**Phylogenetics Bootcamp—Building your gene family dataset**

Part 1: Using BLAST to get an initial set of putative gene family members

Getting Phytozome12:

**The following steps have already been done. These are for your benefit.**

1. Sign onto Apollo and then log on to six.
2. Use curl to login to Phytozome:

* curl 'https://signon.jgi.doe.gov/signon/create' --data-urlencode 'login=USER\_NAME' --data-urlencode 'password=USER\_PASSWORD' -c cookies > /dev/null

1. Get the list of all available files available on Phytozome12 and write them to a file.

* curl 'http://genome.jgi.doe.gov/ext-api/downloads/get-directory?organism=PhytozomeV10' -b cookies > files.xml

1. From this file, you want to get the primary transcript and peptide file names for all samples. To do this, use the grep command:

* grep ".cds\_primaryTranscriptOnly" files.xml > primaryTranscripts\_cds.txt
* grep "protein\_primaryTranscriptOnly" files.xml > primaryTranscripts\_protein.txt

1. The “&” symbol in these names needs to be changed from “&amp” to “&”.

* perl -pi.bak -e 's/\%amp\;/&/g' primaryTranscripts\*.txt

1. There are 72 (as of 04/17/2017) file names in each of these directories. In order to download these files, we need to do a little scripting.