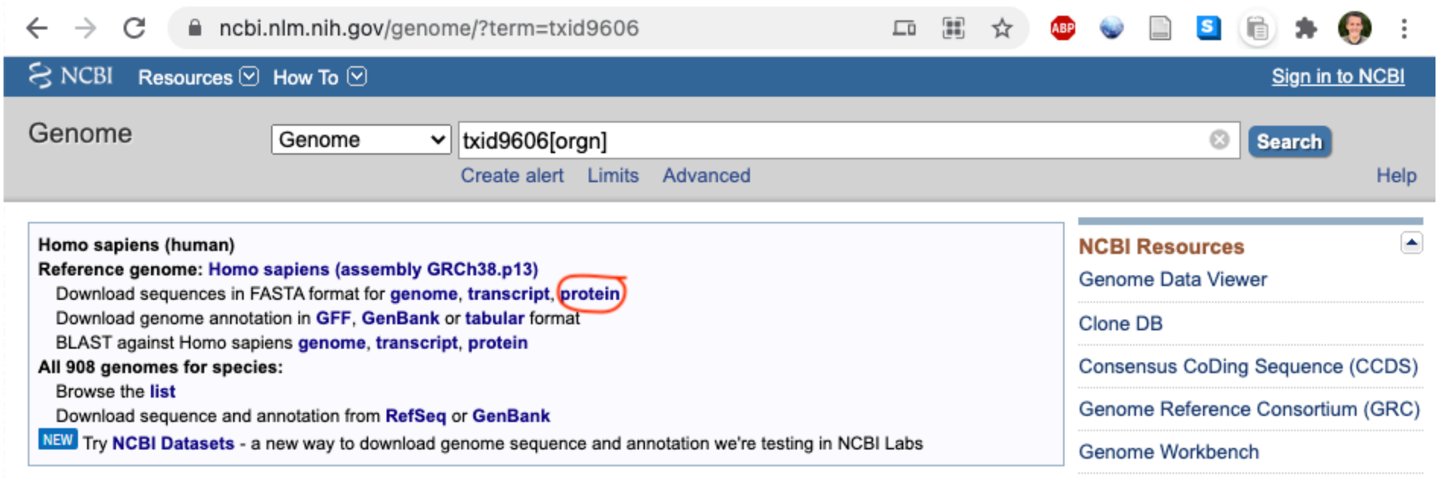
Environment for ASR in the Harms lab

Install necessary software

- Configure a scientific computing environment in python. If you have not done so already, I recommend miniconda. Instructions to set up the environment are here. You'll need jupyter, numpy, and pandas at a minimum.
- Copy contents of this directory to working location.
- Install biopython. On a terminal, type `conda install -c bioconda biopython`.
- Install muscle. On a terminal, type `conda install -c bioconda muscle`. (This may not work in windows. If it fails, you can download an installer from the linked muscle website).
- Install blast. On a terminal, type `conda install -c bioconda blast`. (This may not work in windows. If it fails, you can download an installer from the linked ncbi website).
- If you're on windows, install the ubuntu subsystem. This will allow you to easy use the bash tools we use in the tutorials. (If you're on macOS or linux, you already have those tools.)
- Install FigTree for viewing trees.
- Install AliView for editing alignments.
- We'll use raxml to generate our trees and ancestors. These can be installed locally, however, I would recommend running them on a high-performance computing environment.

Create a local copy of the human proteome for reverse BLASTing

1. In a browser, navigate to: https://www.ncbi.nlm.nih.gov/genome/?term=txid9606
2. Click the circled link below to download the human proteome as a zipped file (~20 Mb)



Homo sapiens (human)
Reference genome: **Homo sapiens (assembly GRCh38.p13)**
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Homo sapiens **genome**, **transcript**, **protein**
All 908 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**
NEW Try **NCBI Datasets** - a new way to download genome sequence and annotation we're testing in NCBI Labs

NCBI Resources
Genome Data Viewer
Clone DB
Consensus CoDing Sequence (CCDS)
Genome Reference Consortium (GRC)
Genome Workbench

3. Place the file in a working directory. Uncompress it and convert it into a BLAST database. Note, the name of the `.gz` and `.faa` file might be slightly different as the proteome versions on NCBI are continually updated. On the command line, run:

```
gunzip GCF_000001405.39_GRCh38.p13_protein.faa.gz
makeblastdb -in GCF_000001405.39_GRCh38.p13_protein.faa -dbtype prot -out GRCh38
```

This will create a set of files like `GRCh38.phr` and `GRCh38.pot` in your working directory. If you're pressed for space, you may delete the initial `.faa` file at this point.