Environment for ASR in the Harms lab

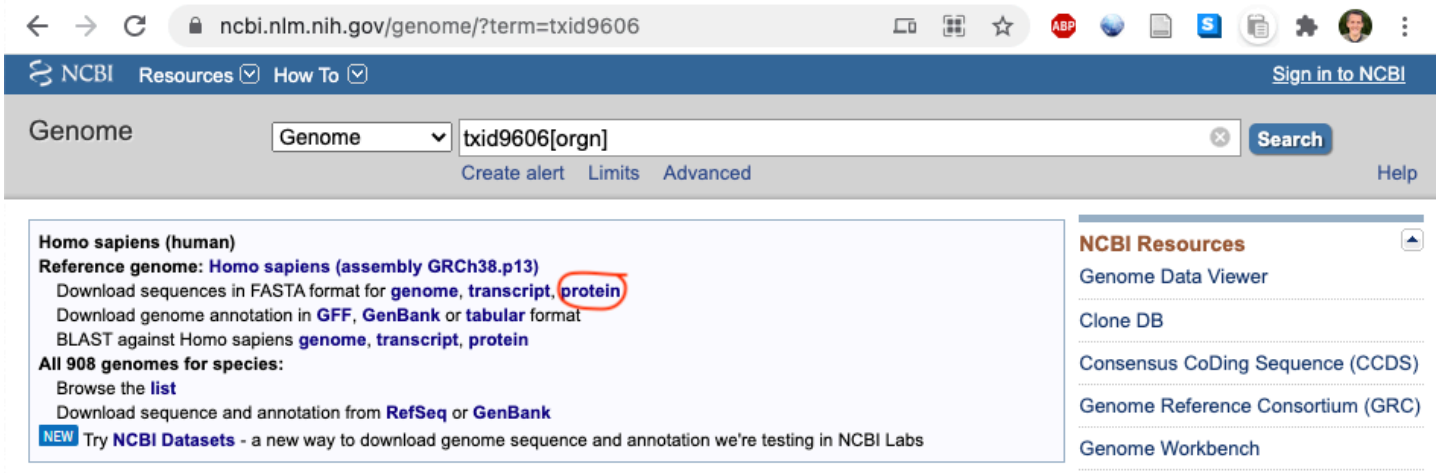
Install necessary software

The following should be run on your own computer.

- Install FigTree for viewing trees.
- Install AliView for editing alignments.
- 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.
- On linux and macOS:
  - Install biopython. On a terminal, type `conda install -c bioconda biopython`.
  - Install muscle. On a terminal, type `conda install -c bioconda muscle`.
  - Install blast. On a terminal, type `conda install -c bioconda blast`.
- On windows:
  - Install the ubuntu subsystem. This will allow you to easy use the bash tools we use in the tutorials.
  - Open a *conda* terminal. Type `conda install -c bioconda biopython`.
  - Try to install muscle via conda. Open a *conda terminal* and type `conda install -c bioconda muscle`. If this fails:
    - Download the latest windows muscle binary from https://www.drive5.com/muscle/downloads.htm.
    - Put it in a convenient location on your computer (maybe make a "programs" folder in your home folder?).
    - Rename the binary from something like `muscle3.8.31_i86win32.exe` to `muscle.exe`.
    - Update your windows path to point to the folder where you put muscle. Instructions are here.
  - Try to install blast via conda. Open a *conda terminal* and type `conda install -c bioconda blast`. If this fails:
    - Download the latest windows blast binary from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. (Look for win64 in the filename).
    - Run the installer. It should install the software in some place like `C:\Program Files\NCBI\blast-2.12.0+\bin\`
    - Update your windows path to point to this folder. Instructions are here.

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)



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. Open a terminal (macOS or linux) or an ubuntu subsystem terminal (windows) and run the following commands:

```
cd TO_WORKING_DIRECTORY
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.

Configure software on high-performance computing cluster

Install miniconda and python support

Open a terminal (macOS or linux) or an ubuntu subsystem terminal (windows). SSH into the cluster:

```
ssh UO_USER_NAME@talapas-ln1.uoregon.edu
```

Start an interactive session on a compute node by:

```
qsub -I -A harmslab
```

When the job starts, run the following code. It will ask various questions. Type "Y" or hit [Enter] as prompted. This will install miniconda, various scientific computing libraries, and the phylogenetics libaries `ete3` and `pastml`.

```
wget https://repo.anaconda.com/miniconda/Miniconda3-py39_4.9.2-Linux-x86_64.sh &&
bash Miniconda3-py39_4.9.2-Linux-x86_64.sh &&
conda install numpy scipy matplotlib pandas &&
pip install pastml ete3
```

Configure system to run raxml binary

SSH into the cluster. Then run the following three commands. This will update your $PATH so when you type a command it looks in the directory that has raxml installed.

```
cp .bashrc .bashrc.bak
echo "export PATH=/projects/harmslab/shared/standard-RAxML/:$PATH" >> .bashrc
source .bashrc
```

Make sure you can run raxml by typing:

```
raxmlHPC
```

This should spit out:

```
WARNING: The number of threads is currently set to 0
You can specify the number of threads to run via -T numberOfThreads
NumberOfThreads must be set to an integer value greater than 1

RAxML, will now set the number of threads automatically to 2 !


 Error, you must specify a model of substitution with the "-m" option
```

Copy scripts from repo to cluster

On your computer, open a terminal (macOS or linux) or ubuntu subsystem terminal (windows). Navigate to the the `asr-protocol` directory. Then run the command:

```
scp -r copy-to-hpc UO_USER_NAME@talapas-ln1.uoregon.edu:
```