```
basedOnGenus <- as.data.frame(tax_table(firstQTaxa)) %>% filter(!str_detect(Genus,
    "uncultured|unclassified|\\bD_5__\\b"))
secondQTaxa <- subset_taxa(firstQTaxa, Genus %in% basedOnGenus$Genus)
```

2.1 Shannon Diversity Index (SDI) and Chao1

We applied *Shannon diversity index (SDI)* to estimate the microbial diversity of Saanich Inlet dataset. It has the following definition:

$SDI = -\sum_i^R p_i \log(p_i)$

where p_i represents the distribution of individuals belonging to the ith species, and R represents the number of distinct species [17]. It can be noted that SDI takes both the richness and abundance information to measure the expected uncertainty about species contained in a sample. The high SDI value suggests that species are evenly distributed while low SDI value implies species are disproportionality situated. SDI value could be zero meaning the sample contains exactly one or no species at all. However, SDI does not directly model the expected richness of a sample and, neither, it represents an accurate estimation of species diversity because the probability distribution of species is not knowable exactly; it is only an estimate from a sample.

In contrast to SDI, *Chao1* could be used to recover approximate true richness:

$Chao1 = S_{obs} + \frac{\alpha}{2\beta}$

where S_{obs} represents the observed richness, α and β indicates the number of different species with exactly one or more than two counts, respectively. The Chao1 method is used to rectify the richness by including the distribution of the rarest species [17].

2.2 General Linear Model

General linear model (LM) [18] is employed to recover interactions between several factors that might be exhibited in Saanich Inlet dataset. In our experiments, we use a single regression model that relates a dependent variable y (abundance) to a single quantitative independent variable x_1 (depth or oxygen), and it has the following form:

$y = \theta_0 + \theta_1 x_1 + \epsilon$

The parameter θ_0 is the y-intercept, which represents the expected value of y when x_1 is zero. The parameter θ_1 is the slope of the regression line, and it represents the expected change (positive or negative) in y (abundance) for a unit increase in x_1 (depth or oxygen). θ_1 could be 0 indicating no effective change with x_1. And, ϵ is the error term and is usually set to 0.

All the parameters could be estimated using ordinary least squares. However, to test the significance θ_1, the following hypothesis testing was formulated: i) The null hypothesis $H_0 : \theta_1 > \gamma$, which asserts that no additional predictive value over and above, contributed by θ_1 and the γ is an arbitrary cutoff probability from t-student distribution ; ii) The alternative hypothesis $H_1 : \theta_1 \leq \gamma$ measures whether x_j has additional predictive strengths.

If the weight of θ_1, referred to as the level of significance (or p-value) and defines as a probability, is below or equal to γ then H_1 was accepted; otherwise, accept H_0.