**EVO-DEVO**

Tls4259
Undergraduate Research 2024

# Triana San Miguel's Lab Notebook

## OVERVIEW

Our project is going to use two different computational approaches WGCNA (Weighted Gene Co-Expression Network Analysis) which identifies groups of genes that are co-expressed across various tissue types and ERC (Evolutionary Rate Covariation) which measures how similar the evolutionary rates of genes are across species.. We are using data from a mammalian organ development dataset which includes transcriptomes from seven organs across different species: human, rat,chicken,rabbit,opossum, rhesus, and mouse.

---

September 9, 2024 - September 13 2024

### Overview and Description of Goal

This week we were given a description of what WGCNA entailed and learned a few important concepts for our research. We opened up our WGCNA folders and began to look inside and view the data that was given to us. We viewed the large dataset inside which had all the species and their corresponding genes in each column. As we viewed the datasets we realized our next step was to copy and import the ERC data set which we then filtered out Human Ids to place in a table. Dr.Young informed us that there should be one corresponding protein ID for each species.

### Conclusions and Output

There were a lot of things that I learned this week. To begin, I learned how to create a path for the importation of the ERC data set to just create a path for filtered protein Ids
/stor/work/FRI_321G_RY_Spring2024/Summer2024/WGNCA_GeneEvolution/data/Clean_ProteinProteinIDs.csv.
We also learned that we should begin thinking about which way to convert the Protein Ids —> Gene Ids. We also needed to see proteins shared across species
/stor/work/FRI_321G_RY_Spring2024/Summer2024/WGNCA_GeneEvolution/data/ERC_ProteinProteinIDs.csv

**Next Steps**

The next step for next week is going to be to generate a metatable which uses the expression data from each of the species and then maybe try and separate the table separating each species by column. Each species column should have separate columns for tissue, developmental stage, and sample number to determine how many samples of each tissue there are for each species.

---

September 16,  2024-September 20, 2024

**Overview and Description of Goal**

This week I mainly focused on the goal of creating expression data for each of the species and creating a meta table based on those. Me and conner split this task and each did half of the species tables. Sal was tasked with trying to split functions and fix the tables but kept running into the issue of having issues with the periods as the column names. Anya was making folder for each species.TDSR (tissue, developmental stage, replicate)

**Conclusions and Output**

This week did have a lot of challenges that were slightly difficult to understand. For example, every time I would try to save the expression data it would save in my home directory instead of the folders, even though I would use the write.csv with the correct path. After I restarted R it ended up working. The following code ended up working library(dplyr) ,  ChickenExpression_data <-read.table("/stor/work/FRI_321G_RY_Spring2024/Summer2024/WGNCA_GeneEvolution/data/Chicken.CPM.txt"), write.csv(ChickenExpression_data, "ChickenExpression_data.csv", row.names = TRUE), and just substitute it each time with the different species. The species metadata is found /stor/work/FRI_321G_RY_Spring2024/Summer2024/WGNCA_GeneEvolution/data/Species_MetaData

**Next Steps**

There are many steps that we are going to work on for next week. We are going to begin to do more research on WGCNA and try to understand the scripts that were provided to us. We also need to clean out protein ids that have weird names like SynthaseSubunitSourceRGDSymbolAcc1311560". We also need to prepare for next weeks presentation.

---

September 23, 2024- September 27, 2024

**Overview and Description of Goal**

This week was focused on our presentation and preparing to show Dr. Young where we were in our project. We knew we were not tasked with presenting answers to our project but just showed how far along we were. On Tuesday we focused on the structure of our powerpoint and who was going to present which parts. In the mentor meeting we continued to focus on our slides. We made sure to show that only 167/2573 parts of the rows  had working protein ids on the metadata table.

**Conclusions and Output**

I did not run much code during this week and instead focused more on creating the slides and also getting ready to prepare them for the class. We did however get access to the entirety of amikahs code from last year that we were planning to alter to fit our own WGCNA that we were planning to run. We began to look at it and see that while some aspects were the same (The WGCNA progress) there were some aspects that were completely different such as some of the filtering processes that were used. Dr. Young also gave us some feedback after watching our presentation, we needed to match up gene IDs to protein IDs - make a table with gene name, gene ID, and protein it codes for (protein ID and or protein name) , we could also try Ensembl - put in all gene ids and then download the corresponding protein IDs Biomart r program - put in genes and then it will give corresponding protein IDs.

**Next Steps**

There were many next steps that we needed to add to be able to make our project go smoother and to make it more concise. To begin, we needed to add more goals to our slides and think about how we will be able to run WGCNA in the future by using amikahs code. We also needed to see what the Biomart function would be able to do and if we could run it in an efficient manner. We also had to begin thinking about the lack of Chicken Data and how this could affect our project in the future and just to speak to our collaborators.

---

September 30, 2024- October 4, 2024

**Overview and Description of Goal**

There were many tasks that were focused on this week. To begin, Conner was trying to utilize Biomart but it was not working with the amount of genes that we were working with, due to this we had to split up the gene of each species from the expression data and split it into 5,000 increments that would be able to fit into the Bio mart website. This caused me so much time and energy because there would be times that it would run and split the data correctly but other times it would not run and then again would not be put into the correct folder so I would have to then move the dataset manually which would take a long time as well due to the path sometimes not loading for me.

**Conclusions and Output**

The code that was finally able to work for me was as follows :

```
protein_ids <- human_gene_column  # Replace with your actual protein IDs vector or dataframe column

chunk_size <- 5000
num_chunks <- ceiling(length(protein_ids) / chunk_size)
for (i in 1:num_chunks) {
  start_index <- (i - 1) * chunk_size + 1
  end_index <- min(i * chunk_size, length(protein_ids))
  protein_chunk <- protein_ids[start_index:end_index]
  write.csv(protein_chunk, paste0("protein_subset_", i, ".csv"), row.names = FALSE)
}
```

This code ended up working and I used it on all the species. By doing this Conner was finally going to be able to utilize the Biomart website and match up the genes to their corresponding protein ids. The path to the data is as follows
/stor/work/FRI_321G_RY_Spring2024/Summer2024/WGNCA_GeneEvolution/data/ENSM_only/Human/Human_1_to_5000/
We use this path for every species and just change the number of the gene count

**Next Steps**
The steps for next week are going to be try to finish up Bio mart and get all the corresponding Protein Ids for each of the genes. This has been very difficult because even after splitting all the genes Bio mart has still been very hard to use and has not been able to be installed on anyone's computer and the website only works on Conners computer. Sal is also planning to finish up looking at amikahs code.

October 7, 2024- October 11, 2024
**Overview and Description of Goal**
This week after I finished separating all of the gene expression sets into 5,000 and then I began trying to get Bio mart to work on my computer. I first tried to get it to work on my website but it just continued to crash. However I tried to install it using the R studio package and after a while it began to work. This allowed me to begin adding all 5,000 sections using the ensembl genes and create 5,000 increments of the peptide protein matching to the gene.

**Conclusion and Output**
In conclusion, after I advised Dr. Young about the Bio mart now working on my computer she taught me how to be able to simply just run the entire expression data set to get the matching peptide id. This was after I already had separated them of course but it still allowed me to be able to learn and now we could get one metadata set with the entire gene to protein code instead of

having 5,000 increments. This saved us a lot of time and we were able to get Bio mart over and done with

**Next Steps**

Our next steps are going to be trying to work on more Chicken id orthologs  and Dr. Young said that we could begin working on those next week. We are also about to finish up analyzing all the code amikah gave us and finally finish up understanding the WGCNA code

---

October 14, 2024 - October 18, 2024

**Overview and Description of Goal**

This week, I began trying to run WGCNA for the first time, focusing on using the gene expression data and metadata we generated previously. The main objective was to try to start a species WGNCAs. However, I encountered several challenges in the process, including problems with data normalization, module clustering, and memory allocation errors in R.

**Conclusion and Output**

In conclusion I definitely need to try and learn more about the process of filtering, count matrices and also learn more about each step of WGCNA itself. Since amikahs code is similar but also very different from ours I need to prepare the code in a different manner which might be slightly difficult but manageable.

**Next Steps**

My next steps are going to be to try and talk to the team to make sure we are all on the same page with the code. I am also going to try and talk to amikah more and see if I can figure out some places where I am going wrong.

---

October 21 2024- October 25

**Overview and Description of Goal**

This week, I continued to try and run WGCNA and we also started to ask what would be the best way to run WGCNA in regards to species within tissue and if we would need to start looking at any type of pairwise function. We also started to see if there were genes IDs that were expressed more than one time.

**Conclusion and Output**

In conclusion I have now realized that using chat GbT for code input and output had been a mistake due to the errors that I kept on getting and it not understanding how to load in some libraries the correct way so now I need to just completely focus on amikahs code and just ask her for help instead of trying to do it with chat

**Next Steps**
The Next steps would try and understand where I went wrong with chat GBTs code and instead put more effort into understanding each of amikahs lines of code because hers will be more useful in the long run.

---

## October 28, 2024 – November 1, 2024

**Overview and Description of Goal**

This week, I continued making progress with WGCNA but encountered some unexpected challenges. My initial goal was to run WGCNA smoothly and refine the analysis of gene expression within tissue types across species. However, issues arose during the process that required troubleshooting and re-evaluating my approach.While running WGCNA, I realized that my metadata had been converted into a non-numerical matrix. This error required me to take a step back and reformat the metadata into a numerical matrix for proper analysis. Upon resolving this, I discovered a deeper issue my actual dataset was incorrect. This necessitated merging the correct datasets to ensure accurate results moving forward.

**Conclusion and Output**

Although these setbacks delayed the progress I initially planned, they provided valuable insight into the importance of dataset integrity and proper formatting. Correcting these foundational issues will ensure more reliable results as I continue the analysis.

**Next Steps**

- Verify the integrity of the newly merged dataset before re-running WGCNA.
- Cross-check metadata to prevent further formatting errors.
- Seek assistance from teammates or resources as needed to ensure accuracy and efficiency in the analysis process.

---

## November 4, 2024 – November 8, 2024

**Overview and Description of Goal**

This week, I was finally able to run WGCNA without any major issues. While working through the code, I realized that I had misunderstood an important aspect of our dataset—it was already normalized. Because of this, I had to remove a lot of the filtering steps from Amikah's code that were unnecessary for our analysis. Once I made these adjustments, the process ran smoothly, and I was able to generate results for the first time.

**Conclusions and Output**

This week was a big learning experience. I learned to be more cautious about understanding the preprocessing status of our dataset before adding extra steps that could complicate the workflow. Removing unnecessary steps from the pipeline simplified the process and allowed WGCNA to run as expected. Now, I have a clearer path for interpreting the results.

**Next Steps**

For next week, my goal is to begin analyzing and trying to understand the results from WGCNA. I plan to look into the generated modules, examine their biological significance, and see how they align with our research questions. I'll also make sure to document each step for reference and potential troubleshooting.

---

## November 11, 2024 – November 15, 2024

**Overview and Description of Goal**

This week, I focused on trying to interpret the results generated from WGCNA. Connor and I spent time analyzing the box plots from our data, but we both ran into some issues. The box plots showed very slight positive or negative trends, which seemed unusual and didn't align with what we expected. This has made us rethink some of our initial analysis steps and focus on troubleshooting these issues.

**Conclusions and Output**

Although we weren't able to fully interpret our results this week, we identified a potential problem with how the data was being visualized in the box plots. This will need further investigation to determine whether it's an issue with the data itself, the code, or the interpretation process. It's clear that we'll need additional input to resolve this.

**Next Steps**

- Meet with Dr. Young to discuss the box plot issues and clarify what might be causing the unexpected trends.

- Revisit the data preprocessing steps to ensure there aren’t underlying issues affecting the results.
- Continue collaborating with Connor to refine our approach to analyzing and interpreting the WGCNA results.