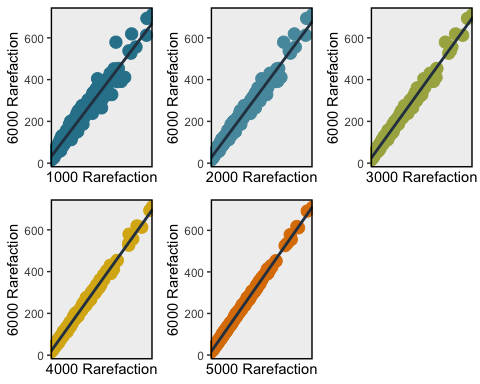
Microbiome Analyses Walkthrough

This is an [R Markdown](http://rmarkdown.rstudio.com) walkthrough of the data analyses associated with **Fuess et al. 2020a**. Follow the step by step instructions below to go from supplied data to results presented in **Fuess et al. 2020a**. Source codes and raw data are avilable via [GitHub](https://github.com/lfuess).

## Diversity Analyses

We will start with the analyses of correlations between diversity and gene expression. This uses data from previously published TagSeq (**Fuess et al. 2020b**) and microbiome (**Ling et al. 2020**) analyses. All necessary data is included in this repository.

To start our analyses, we first need to choose an appropriate rarefaction level for all later analyses. We will identify the lowest rarefaction at which rank orders are generally preserved. You will use the code **Rarefaction\_Graphs.R** for this portion



Looking at these graphs, the 2000 Rarefaction level will work well for our purposes. Next we move in to generating files for correlations and WGCNA. to do this we are going to use the following file **Diversity\_InitialMatrix\_Generation.R**. There will be no graphical output from these, just files that we later use for correlations, etc. Outputted files include:

* **DiversityData2000.csv**, a file with sample names and alpha diversity means for all 387 samples shared between the two data sets; there will be N/As in here for diversity in the case of samples that had microbiome data, but did not meet cutoff for calculating diversity at a rarefaction level of 2000. This will not affect correlations.
* 3 diversity correlation matrix files, subdivided into chunks to make correlation analyses easier. You can increase or decrease these chunks as needed. A highpowered computer will be needed for further analyses of these.

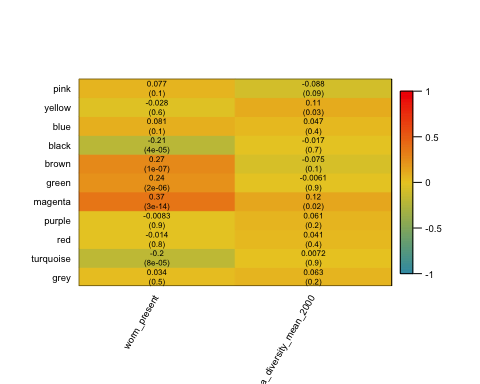
### WGCNA Network Build

Now the data splits into two analyses: the correlations and the WGCNA analyses. We will start with the WGCNA analysis. WGCNA will build a co-expression network that we can use to correlate to traits of interest (diversity, and later family proportions). We’ll make the network, export it, and then correlate the network to infection and alpha diversity. When you finish you’ll have two new files that we’ll use later:

* **Stickeblack-all-networkConstruction.RData**, data for the network build. You’ll need this for the correlation we do next and for family correlations.
* **WGCNA\_Diversity\_Corrs.csv**, a file with all the info about the WGCNA network modules; lists modules, what module they’re in, and membership to all network modules

We start by building a network using the code: **WGCNA\_NetworkBuild.R**. We can then look for correlations between that network and microbial diversity using the code: **Diversity\_WGNCA\_Correlations.R**.

### Correlation Tombstone Plot



Two modules are correlated to Alpha Diversity: **yellow** and **magenta**. Next we need to figure out what is in each of these two modules. Luckily we have a sript for parsing that **WGCNA\_Diversity\_Corrs.csv** file. The **Diversity\_WGCNA\_GOMWU\_Parse.R** script will output info for GOMWU (gene ontology analyses) and a file with a list of all the contigs in this module. That’s what we use to describe each one. So from this you end up with four files:

* **magenta\_transcripts.csv**, list of transcripts in the purple module
* **yellow\_transcripts.csv**, list of transcripts in the yellow module
* **magenta\_kme\_GOMWU.csv**, input for GOMWU for the purple module
* **yellow\_kme\_GOMWU.csv**, input for GOMWU for the purple module

Ok, so now we want to use gene ontology to characterize these two modules. We’ll start with the mgenta module and use the GO-MWU R code for this (need to insert referene here). I’ve included the annoation data from **Fuess et al 2020b** which we’ll use for this script as well as all the relevant GO-MWU scripts in the GitHub Repository. Here are the results for the purple module, biological processes:

## 40 GO terms at 10% FDR

## GO terms dispayed: 23   
## "Good genes" accounted for: 25 out of 33 ( 76% )

You can edit those scripts to analyze any module, or cellular component and molecular function enrichment as well if you so desire. But generally speaking, this wraps up the WGCNA portion of the diversity data. Now we will move on to correlations.

### Diversity Correlation Analyses

The code for running the correlation analyses is in the file **DivMatrixServer\_JustP.R**. This script is written for three subsets, but can be adjusted for more or less. It will return a matrix of p-values for each of the subsets of the matrix. I highly recommend doing this on as powerful a computer or server. The scale of these matrixes are massive.

After running the massive correlation matrix we can then parse out the data. We do this using the script: **ProcessingGeneCorrs\_Diversity.R**

This will return the following files:

* **SigCorrResults.csv**, list of all significant genes with the p-vaules included

Next we need to generate a matrix for calculating Tau values for each of these significant genes. We will use the Tau values for further gene ontology analyses. The script that generates the Tau matrix is: **Diversity\_TauMatrix\_Generation.R**

This will return the following files:

* **Diversity\_Tau\_Matrix.csv**, Matrix which can be analyzed to obtain tau values of correlation between significant genes and alpha diversity

Next we want to calculate tau values for every significantly correlated gene. Again I did this on a high powered computer and I used this script: **DivMatrixServer\_JustTau.R**. It returns the following files:

* **Diversity\_Corr\_results\_tau.csv**, Matrix of tau values for each of the 1931 significant genes and alpha diversity.

Moving on, we now need to parse this file and generate input for GO-MWU so we can figure out what biological processes are significantly enriched within this group of genes which are significantly correlated to alpha diversity of the microbiome.

To do this we use this script: **Diversity\_Tau\_Results\_Processing.R**. This will generate the following files:

* **Diversity\_SigGenes\_TauValues.csv**, List of all 1931 significant genes and their respective tau values.
* **Diversity\_Taus\_GOMWU.csv**, input for GOMWU analyses of gene ontology enrichment of the genes which are singificantly correlated to diversity.

### Diversity Correlation GO-MWU Analyses

Last step for diversity section! We need to do gene ontology enrichment of the genes which are significantly correlated to microbiome alpha diversity. We’ll just do biological processes, but the script is easily modified to do cell component and molecular funciton. To do this, just run the folowing script: **GO\_MWU\_Diverstiy\_Taus.R**

## Family-Level Analyses

Moving on to analyses for the second half of the paper, where we shifted gears and looked at correspondance between gene expression and proportion of specific microbial families. To do this, we will start with the correlation analyses so we can narrow down the list of families to those which are correlated to the highest number of genes. To start we will first take the raw OTU count data from **Ling et al. 2020** and calculate out proportions for each family. Then we will merge this with our expression data to create our correlation matrixes. Again we subset these, 5 subsets this type, though that is easily adjusted. We will do this using the following script: **FamilyCorrMatrixPrep.R**

This will output the following files: \* **FamilyPropData.csv**, Matrix of all samples with infection status (for WGCNA matrix), and proportion of all detected microbiome families. \* 6 Correlation matrixes each with roughly ~5000 genes, and data for all 306 microbial families.

### Family Correlation Analyses

Ok so the next step is to calculate the p values for all correlations. Here again a high powered computer comes in handy! The script to run this is: **FamilyMatrixServer\_JustP.R**. This will return 6 files with correlation results from each subset that we can then parse. For families we are going to parse the results and then just take the top 5% of families (those which are correlated to the most genes) for futher analyses (tau calculations and WGCNA).

**Side Note**: If you’re curious about the analyses we ran to justify taking the top 5% of families, you can check out the code **Distribution\_Tests.R** where we investigate the distribution of p-values, etc.

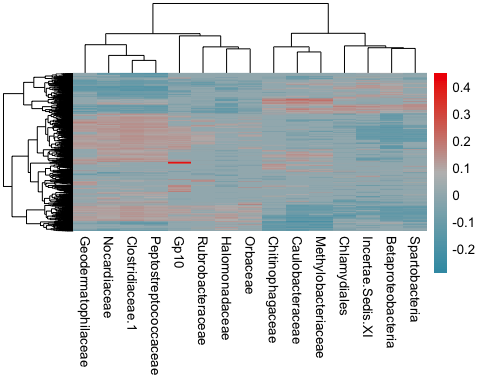
To parse the files and obtain the matrixes for generating tau values for these 15 (top 5%) families, you can run the following script: **Fam\_CorrRes\_Processing.R**.

This will output the following files: \* **top5families.csv**, List of the top 5% most correlated families (15 total) \* 15 files, one for each family, with list of significant genes correlated to that family \* 15 matrixes, one for each family, for generating tau values for all genes which are significantly correlated to those families

Again, we move to a high powered computer to calculate these tau values. You can use the following file, which will automatically generate tau values for all your genes in each of the 15 files: **Families\_Tau\_Server.R**.

Next we need to process the Tau results for GOMWU analysis. We do this using the file: **Family\_Tau\_Processing\_GOMWU.R** This will output 15 files, one for each significant family, that can then be input into the respective GO-MWU files to obtain information on gene ontology enrichement.

We can also use the output of these files to make Figure 3 from the paper. The script is **Family\_Tau\_Heatmap.R**.



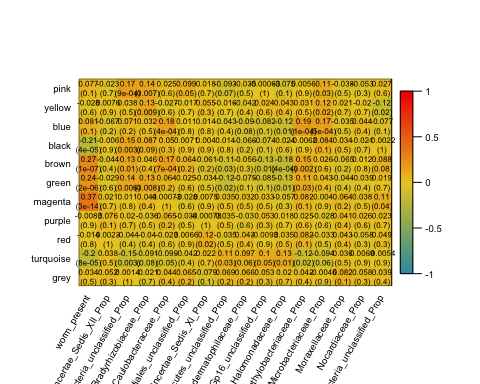
### Family Correlation GO Analysis

So after that we run gene ontology enrichment analyses on each of the 15 families. All the scripts and results can be found in the GitHub directory.

The output from gene ontology enrichment analyses of each family can be parased with the file **GOMWU\_Results\_Parse.R**. We then hand annotate the files and create the file **Family\_GO\_Bar.csv** which can be used to make Figure 4 from the paper using the following code: **Family\_GOs\_Sum\_Plot.R**

### WGCNA Analysis for Families + GOMWU

Finally, we want to run WGCNA analysis of the correlations between family proportions and the modules from the network we built earlier. This is fairly simple as the network has already been constructed. We just load the previous RData which has the network structure, and run a quick correlation to generate the tombstone plot.



So every module **except** the purple is correlated to at least one family. We can look at what’s in these modules using the same scripts as above ( **Diversity\_WGCNA\_GOMWU\_Parse.R**) In the end you will end up with the following two files for every module: **1** list of transcripts in the module, and **2** input for GOMWU for the module. You can use the general GOMWU script from above to investigate gene ontology enrichment for these additional modules as desired.

This wraps up the WGCNA portion of the family data and our analyses in general!!!

### SEM Analysis for Co-varriates

The last part of our analysis is to check the robustness of our correlations to covarriates. We do this using path analysis that accounts for several factors including fish sex, mass, infection status, etc. Most of this analysis is contained in two scripts each for Diversity and Family.

First we use this script to create the necessary csv file and automatically generate a large script that will run SEM analysis on all genes signficiantly correlated to diversity.

We can then run the resulting script to get our SEM results.

And then we use this script (**Path\_Analysis\_Results\_Parse.R**) to parse the results and generate a list of the genes which remain signficant for covarriance between microbiome diversity and transcript abundance following SEM analysis. Note it may need to be tweaked depdendent on what output you’re parsing (family or diversity).

Finally we do the same for each family, using similar scripts; I’ll show an example here for Spartobacteria

And then run the resulting code and parse the results following the script listed above.

That’s it for our data. Please feel free to contact me with any further questions you may have regarding the data or code listed here.