Running OrthoFinder with protein sequences from the DataStore:

OrthoFinder is great and has a very detailed walkthrough at their GitHub repo:
https://github.com/davidemms/OrthoFinder

This walkthrough is just designed to get you running OrthoFinder in an Atmo instance:

1. First we need to get a pesky dependency installed. This dependency is FastTree and is necessary for OrthoFinder to start! Install by running the following code:

```
###
sudo apt-get install fasttree
###
```

- When prompted, put in your CyVerse password.

```
###

cd /usr/bin

sudo su

mv fasttree FastTree

exit

export PATH=$PATH:/usr/bin

cd /scratch

###
```

- Test that FastTree is installed and that you can access it by running

```
###

FastTree –h

###
```

2. Now let's grab those protein sequences that we exported to the DataStore earlier. If you haven't already initialized icommands, follow this tutorial (commands shown below):
https://learning.cyverse.org/projects/data_store_guide/en/latest/step2.html

```
$ iinit
One or more fields in your iRODS environment file (.irodsEnv) are
missing; please enter them.
Enter the host name (DNS) of the server to connect to: data.cyverse.org
Enter the port number: 1247
Enter your irods user name: #your_cyverse_username
Enter your irods zone: iplant
Those values will be added to your environment file (for use by
other i-commands) if the login succeeds.
```

```
Enter your current iRODS password: #your_cyverse_password
```

###

```
icd coge_data
iget yourproteinfile.fasta simplername.fasta
```

###

- Repeat the iget command for each fasta file, using copy and paste (makes life easier, but it’s easier to do it with the mouse by right clicking).

###

```
mkdir protein_sequences
orthofinder –t 8 –f protein_sequences
```

###

- This will run OrthoFinder and will take a bit of time. You can run this in the background if you choose by killing this process (control/command +C) and then relaunching with the following code, which will continue the process in the background

###

```
nohup orthofinder –t 8 –f protein_sequences >log.txt &
```

###

- You can check to see if your job is running by typing “htop” which will give you a visual of what is running on your instance. Press the F10 button to exit htop.
- Or you can view the log.txt file by running:

###

```
tail log.txt
```

###

- Once OrthoFinder is finished, check out the output info by navigating to the right folder:

###

```
cd protein_sequences/Orthofinder/Results_***/Comparative_Genomics_Statistics/
```

###

- Where *** is today’s date (i.e., Results_Jul23/)
- Here you can browse the different .tsv files – I would recommend:

###

```
nano Statistics_Overall.tsv
```

###

ctl +x to quit