Final_Writeup_Bi610

Dane Dewees

December 8, 2017

Statistical Methods

File 1 - Survival

Using statistical packages in R, we parsed through 4 seperate datasets that presented raw data on how the presence of microbiota as well as the genetic background/variation of each host may influence phenotypic variation among the larval stickleback. The *first file contains information about whether each studied fish survived to 14 days post fertilization (or died)*. Utalizing a generalized linear model (glm) for the response variable of *survival*, it gave us the option of a binary aproach (yes or no) to test if assumptions were met. Various observations were viewed as independent.

File 2 - Sex

For the second file, which *contained information about whether each experimental fish was female or male*, we also used glm in terms of analyses for the distribution and inlfuence of sex. The sex ratio, which was measured by the response variable in the sex data being binary (male or female), we assumed the normality assumption was not met. The most basic generalized linear models were fit intially to predict *Sex* given *Population* and *Microbiota*.

File 3 - TAGs

For the third file, which *contained HPLC measurements of triacylglyceride (TAG) concentration in ug/mg for whole bodies of 14 dpf stickleback larvae.*. There was a temporal factor in this dataset that showed that the 14th day post-fertilization, there were sacrificed and processed for lipid concentration given the predictor variables assed above. The three variables were of a balance design, and thus we ran a Model I Fixed effects factorial ANOVA. Determining that *Sex* was the predictor variable causing the unbalanced nature of the experiment, we removed it from dataset (following confirmation using power analysis on the *Sex* variable given the outcome of insignificant results for both *Population* and *Microbiota* on the variable in the *sex_data_set*).

File 4 - Gene count

For the fourth and final file, which contained *transcript abundance measures from RNA-seq data for individual 14 dpf fish intestines (with data for 300 genes expressed in the gut were included).*. Two different methods of non-metric multidimensional analysis was done on this dataset (PCoA and nMDS). Disimilarity matices were generated and assessed using the Bray-Curtis distance. Population, Microbiota, and Sex were of interest here. Given the dimensional data set, we ran multidivariate statistics to determine the distribution of gene expression given the categorical variables stated above. PCoA & nMDS analysis was ran for the data set in order to ensure the accurecy of gene count expression.

Results

Objective from the four datasets was to quantify gene expression from the stickleback fish in ~300 genes based on the differences in both *Populations* and *Microbiotas*. Isolating the other factors that may be infleuntial to expression levels were mortality rates, lipid concentration, and sex were used to better understand the process of how Microbiota influences larval stickleback biology.

Microbiota showed statistical significance with regards to the effect on *Survived* or not. (z = -2.235, p = 0.0254). Using the Chi-squared statistic supported the effect of *Microbiota* on *Survived* (X^2 = 5.2407, p = 0.02206). *Population* did not show a direct influence on the effects of survival rate (Supp. Table 1) in tested samples (z = 0.378, p= 0.705). Again, using the Chi-squared statistic helped support that there was not enough of an effect from *Population* on *Survived* (X^2 = 0.14343, p = 0.7049). The interaction between *Microbiota* and *Population* did not show any significance with regards to survival (X^2 = 0.2455, p = 0.6203) when referencing tables **supplemental tables 1:3a-c.**

Population, *Microbiota*, and the interaction between the two is not statistically significant in terms of having an effect on *Sex* (Z = -0.675, P = 0.499), (Z = 1.118, p = 0.263), (Z = 1.121, p = 0.262) respectively when referencing **supplemental tables 4a-d.**

Highlighting a clear distribution when referencing the significance of both *Population* and *Microbiota* with regards to TAGs freqeuncy is shown in both **Figures 1&2**. Each reperesent the difference across population's infleunce and microbiota type inlfuence on Tags abundance. **Figure 2** highlights the linear combinations of Population and Microbiota. This distinction between *Populations ~ Microbiota* in comparison to germ free shows the distribution across said factors. The interaction effects model showed a statistically significant effect on lipid content (F = 230.11, p < 2e-16). The interaction plot in **Figure 3** shows the comparison of both the main effects and interaction effects models side by side as they relate to the response variable of interest. *Population* and *Microbiota* seperately showed statistical significance with TAG concentration (F = 133.44, p < 2e-16, and F = 57.63, p = 6.64e-11). The effects gave an R2 of 0.841, which shows that 84.1% of the variation in TAGs concentration may be due to the interaction effects model of *Population ~ Microbiota*.All three of the terms in the interaction effects model have statistically signficant effects on TAGs expression (**Sup. Table 1**). One can conclude that the interaction between population and microbiota type (germ-free) has and influence on the influx of lipid expression (TAGs) in stickleback fish (**Figure 3**).

Figures 4-6 illustrate the PCoA charts associated with *Population*, *Microbiota*, and the interaction between the two in ordination plots. **Figure 4** shows seperate clustering with regards to Boostrap (Bt) & Rabbit Slough (RS) *Population*. BS seem to cluster closer in distance than RS. **Figure 5** is similar but shows *Microbiota* treatment. Distance is more variable and there does seem to be a bit of overlap given the estimation of PCoA 1 and PCoA 2. **Figure 6** shows the interaction ordination plot between the two factors and highlights specific trends.

Referencing **Figure 7** shows the Red1 dots corresponding to the Bt *Population*, and slateblue2 corresponding to the RS *Population*. There are two distinct clusters but it seems that the clustering is not influenced based on *Population* alone. As you can see from **Figures 6-8**, the nMDS analysis via spiderplots highlights the clustering variation between groups while still maintaining similar structure to the PCoA chart (**Figure 6**) shown before. The distinct clustering in each plot was of course supported by significant effect values on gene expression from the use of permutation tests on the nMDS

dimensions. *Microbiota* showed 8.8% variaion in the interaction plot as well as in gene expression (F = 5.39, p = 0.0019). *Population* seemed to also show similar trends as *Microbiota* with regards to gene expression infleunce (F = 5.39, p = 0.0019). Localizing the predictor variables gave infleunce to over ~15% of the variation in gene expression. **Table 1** shows that Microbiota, Population, and the interaction between the two showed statisitcal signifiance with having an effect on the derived variables nMDS 1 and nMDS 2 (Pr(>F) 0.000999, 0.0000999, and 0.028971 respectively). This also showed the disimilarity effects when looking at the interaction between the two factors. The disimilarity came out to be roughly 8.8% for *Microbioata* and 15% for *Population*.

Figure 1

```
grid.arrange(Pop_ggplot, Micro_ggplot, ncol=2, nrow=1)
```

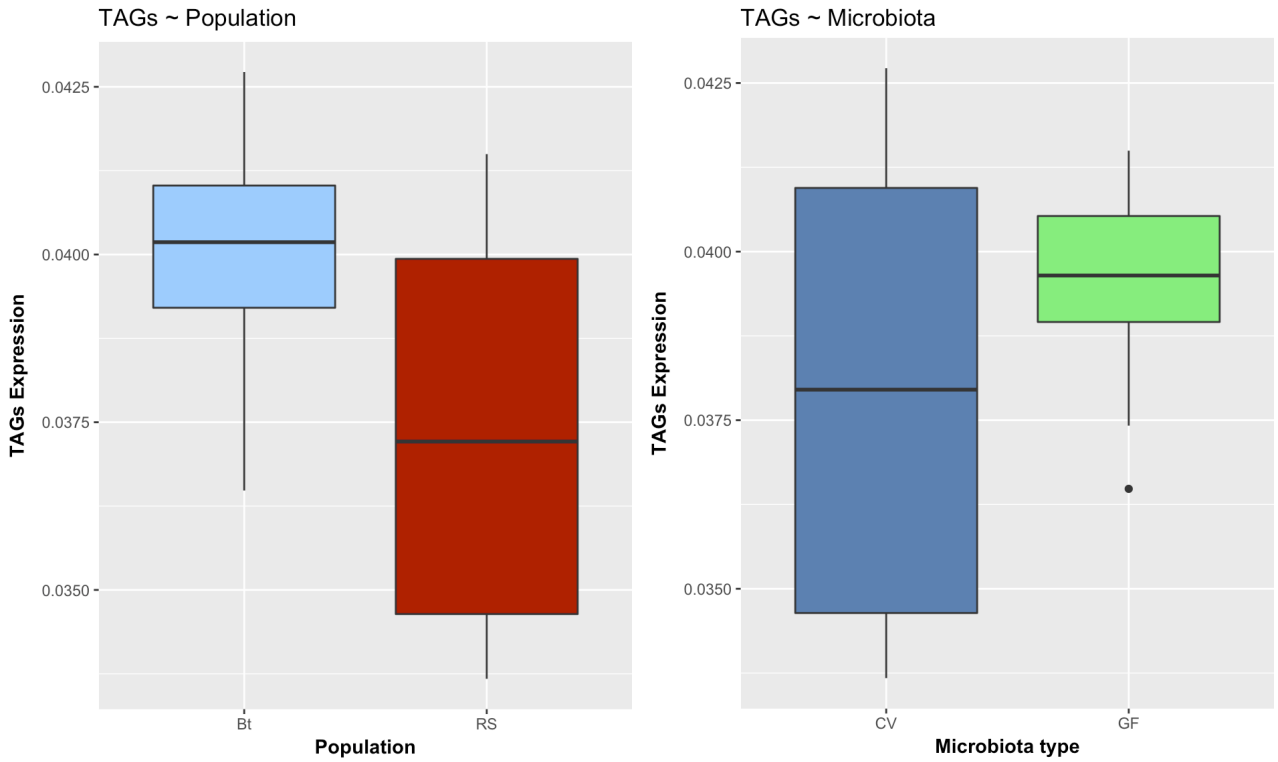
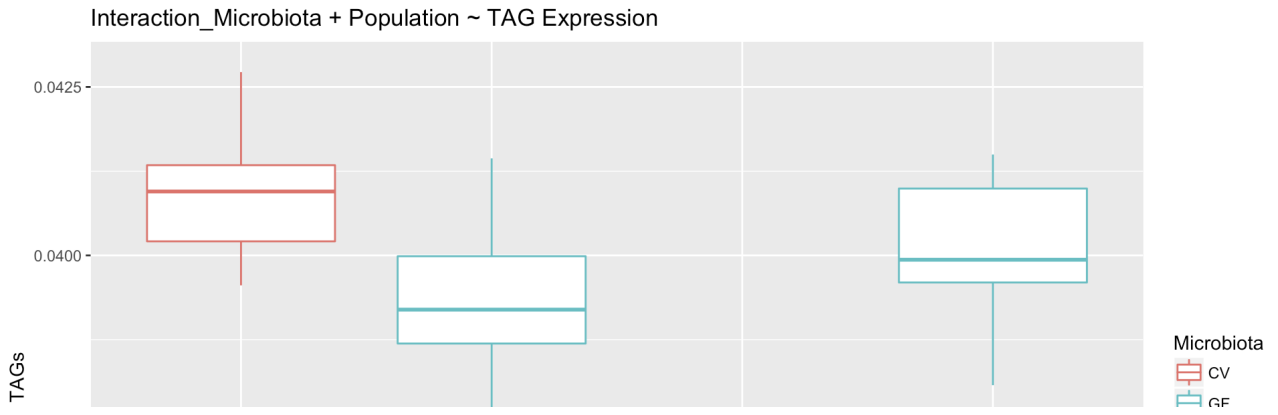
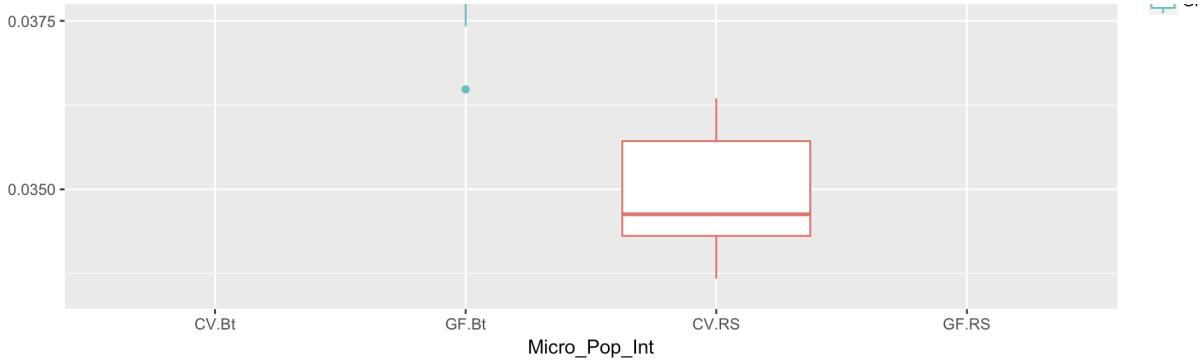


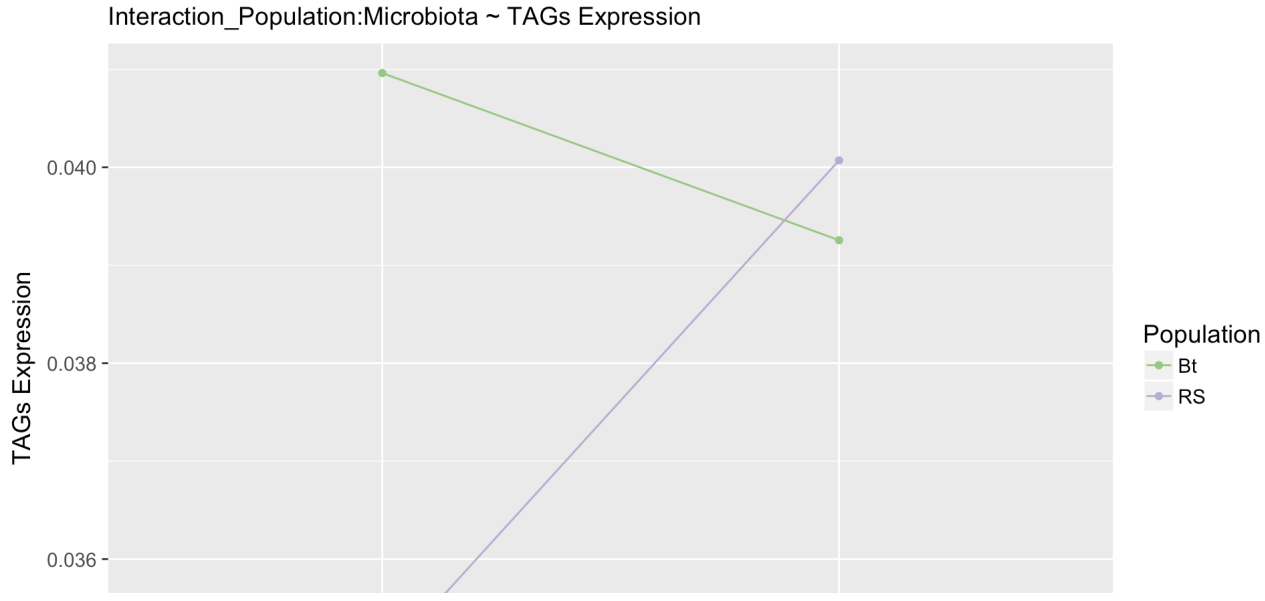
Figure 2

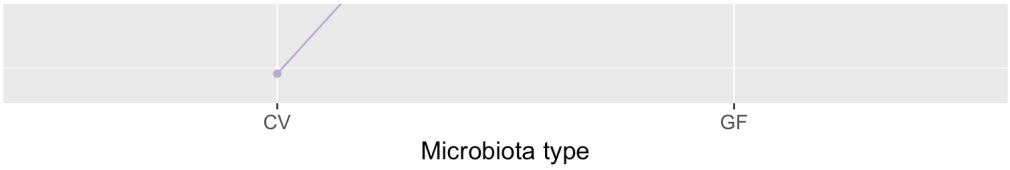





```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                    TAGs Expression
##                           Main Effects        Interaction Effects
##                               (1)                    (2)
## ---------------------------------------------------------------
## (Intercept)             0.039*** (0.0004)       0.041*** (0.0002)
## Population              -0.003*** (0.0004)     -0.006*** (0.0003)
## Microbiota              0.002*** (0.0004)      -0.002*** (0.0003)
## Population:Microbiota                           0.007*** (0.0005)
## ---------------------------------------------------------------
## Observations                  80                     80
## R2                           0.384                  0.847
## Adjusted R2                  0.368                  0.841
## F Statistic           24.032*** (df = 2; 77) 140.394*** (df = 3; 76)
## ===============================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```

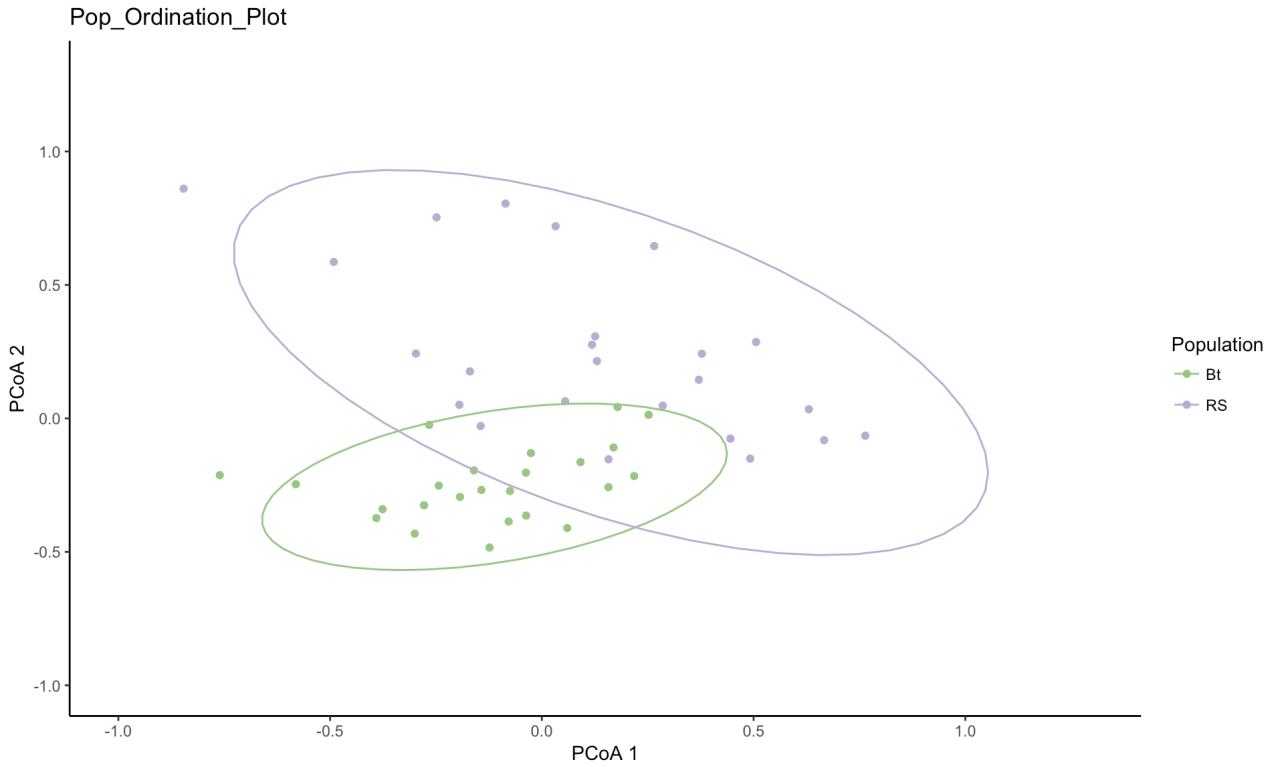
Figure 3





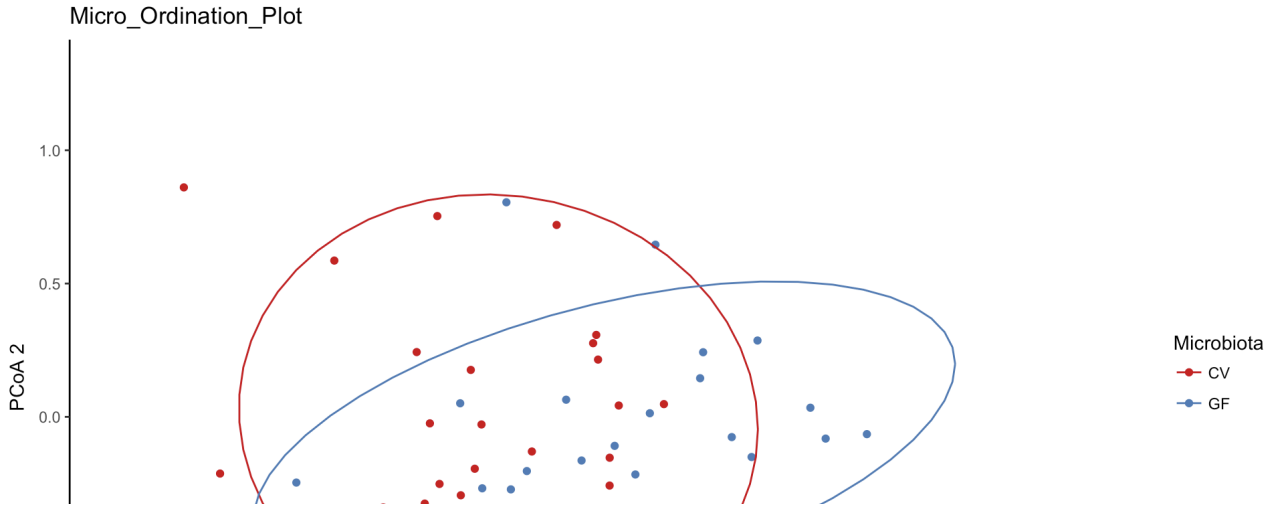
RNA_seq data

Figure 4



Two distinct clusters, bootlake `Population1 seem to cluster closer in distance than Rabbit Slough.

Figure 5



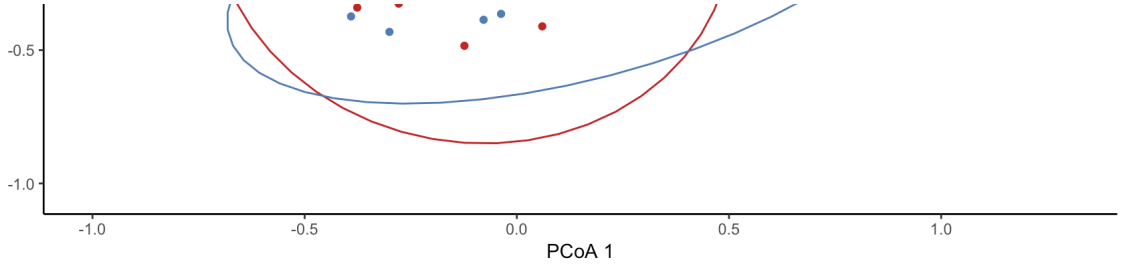
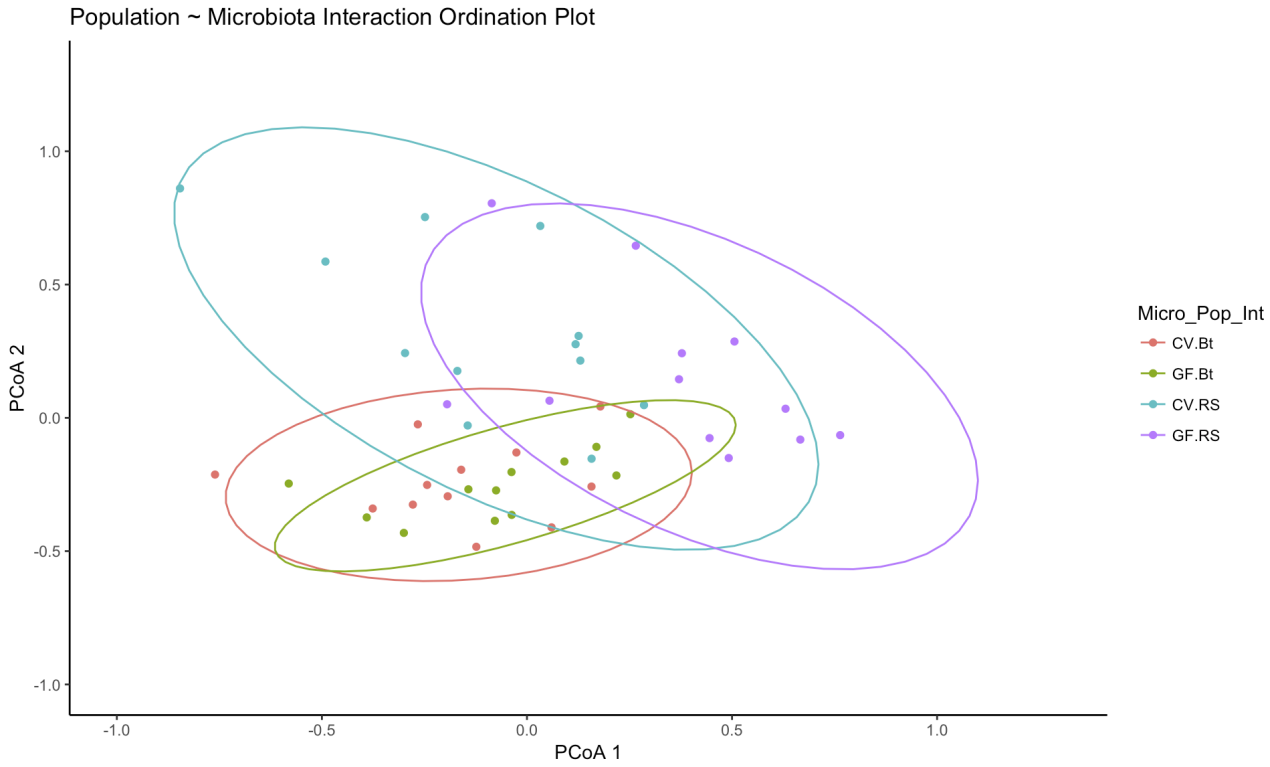
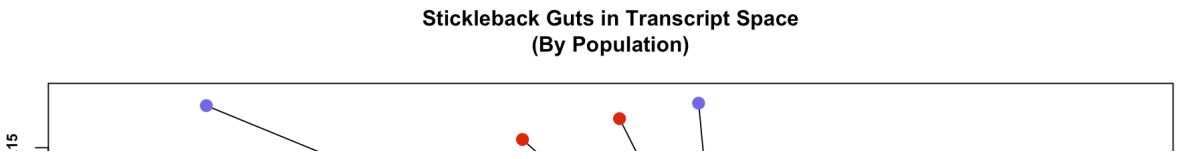


Figure 6




```
## initial  value 14.051550
## iter   5 value 10.786424
## iter  10 value 10.296052
## iter  15 value 9.949206
## iter  20 value 9.725984
## iter  20 value 9.721272
## iter  20 value 9.716605
## final  value 9.716605
## converged
```

Figure 7



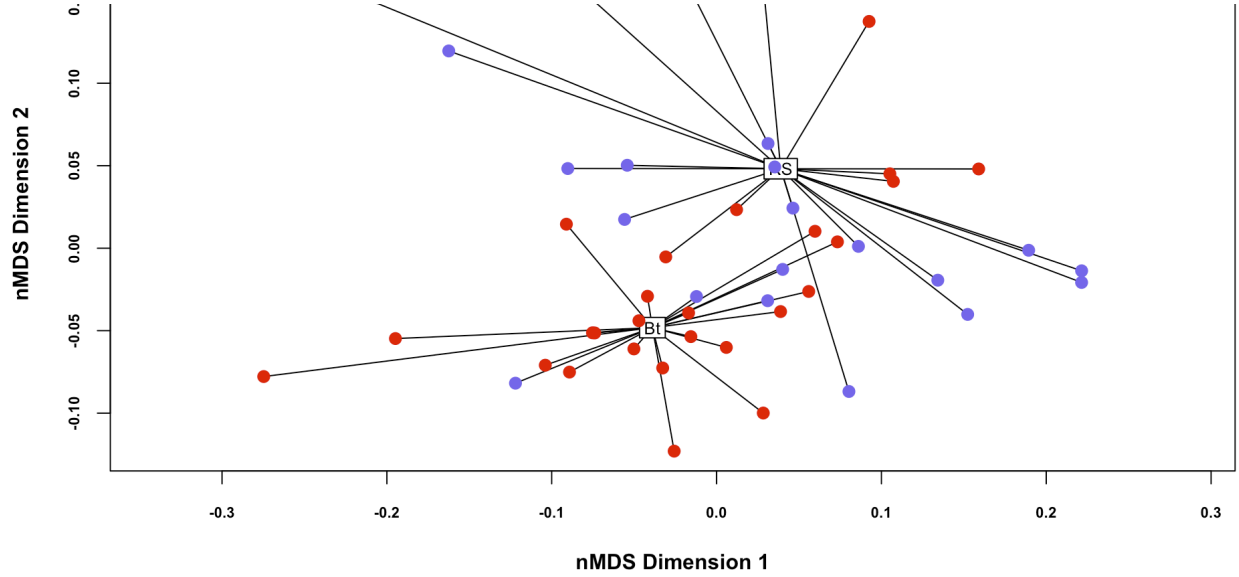


Figure 8

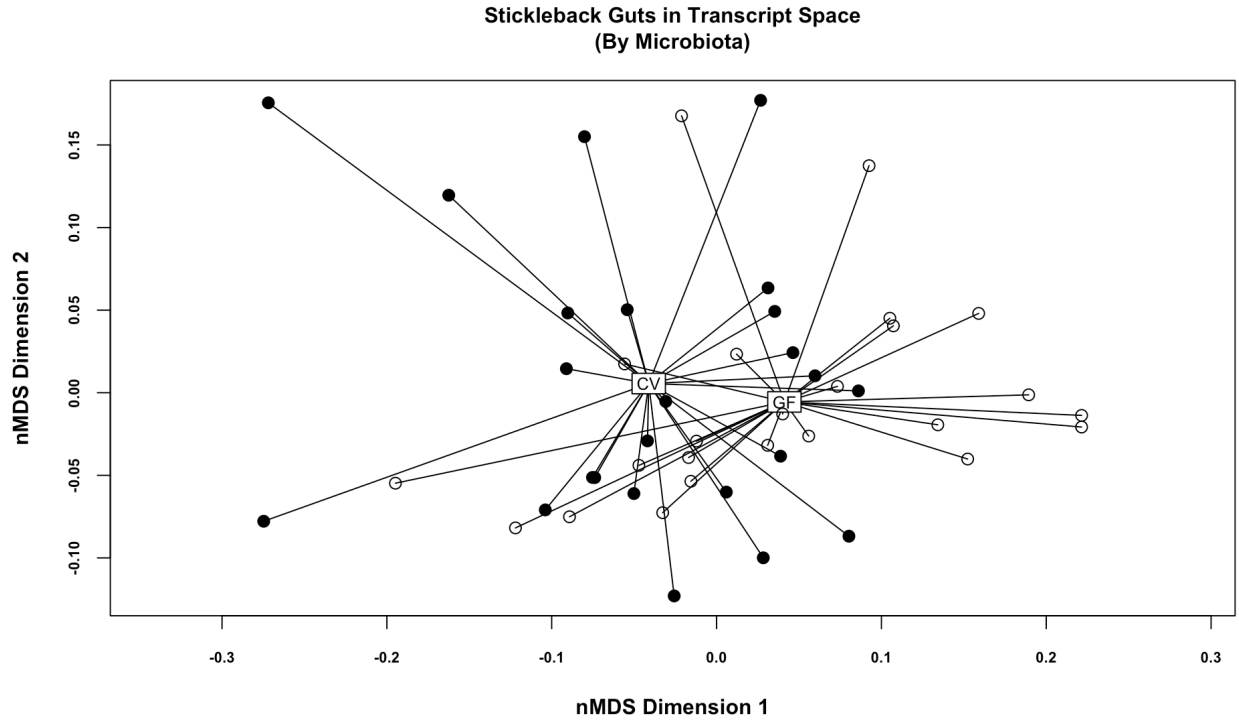
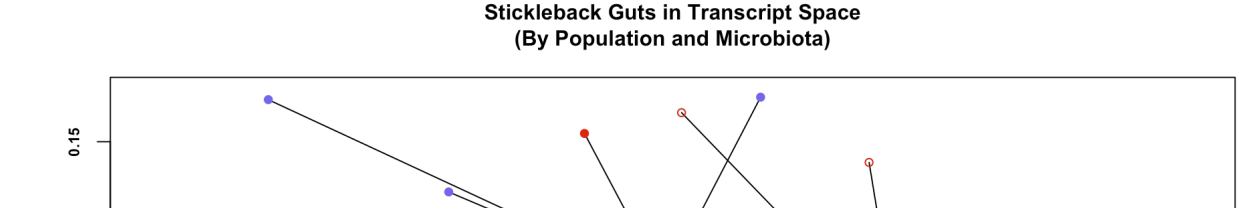


Figure 9



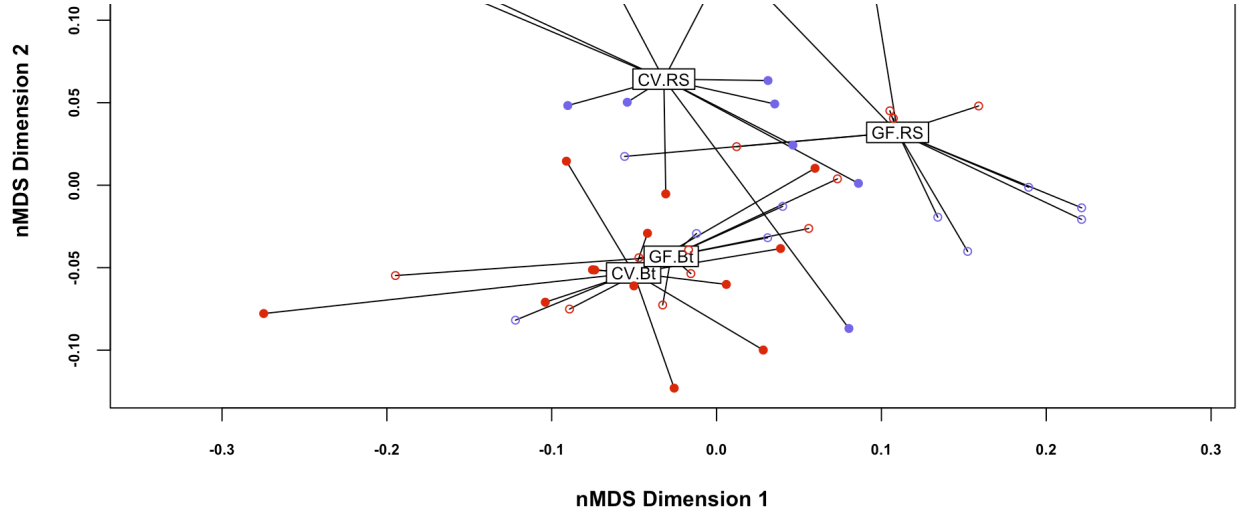


Table 1 - Permutation test

```
##
## Call:
## adonis(formula = vare.dis ~ Microbiota * Population, data = otu.env,      pe
rmutations = 1000)
##
## Permutation: free
## Number of permutations: 1000
##
## Terms added sequentially (first to last)
##
##                       Df SumsOfSqs  MeanSqs F.Model      R2   Pr(>F)
## Microbiota             1   0.06007 0.060070  5.3902 0.08809 0.000999 ***
## Population             1   0.10351 0.103514  9.2884 0.15181 0.000999 ***
## Microbiota:Population  1   0.02795 0.027948  2.5078 0.04099 0.024975 *
## Residuals             44   0.49035 0.011144         0.71911
## Total                 47   0.68189                  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion

When looking at unique features in the Gnotobiotic organisms, one must understand how they are characterized by experimentally controlled assemblages of microbes. Isolating the microbiota in the intestines of stickleback fish while factoring in sex ratios, lipid content, and transcript abundance, one can infer differences across the distribution of gene expression in sticklaback fish. According to the literature, there is this mutual understanding that if fish is germ-free, it has a harder time ingesting food [2,3]. However, the data generated from Dr. Cresko's lab contradicts that. Since this data from lipid concentration and RNA_seq data were looked at concurrently, lipids or stickleback in the lab are shown to be fatter and bigger in this data than other studies. Utlizing the software R [1], we were able

to apply statistical analysis and various computations in this paper. While looking at the RNA_seq data and running both PCoA and nMDS analysis, we found a pattern that fit both tests with regards to clustering between each factor and it's corresponding response variable (i.e., gene expression). The general trend showed across both the preliminary analysis in the intitial datasets and then in RNA seq dataset that both microbiota and population had a role in gene expression. When referencing **Figure 6** one can see four clusters given the four linear combinations of *Population* and *Microbiota*. Germ free *Microbiota* stickleback in the bootlake *Population* cluster most closely. The Rabbit Slough *Population* given the *Microbiota* CV treatment seemed to have the most variation (most distance between points) explained by PCoA 1 and PCoA 2. This is later confirmed in the nMDS analysis and shows similar trends with **Figure 9**. Given that the microbioata had the most signifiant effect on lipid concentration, we would also expect to see that in the RNA sequence data. As shown above, microbioata alone has been established as an infleuncer in lipid production. The difference across the various Microbiobiota tyes showes that a controlled rather than germ free environment could play a role in the natural production of lipid expression. As mentioned in [3] environmental factors such as intestinal microorganisms and diet can help represent attractive targets for localizing and more importantly control dietary lipid absorption and energy balance. When you factor in various populations, however, you see less of a trend with regards to an influx of lipid concentration (especially when referencing the Bt *Population*). Comparing the overall effect of *Population* & *Microbiota* with regards to disimilarity, one would have to reference the overall percentage that each factor had on gene expression. Cross referencing that data with the lipid concentration, we found that *Population* had a greater influence on gene expression when factoring in the interaction between the two than did *Microbioata*. This raises questions about the zebrafish at a younger stage of life and how controlled assemblages of microbes may influence host-microbe interactions. Seeing that this data highlights the influence of germ-free environments and it's role on the ingesting food, one might ask what other paramters are involved that allows the species to ingest despite having a diverse microbiome?. That is what other parameters might be playing a role that was not necesarily included in the model but may be infleuntial in the outcome. Referencing immune response cells and production of these cell types has previously been shown to be influenced by microbial cells. Many microbioal pathogenesis studies in this species have concentrated on the embryonic stages to better understand the infleunce of immune response cell production durin hematopoietic development [4]. Future studies could factor in a cluster of these cell types to see what are upregulated/downregulated and cross-reference them the data generated in this model to confirm whether or not that *Microbioata* and *Population* do in fact have a role in lipid gene expression. It's been establish that groups of immune cells undergo challenges during early stages of life for the stickleback fish which in turn could alter the proliferation, differentiation, and/or maintenance of hematopoietic immune cells[4]. Extending this model into further factors that were not included could help infer the difference between the gene expression levels and the stickleback fish microbiomes.

Literature Cited

[1] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.Rproject.org/ (http://www.Rproject.org/).

[2] Rawls et. al., (2006). Reciprocal Gut Microbiota Transplants from Zebrafish and Mice to Germ-free Recipients Reveal Host Habitat Selection, Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108 USA.

[3] Semova et. al., (2012). Microbiota Regulate Intestinal Absorption and Metabolism of Fatty Acids in the Zebrafish, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[4] Katner., M. Rawls., J. (2002). Host–microbe interactions in the developing zebrafish, Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Supplemental Material

Sup. Table 1

```
##
## Call:
## glm(formula = Survived ~ Microbiota, family = binomial, data = Gacu_survival
_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9479   0.5701   0.5701   0.8446   0.8446
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.8473     0.2440   3.473 0.000515 ***
## MicrobiotaGF   0.8873     0.3969   2.235 0.025395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.61  on 159  degrees of freedom
## Residual deviance: 165.37  on 158  degrees of freedom
## AIC: 169.37
##
## Number of Fisher Scoring iterations: 4
```

Sup. Table 1a

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         159     170.61
## Microbiota  1   5.2407       158     165.37  0.02206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sup. Table 2

```
#Survival ~ Population
Pop_Model <- glm(formula = Survived ~ Population, family = binomial, data = Gac
u_survival_df)
summary(Pop_Model)
```

```
##
## Call:
## glm(formula = Survived ~ Population, family = binomial, data = Gacu_survival
_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7600   0.6912   0.6912   0.7364   0.7364
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.3099     0.2733   4.793 1.64e-06 ***
## PopulationRS  -0.1435     0.3791  -0.378    0.705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.61  on 159  degrees of freedom
## Residual deviance: 170.47  on 158  degrees of freedom
## AIC: 174.47
##
## Number of Fisher Scoring iterations: 4
```

Sup. Table 2a

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         159     170.61
## Population  1  0.14343       158     170.47   0.7049
```

Sup. Table 3

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Gacu_survival_df$Survived
##
## Terms added sequentially (first to last)
##
##
##                                                          Df Deviance
## NULL
## Gacu_survival_df$Population                               1   0.1434
## Gacu_survival_df$Microbiota                               1   5.2455
## Gacu_survival_df$Population:Gacu_survival_df$Microbiota   1   0.2455
##                                                          Resid. Df
## NULL                                                           159
## Gacu_survival_df$Population                                    158
## Gacu_survival_df$Microbiota                                    157
## Gacu_survival_df$Population:Gacu_survival_df$Microbiota        156
##                                                          Resid. Dev
## NULL                                                        170.61
## Gacu_survival_df$Population                                 170.47
## Gacu_survival_df$Microbiota                                 165.22
## Gacu_survival_df$Population:Gacu_survival_df$Microbiota     164.98
##                                                          Pr(>Chi)
## NULL
## Gacu_survival_df$Population                                0.7049
## Gacu_survival_df$Microbiota                                0.0220 *
## Gacu_survival_df$Population:Gacu_survival_df$Microbiota    0.6203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

File 2

Sup. Table 4

```
#Model Sex ~ Population
sex_model_pop <- glm(formula = Sex ~ Population, family = binomial, data = Gacu
_sex_df)
summary(sex_model_pop)
```

```
##
## Call:
## glm(formula = Sex ~ Population, family = binomial, data = Gacu_sex_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.354  -1.220   1.011   1.135   1.135
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.1001     0.3166   0.316    0.752
## PopulationRS   0.3054     0.4521   0.675    0.499
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 109.65  on 79  degrees of freedom
## Residual deviance: 109.19  on 78  degrees of freedom
## AIC: 113.19
##
## Number of Fisher Scoring iterations: 4
```

Sup. Table 4a

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Sex
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          79     109.65
## Population  1  0.45761        78     109.19   0.4987
```

Sup. Table 4b

```
#Model Sex ~ Microbiota
sex_model_micro <- glm(formula = Sex ~ Microbiota, family = binomial, data = Ga
cu_sex_df)
summary(sex_model_micro)
```

```
##
## Call:
## glm(formula = Sex ~ Microbiota, family = binomial, data = Gacu_sex_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.354  -1.220   1.011   1.135   1.135
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.1001     0.3166   0.316    0.752
## MicrobiotaGF   0.3054     0.4521   0.675    0.499
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 109.65  on 79  degrees of freedom
## Residual deviance: 109.19  on 78  degrees of freedom
## AIC: 113.19
##
## Number of Fisher Scoring iterations: 4
```

Sup. Table 4c

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Sex
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          79     109.65
## Microbiota  1  0.45761        78     109.19   0.4987
```

Sup. Table 4d

```
#Specifiy a binomial error distribution
sex_main_model <- glm(Gacu_sex_df$Sex ~ Gacu_sex_df$Population + Gacu_sex_df$Mi
crobiota, family = binomial, data = Gacu_sex_df)
summary(sex_main_model)
```

```
##
## Call:
## glm(formula = Gacu_sex_df$Sex ~ Gacu_sex_df$Population + Gacu_sex_df$Microbi
ota,
##     family = binomial, data = Gacu_sex_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4232  -1.2871   0.9501   1.0715   1.2000
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.0529     0.3892  -0.136    0.892
## Gacu_sex_df$PopulationRS    0.3071     0.4535   0.677    0.498
## Gacu_sex_df$MicrobiotaGF    0.3071     0.4535   0.677    0.498
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 109.65  on 79  degrees of freedom
## Residual deviance: 108.73  on 77  degrees of freedom
## AIC: 114.73
##
## Number of Fisher Scoring iterations: 4
```

```
#anova(mainModel, test = "Chisq")

#plot(mainModel)
```

Sup. Table 5

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.6023 0.6155
##       76
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Gacu_lipids_df$TAGs
## W = 0.89626, p-value = 8.498e-06
```