**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=PhytozomeV12' -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
* If you plan on using this Perl one-liner to alter files in the future, use the website Rubular to help prepare your regular expression: <http://rubular.com/>

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.
2. The script bulk\_download\_JGI.pl is available in the Github repository <https://github.com/mrmckain/Teaching/tree/master/Phylogenomics_Bootcamp/Gene_Family_Building>.

* To run the script: perl bulk\_download\_JGI.pl primaryTranscripts\_protein.txt
* Use this with both files we created above.

1. Once downloaded, unzip the files so they can be combined into a single file.

* gunzip \*gz

1. To use these CDS/protein files to make a BLAST database, we first need to verify that the sequence names are the same. We will also want to remove anything after a space in the header since BLAST ignores this information.

* To do this, use the following command: grep ">" -m 1 \*fa
* This will report the first match of “>” in each file that ends with “fa”.
* The CDS and protein files will be back-to-back, so you can scroll through the output easily to see if they match.

1. Looking at the ID before the first space in the header, we can see that many of the CDS and protein files have matching names. We want to focus on those that do not:

* Acoerulea\_322\_v3.1
* Aoccidentale\_449\_v0.9
* Bdistachyon\_314\_v3.1
* BrapaFPsc\_277\_v1.3
* Bstacei\_316\_v1.1
* Bstricta\_278\_v1.2
* Cgrandiflora\_266\_v1.1
* Dsalina\_325\_v1.0
* Egrandis\_297\_v2.0
* Ghirsutum\_458\_v1.1
* Gmax\_275\_Wm82
* Kfedtschenkoi\_382\_v1.1
* Klaxiflora\_309\_v1.1
* Macuminata\_304\_v1
* Mesculenta\_305\_v6.1
* Mguttatus\_256\_v2.0
* Mpolymorpha\_320\_v3.1
* PdeltoidesWV94\_445\_v2.1
* Ppatens\_318\_v3.3
* Ppersica\_298\_v2.1
* Ptrichocarpa\_444\_v3.1
* Pvirgatum\_273\_v1.1
* Pvirgatum\_450\_v4.1
* Pvulgaris\_442\_v2.1
* Sbicolor\_313\_v3.1
* Sfallax\_310\_v0.5
* Spurpurea\_289\_v1.0
* Sviridis\_311\_v1.1
* Vcarteri\_317\_v2.1
* Zmays\_284\_Ensembl-18\_2010-01
* ZmaysPH207\_443\_v1.1
* Stuberosum\_448\_v4.03—EXCEPTIONALLY DIFFICULT

1. We can use the patterning of these make two Perl one-liners to fix issues. We are going to change the protein files to match the CDS file.

* perl –pi.bak –e ‘s/\.p(\s+)/$1/g’ \*protein\_primaryTranscriptOnly.fa
* perl –pi.bak –e ‘s/\_P(\d+)\s/\_T$1\s/g’ Zmays\*
* perl –pi.bak –e ‘s/ GSMUA\_Achr(\d+)+P/ GSMUA\_Achr$1T/g’ Macuminata\_304\_v1.protein\_primaryTranscriptOnly.fa

1. Non-matching IDs is not the only problem we have. The following have no unique identifiers to associate them with their origin:

* MpusillaCCMP1545\_228\_v3.0
* MspRCC299\_229\_v3.0
* Olucimarinus\_231\_v2.0
* Rcommunis\_119\_v0.1
* Smoellendorffii\_91\_v1.0

1. When we build our gene trees, it will benefit us if we have species name, or clear identifiers, for each sequence. To do this, we are using the script “fix\_names.sh”.

* To run, simply type: sh fix\_names.sh
* The script will look at each file that ends in “.fa”.

1. Now that we have fixed all of our samples, we can create a single, concatenated file:

* cat \*protein\*.fa > Full\_Phytozome\_04172017\_ protein\_primaryTranscriptOnly.fa
* cat \*cds\*.fa > Full\_Phytozome\_04172017\_ cds\_primaryTranscriptOnly.fa

15. Create a BLAST database for the protein files from Phytozome.