GEO Database Dataset to Seurat Import Guide:

1. [Create Query String](https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html):
   1. For example this link returns datasets from humans with “Hidradenitis Suppurativa” and RNA in the description published in the last year
      1. [https://www.ncbi.nlm.nih.gov/gds?term=human[organism]+AND+%E2%80%9Dhidradenitis%20suppurativa%E2%80%9D+RNA+2023/01:2024/07[Publication%20Date]](https://www.ncbi.nlm.nih.gov/gds?term=human%5borganism%5d+AND+%E2%80%9Dhidradenitis%20suppurativa%E2%80%9D+RNA+2023/01:2024/07%5bPublication%20Date%5d)
2. Write down relevant accession codes (GSE\*\*\*\*\*\*) onto a txt file separated by newline
3. Downloading Data
   1. Programmatically:
      1. Inside your command line, run the following command:

*python geo\_dl.py <path to accession code txt file> <data output folder>*

* 1. Manually:
     1. Go to the webpage of each dataset. On the bottom there should be a download link for a tarball

A close-up of a computer screen

Description automatically generated

* + 1. You can use 7zip or gunzip in linux to extract each dataset. Don’t extract subfolders as Suerat expects to read in a gz archive.
    2. A 10x genomics single cell RNA dataset is a combination of a barcodes, features, and matrix file.

The barcodes.tsv file contains the unique DNA identifier for each cell which is assigned before sequencing. The features.tsv file contains a list of all genes in the dataset, with each line containing a gene’s unique ID and name. Finally, the matrix.mtx file contains information on each gene’s expression level by cell, with each line containing three values corresponding to the row number of the gene from features.tsv, the row number of the cell from barcodes.tsv, and the expression value respectively.

For Seurat to read in the data, these three files must be organized into their own directory. Group the files together based on their sample ID (first part of file name before the first underscore).

If you see genes.tsv, rename it to features.tsv

* + 1. Finished directory should look something like this:  
       A screenshot of a computer

       Description automatically generated
    2. With each sample containing three files:

A black rectangle with white text

Description automatically generated

1. Reading Data into Seurat:

*library(Seurat)*

*seurat\_obj <- Read10X(‘path to sample’)*