Testing Equality of Mean Diversity Scores Between Male and Female Subjects

11/15/2019

Contents

Summary

1. No statistical difference was found between male and female subjects in the study based on the aggregated level values of Chao, Shannon and Observed Species.

2. Individual-level data (values at each iteration for each skin sample) was also analyzed and found statistical difference between make and female subjects.

3. The skin samples appeared to have several distinct clusters that require biological justification or assess the design of the experiment to see whether there is some information is missing (for example, demographic information associated with the subjects from whom the skin samples were taken).

4. Query: The formation of the data generation process is critical to determine the validity of the models built at the individual iteration level data.

1. Introduction

This analysis note contains exploratory analyses on the ecological indicators of diversity between male and female subjects in the study.

Goal: Are Chao 1 values statistically different for male and female samples? Same for Shannon Diversity, and observed species values.

Several Steps for Analysis

1. We first test the use of the mean of the 10 generated diversity scores by Qiime for each skin sample as the response to fit a linear model to see whether there is a statistically significant difference.
2. Then look at potentially hidden clustering patterns in the skin samples to define a cluster variable to fit a better model if these clusters have biological meaning.
3. Fit a random effect mixed model that allows using more granular data.

We perform the same type of analysis for each of the three ecological diversity indices.

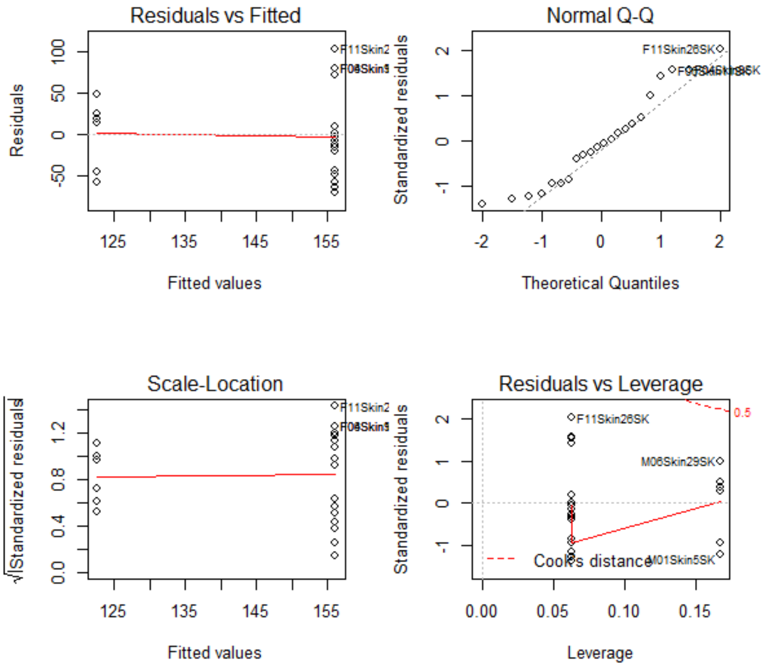
2. Chao1 Diversity Index

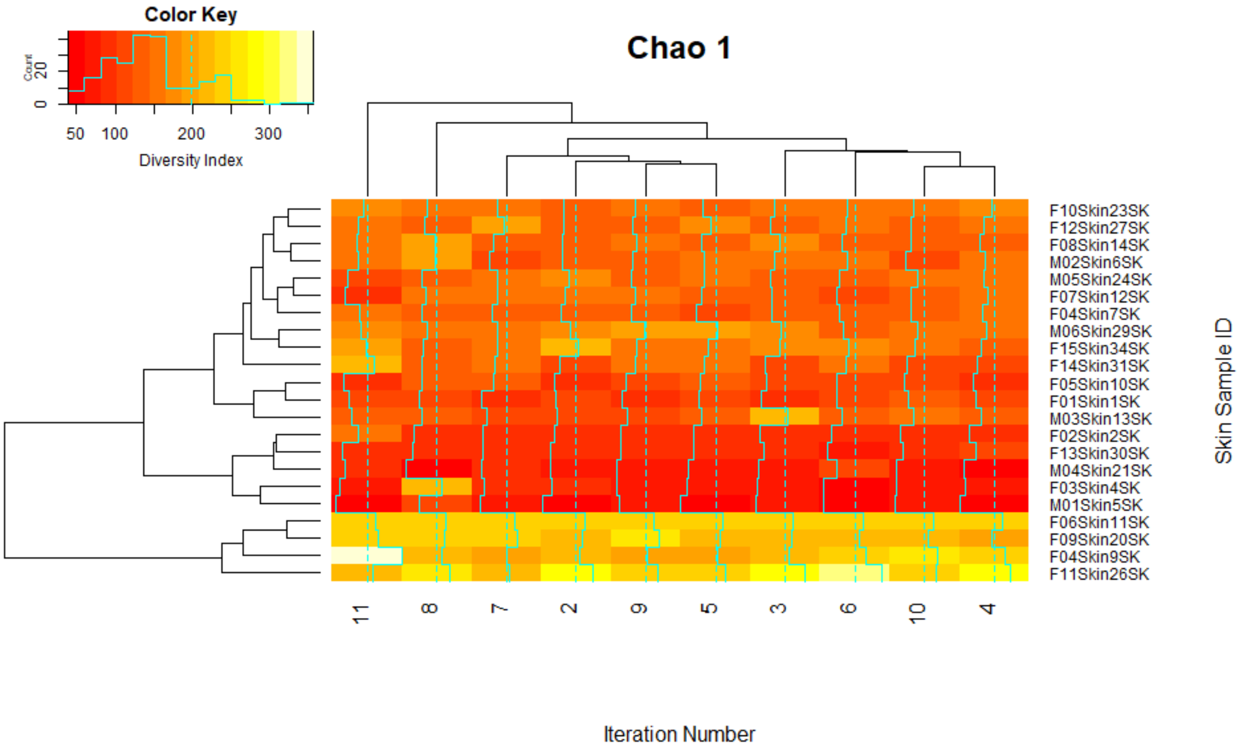
2.1. Aggregated Regression

The standard simple linear regression model was fit to the mean of 10 indices generated from Qiime for each skin sample with gender as a predictor variable.

```
## lm(formula = SK.avg ~ gender)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -70.07 -44.53  -5.30  23.10 103.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   156.13      13.14  11.878 1.63e-10 ***
## genderM       -33.53      25.17  -1.332    0.198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Interpretation: The linear model output shows that the average Chao index score of male subjects is about 33 less than that of female subjects. But it does not achieve the statistical significance (p-value = 0.198).





Interpretation: The above heatmap of the Chao index indicates three significantly different clusters of skin samples:

Cluster 1 (high Chao score): F11Skin26SK, F04Skin9SK, F09Skin20SK, and F06Skin11SK;

Cluster 2 (medium Chao score): M01Skin5SK, F03Skin4SK, M04Skin21SK, F13Skin30SK, and F02Skin2SK;

Cluster 3 (low Chao score): F01Skin1SK, M02Skin6SK, F04Skin7SK, F05Skin10SK, F07Skin12SK, M03Skin13SK, F08Skin14SK, F10Skin23SK, M05Skin24SK, F12Skin27SK, M06Skin29SK, F14Skin31SK, and F15Skin34SK.

2.2. Questions About Data Generation Process

Question: Are these clustered skin samples bearing special biological information?

Another Question: Qiime generated 10 diversity index scores for each skin sample. How these scores are generated? Observed at a different time? different experimental environments? I need to know how these 10 indices were obtained from each skin sample in order to decide whether a random effect model is appropriate for the data.

2.3. Fit A Linear Regression Using Individual-Level Data

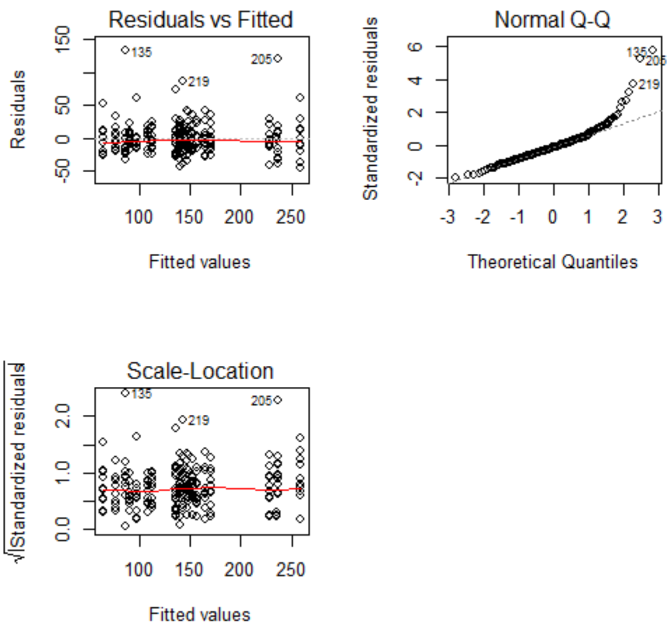
The validity of this model is dependent on the data generation process. I will justify this about I got feedback from your team.

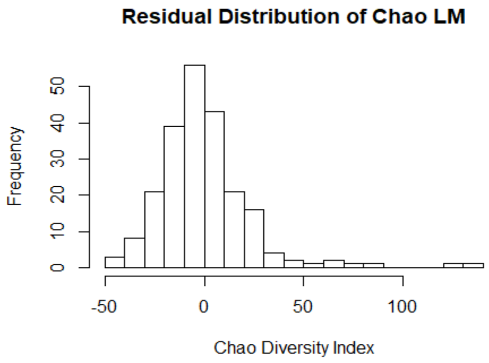
```
##
## Call:
## lm(formula = value ~ gender + skinSK, data = chao.long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.087 -13.603  -1.939   8.940 132.947
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      112.855      7.729  14.601  < 2e-16 ***
## genderM           34.485     10.931   3.155 0.001856 **
## skinSKSkin11SK   122.733     10.931  11.228  < 2e-16 ***
## skinSKSkin12SK    27.007     10.931   2.471 0.014330 *
## skinSKSkin13SK   -11.635     10.931  -1.064 0.288439
## skinSKSkin14SK    35.940     10.931   3.288 0.001194 **
## skinSKSkin1SK     -4.735     10.931  -0.433 0.665360
## skinSKSkin20SK   115.696     10.931  10.584  < 2e-16 ***
## skinSKSkin21SK   -69.699     10.931  -6.376 1.25e-09 ***
## skinSKSkin23SK   44.382     10.931   4.060 7.06e-05 ***
## skinSKSkin24SK   -6.604     10.931  -0.604 0.546431
## skinSKSkin26SK  146.442     10.931  13.397  < 2e-16 ***
## skinSKSkin27SK   40.003     10.931   3.660 0.000324 ***
## skinSKSkin29SK   22.890     10.931   2.094 0.037528 *
## skinSKSkin2SK   -15.529     10.931  -1.421 0.156992
## skinSKSkin30SK  -21.773     10.931  -1.992 0.047760 *
## skinSKSkin31SK   29.818     10.931   2.728 0.006948 **
## skinSKSkin34SK   52.842     10.931   4.834 2.68e-06 ***
## skinSKSkin4SK   -26.802     10.931  -2.452 0.015075 *
## skinSKSkin5SK   -83.435     10.931  -7.633 9.50e-13 ***
## skinSKSkin6SK        NA         NA      NA       NA
## skinSKSkin7SK    22.805     10.931   2.086 0.038234 *
## skinSKSkin9SK   123.512     10.931  11.299  < 2e-16 ***
## ---
```

The output indicates that

(1). There is a statistically significant difference between male and female subjects in terms of the Chao Index. The directions of the difference in individual and aggregate analysis are opposite!

(2). The Chao index across the skin samples seems to have several distinct groups. This is consistent with what we observed in the heatmaps in which distinct clusters were observed.





Residual plots show that the linear regression model fits the data well.

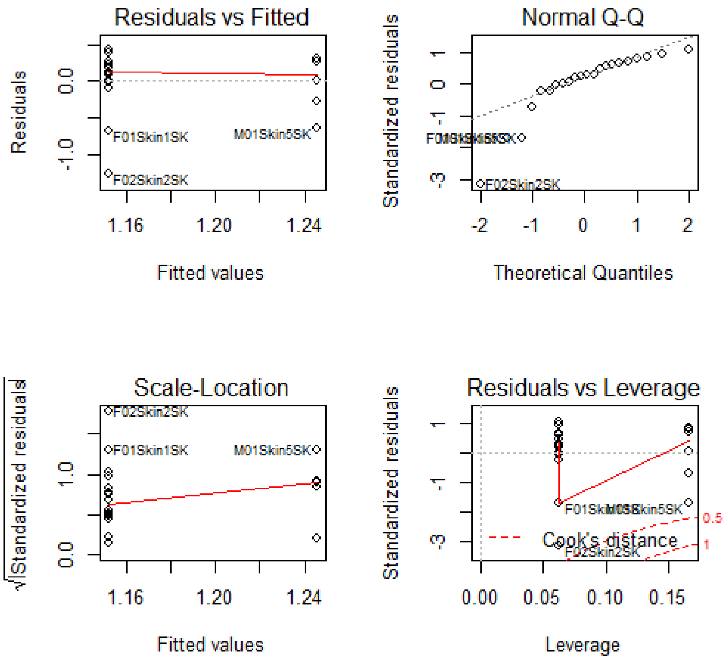
Remark: For Shannon and Observed Species indices, I can repeat the same type of analyses. I will not go to details before obtaining your expert feedback. I will provide basic regression models and heatmaps so you see the same pattern that there is no statistically significant difference between male and female subjects.

3. Shannon Diversity Index

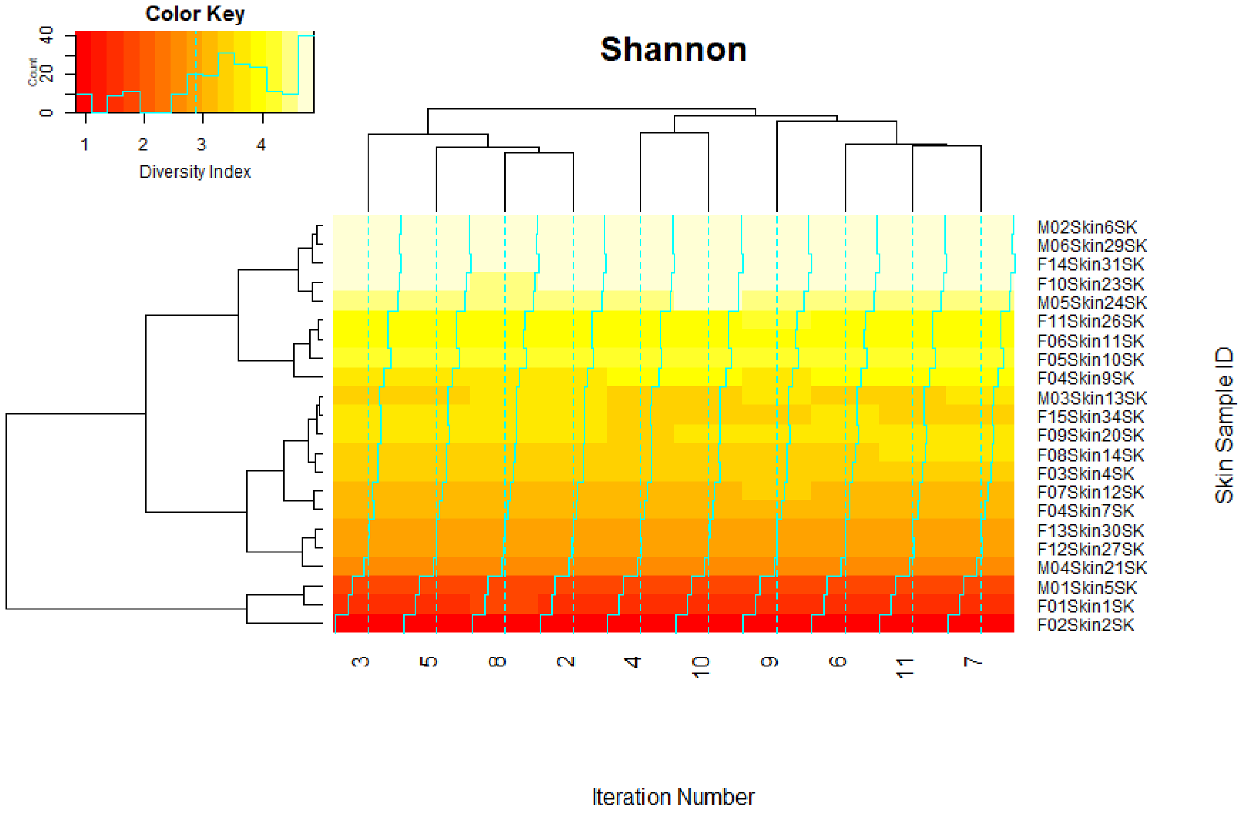
3.1. Aggregated Regression

Similar to what we did in the Chao index, a linear regression based on the aggregate data was fit.

```
## lm(formula = SK.avg ~ gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4897 -0.4477  0.1273  0.7452  1.4723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3787     0.2688  12.570 5.96e-11 ***
## genderM       0.3016     0.5147   0.586    0.564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



Residuals vs Fitted
Residuals
0.0
-1.0
F01Skin1SK
M01Skin5SK
F02Skin2SK
1.16
1.20
1.24
Fitted values
Normal Q-Q
Standardized residuals
1
0
-1
-3
F01Skin5SK
F02Skin2SK
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.0
0.0
F02Skin2SK
F01Skin1SK
M01Skin5SK
1.16
1.20
1.24
Fitted values
Residuals vs Leverage
Standardized residuals
1
-1
-3
F01Skin1SK
M01Skin5SK
Cook's distance
F02Skin2SK
0.5
1
0.00
0.05
0.10
0.15
Leverage



Color Key
Count
40
20
0
1
2
3
4
Diversity Index
Shannon
M02Skin6SK
M06Skin29SK
F14Skin31SK
F10Skin23SK
M05Skin24SK
F11Skin26SK
F06Skin11SK
F05Skin10SK
F04Skin9SK
M03Skin13SK
F15Skin34SK
F09Skin20SK
F08Skin14SK
F03Skin4SK
F07Skin12SK
F04Skin7SK
F13Skin30SK
F12Skin27SK
M04Skin21SK
M01Skin5SK
F01Skin1SK
F02Skin2SK
Skin Sample ID
3
5
8
2
4
10
9
6
11
7
Iteration Number

Interpretation of the model outputs:

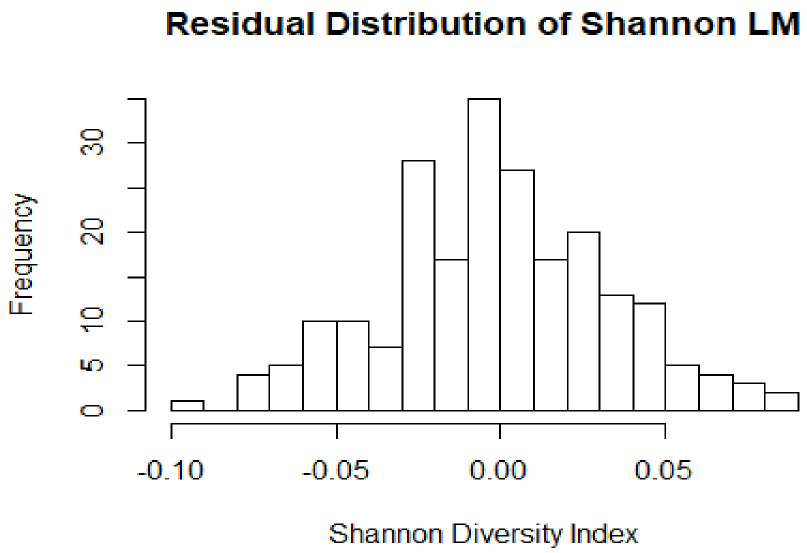
1. No statistical significance was achieved in terms of the difference between male and female subjects in the study. However, we can see the Shannon index of the male group is slightly *higher than* the female group by 0.3016 (p=0.564).

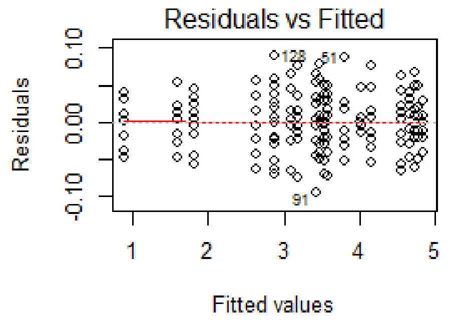
2. From the heatmap, we can see the existence of three clusters of the skin samples based on the Shannon index.

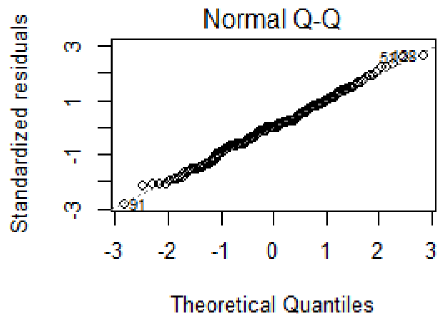
3.2. Regression on Individual-Level Data

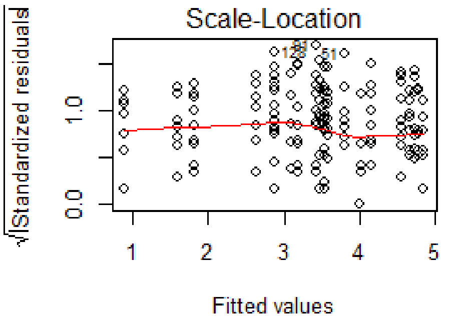
The same comment applies to this model in terms of model validity.

```
##
## Call:
## lm(formula = value ~ gender + skinSK, data = shannon.long)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.096 -0.022  0.000  0.023  0.089
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     4.15400    0.01130  367.746  < 2e-16 ***
## genderM         0.62700    0.01597   39.249  < 2e-16 ***
## skinSKSkin11SK -0.15400    0.01597   -9.640  < 2e-16 ***
## skinSKSkin12SK -0.96000    0.01597  -60.095  < 2e-16 ***
## skinSKSkin13SK -1.24900    0.01597  -78.186  < 2e-16 ***
## skinSKSkin14SK -0.67300    0.01597  -42.129  < 2e-16 ***
## skinSKSkin1SK  -2.55700    0.01597 -160.065  < 2e-16 ***
## skinSKSkin20SK -0.57200    0.01597  -35.807  < 2e-16 ***
## skinSKSkin21SK -2.13700    0.01597 -133.774  < 2e-16 ***
## skinSKSkin23SK  0.50500    0.01597   31.612  < 2e-16 ***
## skinSKSkin24SK -0.22400    0.01597  -14.022  < 2e-16 ***
## skinSKSkin26SK -0.12000    0.01597   -7.512 1.96e-12 ***
## skinSKSkin27SK -1.28300    0.01597  -80.314  < 2e-16 ***
## skinSKSkin29SK -0.03900    0.01597   -2.441   0.0155 *
## skinSKSkin2SK  -3.26500    0.01597 -204.385  < 2e-16 ***
## skinSKSkin30SK -1.28100    0.01597  -80.189  < 2e-16 ***
## skinSKSkin31SK  0.69700    0.01597   43.631  < 2e-16 ***
## skinSKSkin34SK -0.62300    0.01597  -38.999  < 2e-16 ***
## skinSKSkin4SK  -0.71800    0.01597  -44.946  < 2e-16 ***
## skinSKSkin5SK  -2.95500    0.01597 -184.980  < 2e-16 ***
## skinSKSkin6SK        NA         NA       NA       NA
## skinSKSkin7SK  -1.04900    0.01597  -65.666  < 2e-16 ***
## skinSKSkin9SK  -0.35100    0.01597  -21.972  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```









Interpretations:

(1). The model fits the data well. The Shannon index in male group is significantly higher than the female group (p=2E-16).

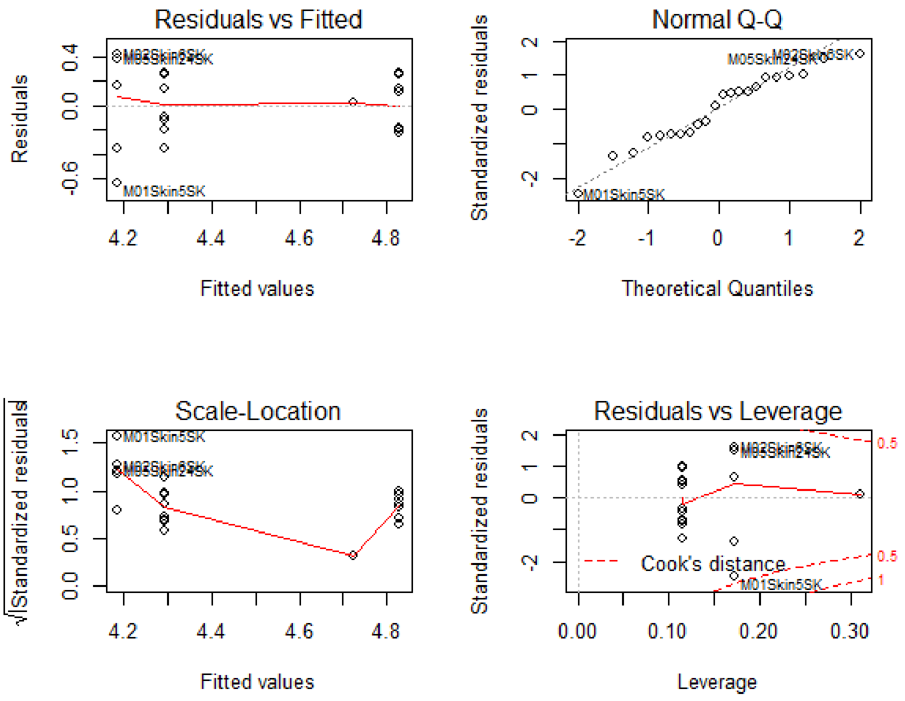
(2). Skin samples are heterogeneous. This is also consistent with the patterns in the heatmap.

4. Observed Species

4.1. Regression on Individual-Level Data

Linear regression model based on the aggregated index scores.

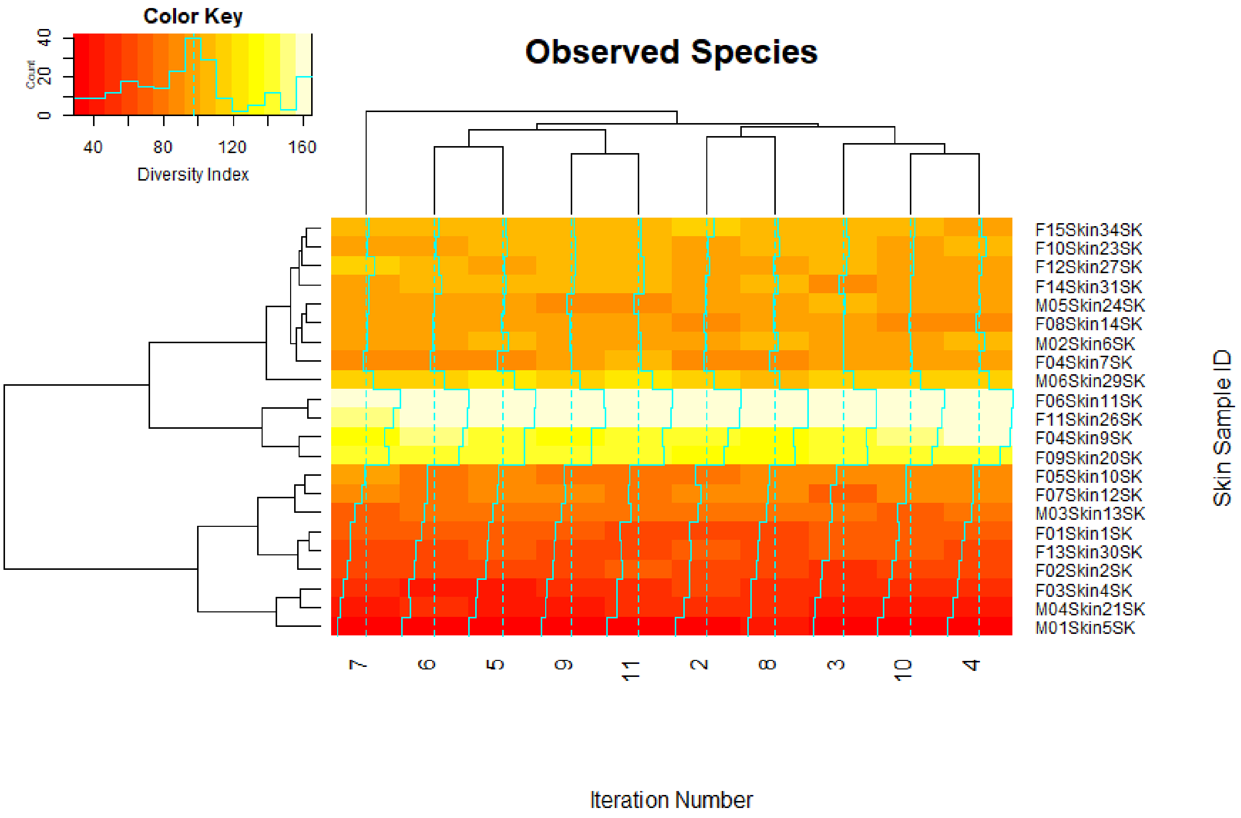
```
##
##
## Call:
## lm(formula = SK.avg ~ gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.444 -28.548  -0.797  20.525  62.756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  101.144      8.526  11.863 1.66e-10 ***
## genderM      -22.894     16.326  -1.402    0.176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



Interpretation:

1. Residual plots indicate a good fit of the model. The observed species value of male subjects is about 23 less than the female subjects (p= 0.176) but did not achieve the statistical significance level.

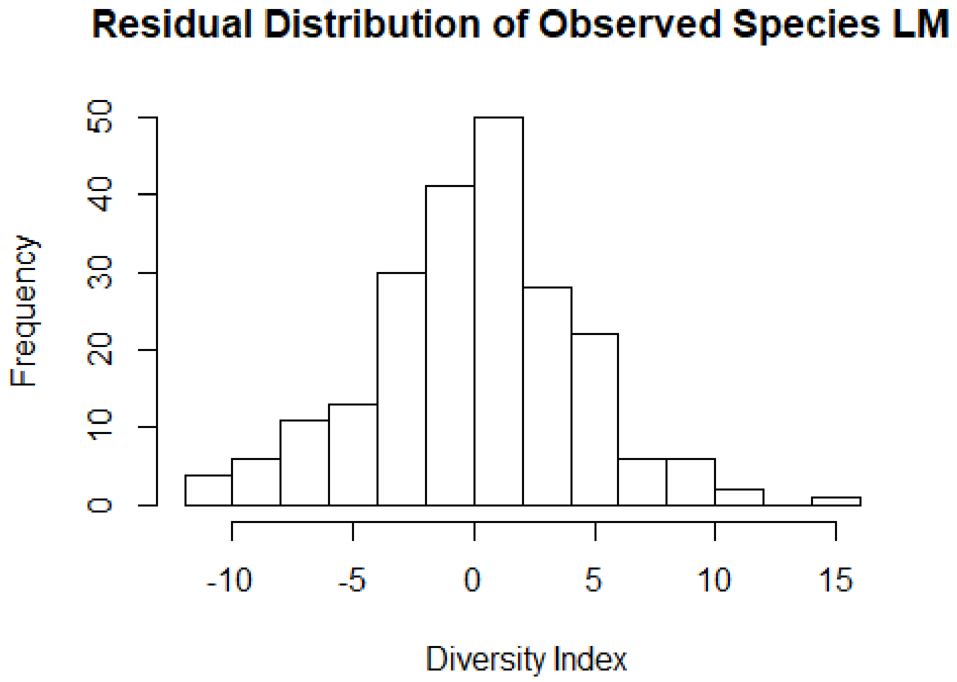
2. The following heatmap gives the same clustering information as shown in the previous analysis.

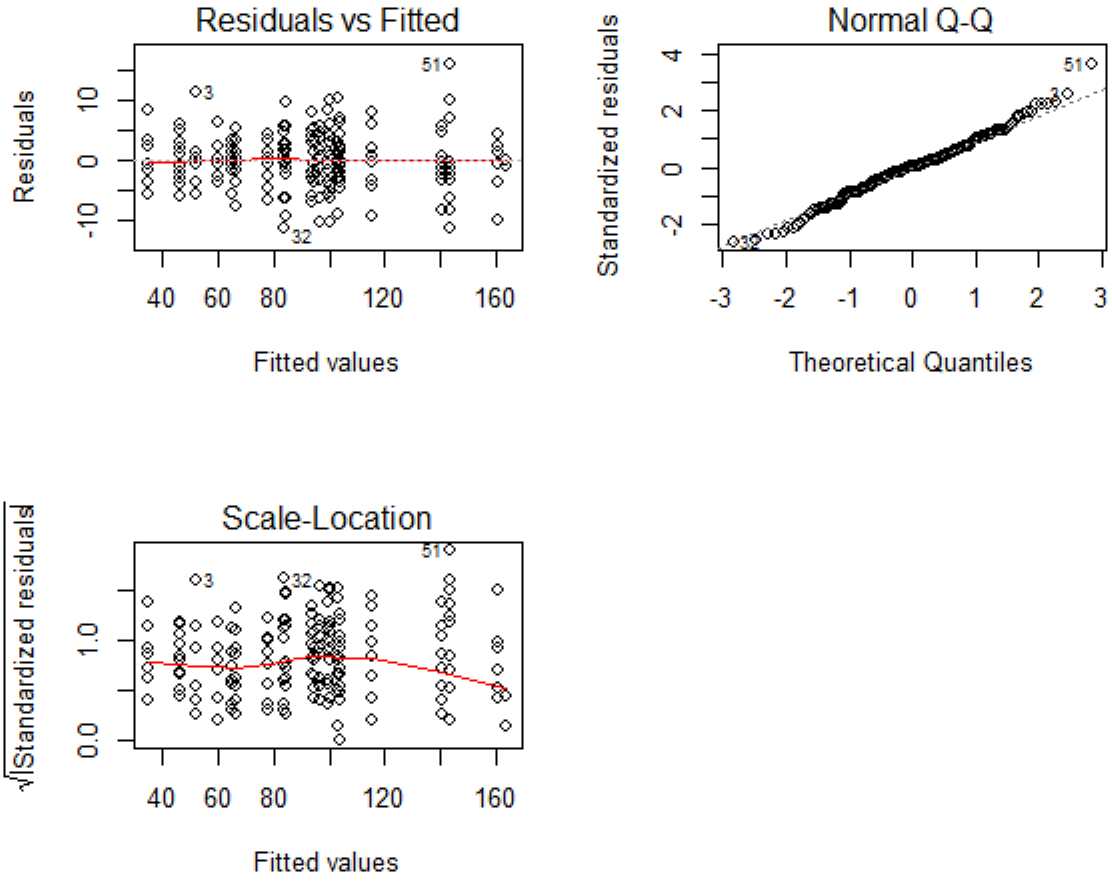


4.2. Regression on Individual-Level Data

The validity issue to be judged!

```
##
## Call:
## lm(formula = value ~ gender + skinSK, data = obs.spec.long)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -11.4   -2.8    0.1    2.6   15.8
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      84.300      1.463  57.617  < 2e-16 ***
## genderM          15.300      2.069   7.394 3.93e-12 ***
## skinSKSkin11SK   79.600      2.069  38.470  < 2e-16 ***
## skinSKSkin12SK   -0.900      2.069  -0.435    0.664
## skinSKSkin13SK  -22.000      2.069 -10.632  < 2e-16 ***
## skinSKSkin14SK   10.500      2.069   5.075 8.92e-07 ***
## skinSKSkin1SK   -17.600      2.069  -8.506 4.39e-15 ***
## skinSKSkin20SK   56.000      2.069  27.064  < 2e-16 ***
## skinSKSkin21SK  -53.500      2.069 -25.856  < 2e-16 ***
## skinSKSkin23SK   19.400      2.069   9.376  < 2e-16 ***
## skinSKSkin24SK   -3.300      2.069  -1.595    0.112    
## skinSKSkin26SK   76.500      2.069  36.972  < 2e-16 ***
## skinSKSkin27SK   18.600      2.069   8.989  < 2e-16 ***
## skinSKSkin29SK   15.600      2.069   7.539 1.66e-12 ***
## skinSKSkin2SK   -24.500      2.069 -11.841  < 2e-16 ***
## skinSKSkin30SK  -19.700      2.069  -9.521  < 2e-16 ***
## skinSKSkin31SK   15.900      2.069   7.684 6.98e-13 ***
## skinSKSkin34SK   19.700      2.069   9.521  < 2e-16 ***
## skinSKSkin4SK   -32.600      2.069 -15.755  < 2e-16 ***
## skinSKSkin5SK   -64.900      2.069 -31.365  < 2e-16 ***
## skinSKSkin6SK        NA         NA      NA       NA    
## skinSKSkin7SK     9.700      2.069   4.688 5.13e-06 ***
## skinSKSkin9SK    58.900      2.069  28.466  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```





Interpretations:

(1). The model fits the data well. The value of Observed Species in male group is significantly higher than the female group (p=3.92E-12).

(2). Skin samples are heterogeneous. This is also consistent with the patterns in the heatmap.