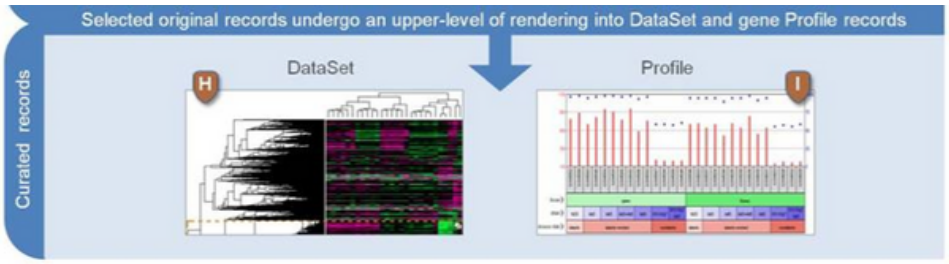
- GEO Dataset is a **study-level** database.It represents a curated collection of biologically and

statistically comparable GEO Samples and forms the basis of GEO's suite of data display & analysis tools.

Not all submitted data are suitable for Dataset assembly. Therefore, all Series records may not have corresponding Dataset record(s).

- GEO Profiles is a **gene-level** database. It stores gene expression profiles derived from curated GEO Datasets.

  -Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a Dataset. Experimental context is provided in the bars along the bottom of the charts. This helps one find if a gene is differentially expressed across different experimental conditions very easily.



Some records on GEO are structured better and presented in the form of Datasets and Profiles.

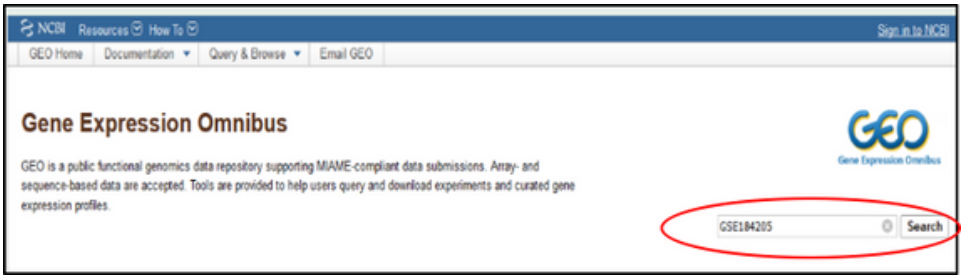
H- Clickable thumbnail cluster image directed to the Dataset record with contains several data analysis tools

I- Clickable Thumbnail chart to enable rapid visual scanning and comparison of multiple Profiles.

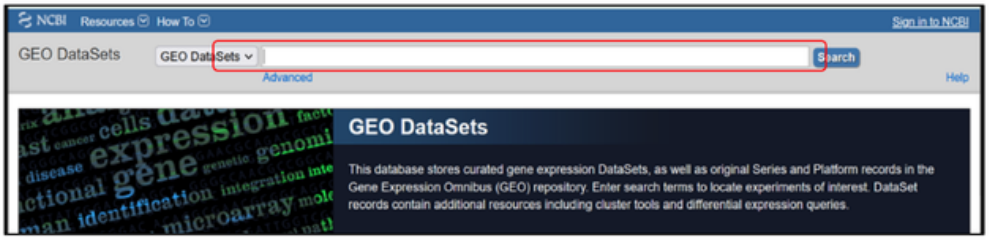
HOW TO QUERY DATA ON GEO:

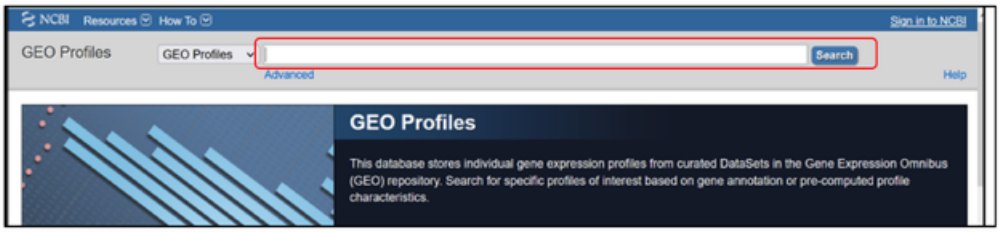
GEO data can be retrieved in several ways:

- **If you have the accession number,** you can use the GEO accession box located on the GEO homepage to look at the specific GEO record.

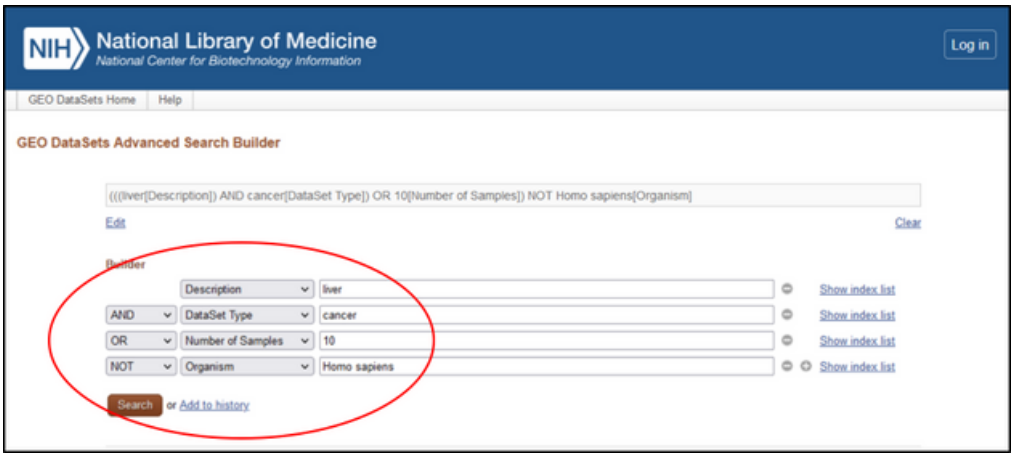


To locate data relevant to your interests(without the accession number), you can search the data on the GEO DataSets and the GEO Profiles Searches may be performed by entering appropriate keywords and phrases into the search box.





However, given the large volumes of data stored in these databases, it is often useful to perform more refined queries using the advanced search in order to filter down to the most relevant data.



It is important to note that all GEO records are not structured to this level. So, you might want to search for publications, find out the accession numbers, and then search on GEO if you want to find specific datasets.

KEYWORDS THAT HELP WITH SEARCH

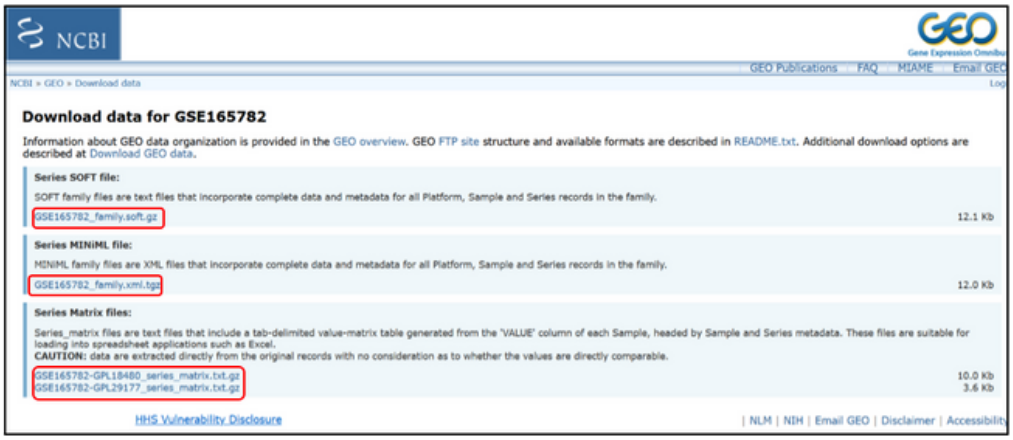
Our curation experts suggest the use of functionality/drug actions such as 'inhibition', 'activation', 'stimulation', etc., to find out datasets related to any drug or disease.

HOW TO DOWNLOAD DATA FROM GEO

Now that we know how to query data on GEO, we need to be able to download and store it for further analysis. There are multiple ways to download the data on GEO. The most straightforward method is given below:

Search for the relevant dataset using the keyword/ the accession number. This is how it shows up. Click on 'Download data.'

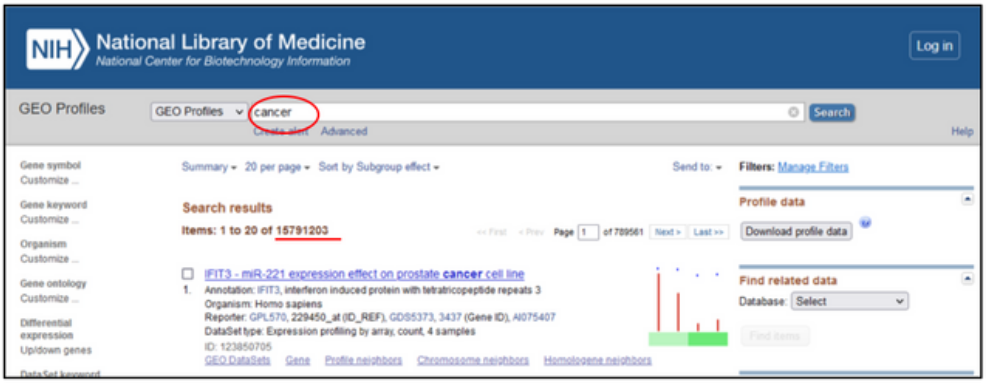
You can download these compressed files in different formats and store it for analysis later.



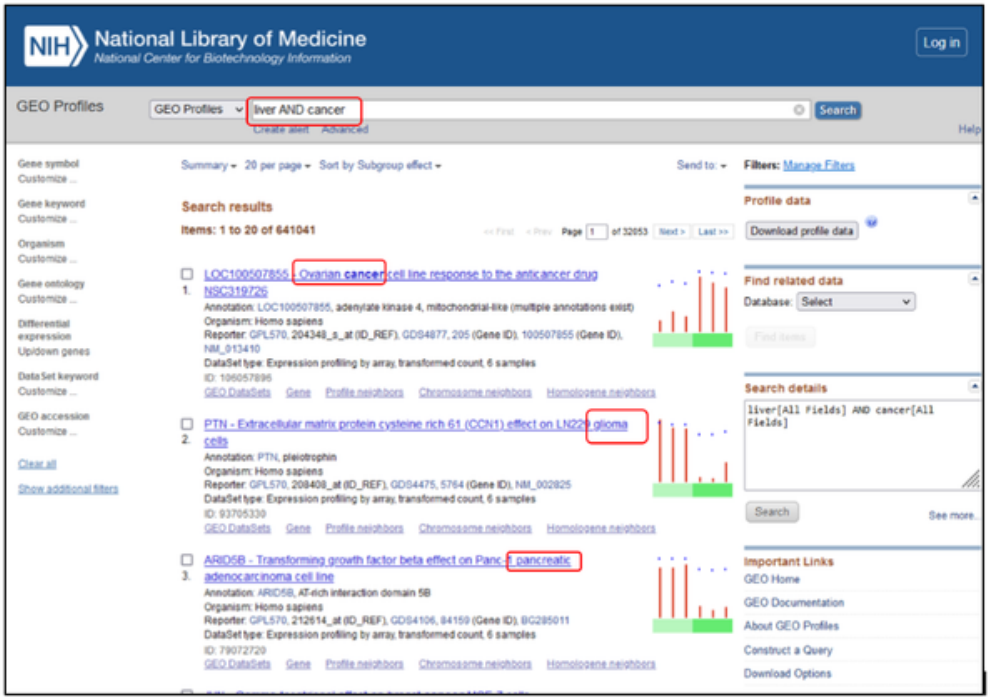
CHALLENGES FACED WHILE USING GEO

- The main challenge while working with GEO is the difficulty of **retrieving data.** Some of the most useful metadata for each dataset in GEO is stored in unstructured English text that is difficult for researchers to utilize effectively. Unless you give correct keywords while searching, the search results might be completely off. Also, using multiple keywords might give vastly different results.

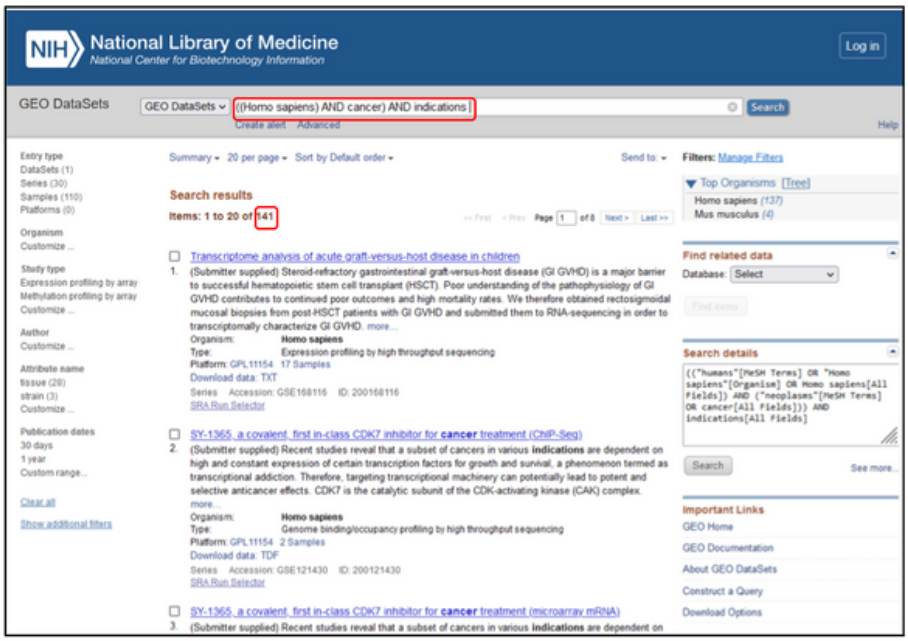
  This can be illustrated through a simple example:
  The search for 'liver' on GEO profiles (gene-level database) turns up 9015752 results; a search for 'cancer' turns up 15791203 results. But a search for 'liver cancer' does not show up any results.



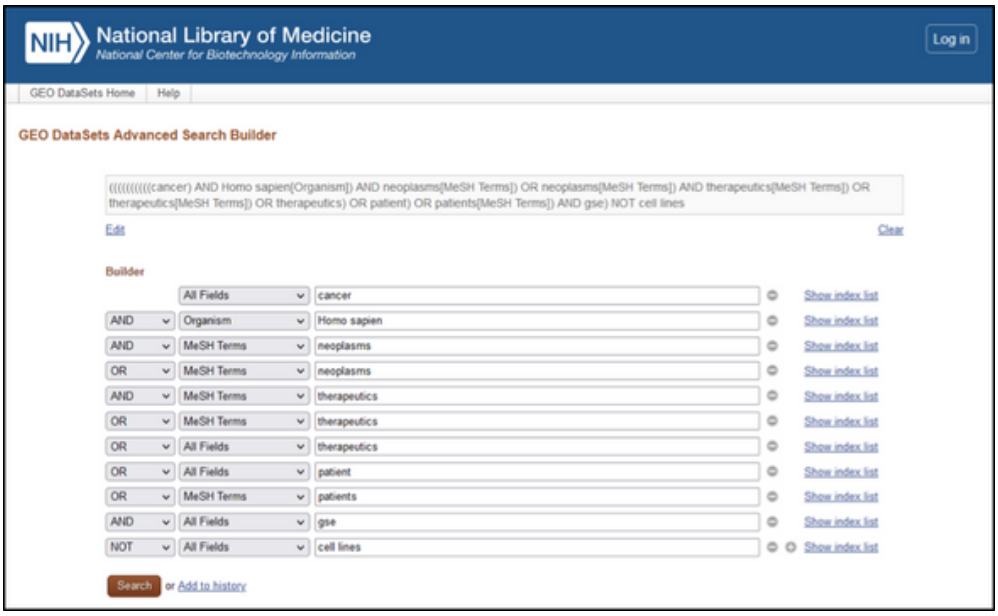
Even by using the advanced search option, one might not be able to optimize the search experience. the search results might not be relevant unless the user defines all the relevant areas.

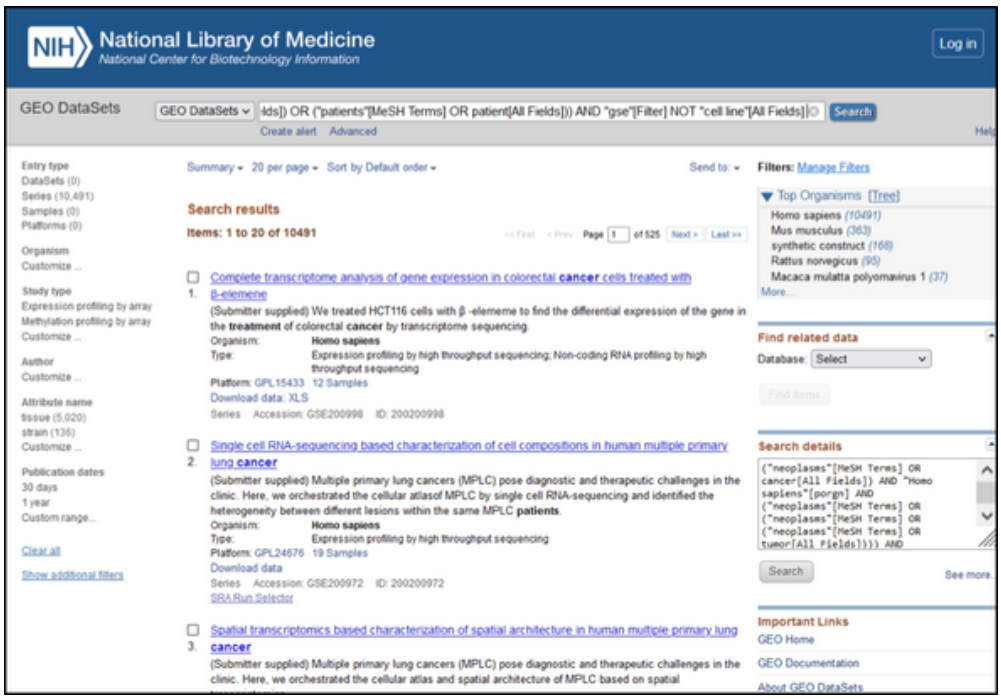


Another example: If you are trying to do a straightforward search to find GEO DataSets (containing sample level data) on human cancer indications this is what you can retrieve:



With a detailed query to streamline the search on GEO, you can improve the results obtained.





If you look at how the query has been structured, you can imagine how much thought has been put in constructing this query. However, one cannot be sure that all the relevant datasets have been retrieved.

- When you download data, it gets downloaded in the form of compressed files. You will know if the relevant data fields are present only once you go through the downloaded files individually. Another challenge is that you must analyze these files individually. You can't load it on a process pipeline because the data is **not standardized.**

- The data on GEO does not follow a particular **ontology.** So, it might be important for you to find out the synonyms and the acronyms/ abbreviations of the keyword of interest to improve your search results.

- Many a time, keyword search can be **misleading** because of wrong metadata tags placed on the records.

WHAT YOU CAN'T DO ON GEO

Though it is an indispensable resource in medical research, there are certain things that cannot be done on GEO. It is important to know these as well so that you need not waste time trying to search for these functionalities.

- Due to a lack of curation, it is difficult to **identify patient details/ clinical data.**

  Metadata annotations are lacking/ incomplete in most cases due to which one has to literally read through the experimental design, description, and the publication (if available) to find out what kind of clinical data is contained in the dataset. This can be done manually if there are 5 publications... but what if there are 1000? Searching for information regarding cell lines or tissues is also not streamlined, and a lot of information is lost while searching for datasets.

- Another important missing feature in GEO is that one can't **search for drugs.**

  If you want to find datasets involving a particular drug or a drug-disease interaction, you must first find out publications that talk about these, look for the GEO IDs in the articles and then search for those on GEO to get relevant datasets. This is a very time-consuming process, and one could miss important datasets if they are not associated with journal publications.

- Making **connected queries.**

  Like checking for experimental data dealing with a particular disease in a specific organism or disease-drug combination is difficult because the metadata is not always well documented or curated.

BONUS TIP FOR ADVANCED GEO USERS

We encourage you to reach out to us if you are a frequent GEO user. We understand that you would face difficulty in **retrieving** relevant datasets, **harmonizing** different data formats, or **finding** the transcriptomics data you are searching for from GEO.

Our data centric ML Ops platform, Polly, hosts over 1.5 million highly curated ML-ready biomolecular datasets from various repositories like GEO, TCGA, etc. This level of curation ensures that you get all the relevant datasets in seconds just by doing a keyword search. Additionally, the curation fields help you filter the data to get very streamlined results. Since the data is highly curated and harmonized, various analyses can also be carried out easily.

Book a demo to know more about how to accelerate your research!