NCGAS Spring Workshop Day 2: KEGG Demo

There are two ways to get KEGG information relatively quickly. If you don't have annotation information, you can submit your amino acids to an online tool called "Ghost Koala". Let's try that:

First, transfer amino acids file to your home computer, using what you learned from Bhavya:

```
#For example:
scp ss93@carbonate.uits.iu.edu:~/Project_Workshop_Fall2018/final_assemblies/annotation/
    transcripts.main.renamed.aa .

#or use WinSCP, etc
```

Then, we just have to submit to GhostKoala (https://www.kegg.jp/ghostkoala/). Fill out the form and hist submit. Two notes on this - first, you can only have one job running per email, and second, you generally have to wait ~12-24hrs for this to run.

In the mean time, let's use the other way to get KEGG terms - from our annotation information! Let's use **awk** to print just the name of the gene and the KEGG term (columns 1 and 12), grab only the lines that have KEGG terms with **grep**, and then reformat them to just have the KEGG term using **sed**.

```
awk '{print $1,$12}' Report_wospaces.xls | grep "KO" | sed 's/KEGG.*`KO://g' > KEGG.list
```

Let's look at those results... hm, some of them appear to be duplicates. Why do you think that is?

To clean them up, we can use **sort** and **uniq**. **uniq** is a command that collapses duplicate lines:

```
sort KEGG.list | uniq > tmp
mv tmp KEGG.list
```

Okay, now we have that file, let's grab just the KEGG terms, with **awk**:

```
awk '{print $2}' KEGG.list > KEGG_only.list
```

And transfer it to our personal computers.

```
scp ss93@carbonate.uits.iu.edu:~/Project_Workshop_Fall2018/final_assemblies/annotation/
    KEGG_only.list .
```

We can use another KEGG tool, map pathway1, to visualize our transcripts in pathways:

Go to: https://www.kegg.jp/kegg/tool/map_pathway1.html
Add in KEGG_only.list -> show pathways

If we want to add colors in our pathways, we can do that with **awk** as well:

```
awk '{print $2, "yellow"}' KEGG.list > KEGG_only.list
```

Transfer the file again:

```
scp ss93@carbonate.uits.iu.edu:~/Project_Workshop_Fall2018/final_assemblies/annotation/
    KEGG_only.list .
```

And submit it to the map_pathway2 tool:

Go to https://www.kegg.jp/kegg/tool/map_pathway2.html
Add in KEGG_only.list -> show pathways

Note: I don’t have a good source for colors that exist in KEGG. If you have one that you’d like to share, please let me know!

Now, if we wanted to show a subset of data on top of our full data set, we can do that as well. Let's pretend our DE genes are listed in the DE.list file (found in annotation directory).

Let's grab the lines that match our DE transcripts from the KEGG list (getting their KEGG terms), and mark them to come up as green:

```
grep -f DE.list KEGG.list | awk '{print $2, "green"}' > DE.KEGG.list
```

The let's grab the rest of the lines (that don't match DE), mark them to come up as yellow, and add them to the same file:

```
grep -vf DE.list KEGG.list | awk '{print $2, "yellow"}' >> DE.KEGG.list
```

Transfer this again:

```
scp ss93@carbonate.uits.iu.edu:~/Project_Workshop_Fall2018/final_assemblies/annotation/
    DE.KEGG.list .
```

And re-enter it into the map_pathway2 tool!

Much of this is also reviewed on our KEGG Pathway blog: https://ncgas.org/Blog_Posts/Ghost%20Koala.php. In fact, our blog is an excellent thing to keep an eye on, as we post common analyses on here frequently!

What are some other subsets of data that might be interesting to visualize?