Microbiome analysis: models, methods and simulations

Zerui Zhang

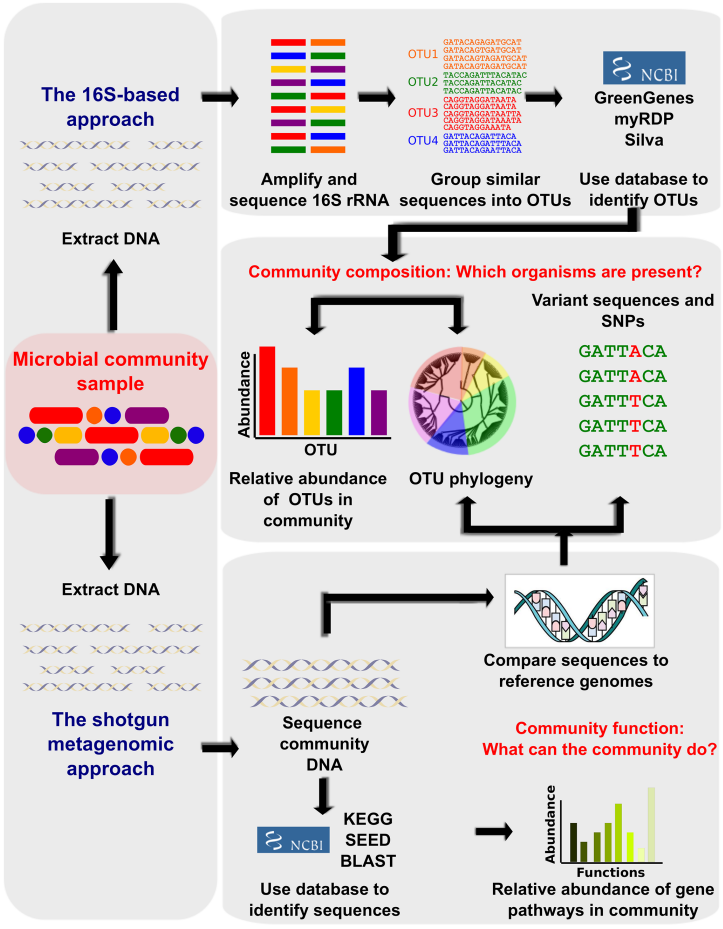
April 16, 2018

1 Introduction to microbiome analysis

1.1 Goal of analysis

1. Taxonomic diversity
2. Functional metagenomics

1.2 Preparations before dealing with data[18]



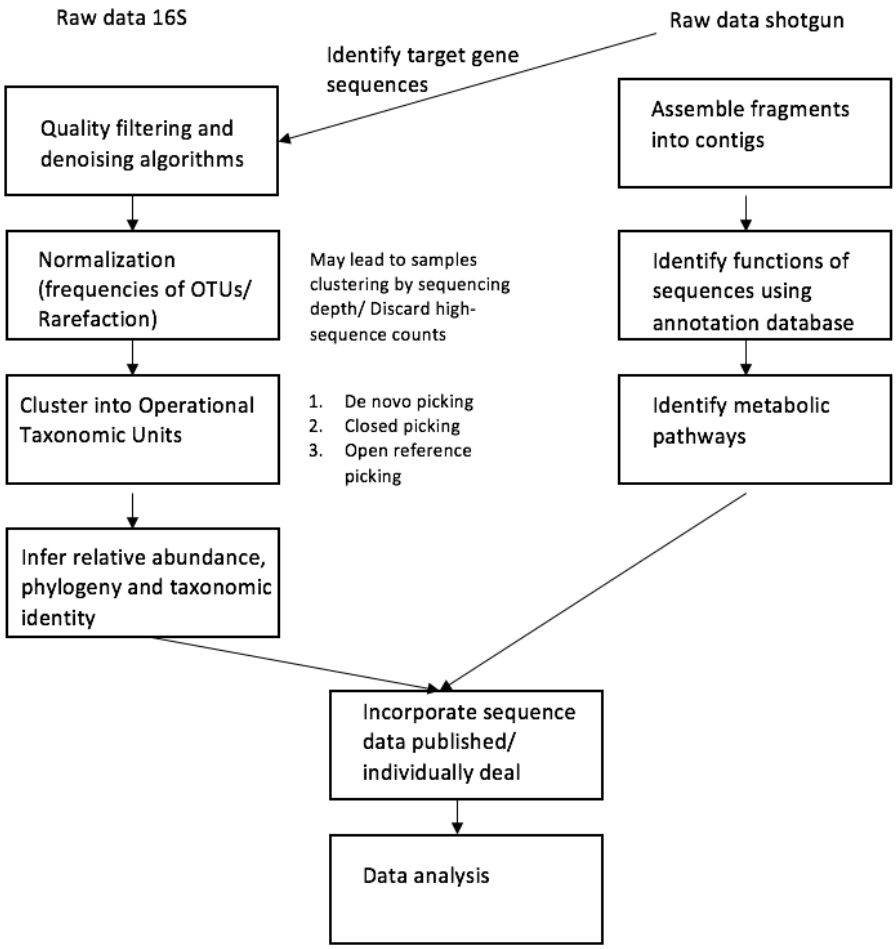
Some sources of variance(with adjustments):

1. Each fragment may be sequenced only once (Evaluated on the basis of a mock community as baseline)
2. Error during choosing marker gene, choosing region to be sequenced, marker extraction, amplification and sequencing (Greatly improved when platform transition from 454-based to Illumina; For

454, parameters like the average quality score of the sequence, number of mismatches in the primers or total length can help filter bad sequencing results; For Illumina, filtering can occur downstream in data analysis, discard the sequences observed only one time or only in a single sample)

3. The abundance of chimeric sequences, or reads in which two unique markers adhere during amplification (Usually equally distributed across all samples, so they will affect estimates of species richness, aka total number of OTUs in a given sample, not estimates of relative species diversity between samples)

Bioinformatics before data analysis



2 OTU count data

2.1 Generation of OTUs

Concept of Operational Taxonomic Unit (OTU) was introduced in the 1960s in the field of numerical taxonomy.[9] The goal was to develop a quantitative strategy for classifying organisms into groups, creating a tree by repeatedly merging the most similiar groups according to the number of traits they had in common. This is an early example of a agglomerative clustering algorithm, before the sequence-based phylogenetic tree construction algorithm (e.g. neighbor-joining, maximum likelihood).

In 16S sequencing, OTUs are typically constructed using an identity threshold of 97%.[10] It is assumed that 97% similarity of 16S sequences corresponded approximately to a DNA reassociation value of 70%, which had been accepted as a working definition for bacterial species.[11] But in NGS sequencing, this threshold would generate many spurious OTUs due to experimental error. We often use de-nove,

closed-reference, open-reference for OTU picking.

α **diversity**: a mearuse of the diversity in a single sample. The simplest measure is richness, the number of species (OTUs) observed in the sample. Other metrics consider the abundences (frequencies) of OTUs, for example to give lower weight to lower-abundance OTUs.

β **diversity**: compares two samples to indicate the similarity or difference between the samples.

2.2 Characteristics:

1. Discrete
2. Count data
3. Overdispersion
4. Sparsity
5. Different-size sample
6. High-dimentional

2.3 Model selection and analysis for OTUs analysis

1. Negative binomial model was fitted in microbiome abundance data analysis, it was also used to test for assessing differences in sequence tag abundance and used for detecting differentially abundant features in clinical metagenomic samples. (deal with overdispersion)

2. Zero-inflated Poission, Zero-inflated Negative Binomial are chosen for modeling the excess zeros. (Hurdle model?)

3. Zero-inflated Gaussian mixture model, motivated by the observed strong correlation between the number of OTUs detected in a sample and the corresponding sequencing depth in 16S DNA studies. It seeks to directly estimate the probability that an observed zero is generated from the detection distribution due to undersampleing or from the count distribution.[13]

4. Dirichlet multinomial model fit multivariate data like microbial metagenomics data. This data can be represented as a frequency matrix giving the number of times each taxa is observed in each sample. The samples have different size, and the matrix is sparse, as communities are diverse and skewed to rare taxa. Most methods used previously to classify or cluster samples have ignored these features. We describe each community by a vector of taxa probabilities. These vectors are generated from one of a finite number of Dirichlet mixture components each with different hyperparameters. Observed samples are generated through multinomial sampling.[6]

3 Some proposed models

3.1 Poisson

Why not simple linear or logarithmic scaling adjustment?
The data of different library imply different levels of uncertainty. e.g., sampling variance of the proportion estimate for each gene / OTU.

Variation in the read counts of features between technical replicates: **Possion** R.Vs. (Poisson process: $T = t_1 + \cdots + t_n, t_i \to 0, P(A) \propto t_i; For \quad t_i, P(A_1 \cap \cdots \cap A_n) = 0; For \quad t_i, t_j, P(A_i \cap A_j) = P(A_i)P(A_j)$)

For geting count data by read-mapping problem:
First consider with binomial distribution, a read lands in a given gene (success) or not (failure):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Imagine situations with unknown trials n, only know the average time of success per interval

$$p = n\lambda$$

$$\begin{aligned}
&\lim_{n\to\infty} P(X = k) \\
&= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} (\frac{\lambda^k}{n})(1 - \frac{\lambda}{n})^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k} (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{n^k} \times e^{-\lambda} \times 1 \\
&= \frac{\lambda^k}{k!} e^{-k}
\end{aligned}$$

where

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^n = \lim_{n\to\infty} (1 - \frac{1}{x})^{x(-\lambda)} = e^{-\lambda}$$

Fit microbiome count data[1]:
With overdispersed, **negative binomial** is applied.
With excess 0's, **zero-inflated models**

3.2 ZIP model (mixture model)[2]

Why not **poisson**? If $X \sim Poisson(\lambda), E(X) = Var(X)$, while for some biological repeats, $Var(X)$ may $> E(X)$, which is *overdispersed*

Why not **binomial**? For binomial, N should be given thus it has a upper limit; $E(X) = np, Var(X) = np(1-p), Var(X) < E(X)$, more restrictions on dispersion than that of Poisson.

$\mathbf{Y} = (Y_1, \dots, Y_n)'$ are independent and

$$Y_i \sim \begin{cases} Poisson(\lambda_i) & with \quad probability \quad 1 - p_i \\ 0 & with \quad probability \quad p_i \end{cases}$$

$$Y_i = \begin{cases} 0 & p_i + (1-p_i)e^{-\lambda_i} \\ k & (1-p_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!}, \quad k = 1, 2, \dots \end{cases}$$

$$E(Y) = \lambda(1-p), \quad Var(Y) = \lambda(1-p)(1+\lambda p)$$

Note that this distribution approaches to $Poisson(\lambda)$ as $p \to 0$
Parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\mathbf{p} = (p_1, \ldots, p_n)'$ satisfy

$$log(\lambda) = \mathbf{B}\beta \quad logit(p) = log(\frac{p}{1-p}) = \mathbf{G}\gamma$$

where $\mathbf{B}$ and $\mathbf{G}$ are observed covariate matrices. β is cofficient matrix. $\mathbf{B}$ is responsible for Poisson outcome in $\mathbf{Y}$, and $\mathbf{G}$ is responsible for excess zeros in $\mathbf{Y}$.

$$\lambda = e^{\boldsymbol{B\beta}}, \quad p = \frac{e^{\boldsymbol{G\gamma}}}{1+e^{\boldsymbol{G\gamma}}}$$

$$f(y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} = \frac{e^{\boldsymbol{G_i\gamma}}}{1+e^{\boldsymbol{G_i\gamma}}} + \frac{e^{-e^{\boldsymbol{B_i\beta}}}}{1+e^{\boldsymbol{G_i\gamma}}} & y_i = 0 \\ (1-p_i)\frac{e^{-\lambda_i}\lambda_i^y}{y_i!} = \frac{e^{-e^{\boldsymbol{B_i\beta}}}(e^{B_i\beta})^y}{(1+e^{G_i\gamma})y_i!} & y_i = 1, 2, \ldots \end{cases}$$

$$l(\gamma, \beta; y) = \sum_{y_i=0} log(e^{G_i\gamma} + e^{-e^{B_i\beta}}) + \sum_{y_i>0} [y_i B_i\beta - e^{B_i\beta} - log(y_i!)] - \sum_{i=1}^{n} log(1+e^{G_i\gamma})$$

When $\mathbf{B} = \mathbf{G}$ and λ and p are not functionally related, **ZIP** requires twice as many parameters as Poisson regression.($log(E(Y|X) = \beta_0 + \beta_1 X)$)
When $\mathbf{G}$ is a column of ones, **ZIP** requires one more parameter than Poisson regression.

3.3 NB model(gamma-poisson model)[3]

$$Y \sim NB(\mu, \theta) \quad f(y; \mu, \theta) = P(Y = y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}(\frac{\theta}{\mu+\theta})^\theta(\frac{\mu}{\mu+\theta})^y$$

Where μ is mean, θ is shape parameter which quantifies the amount of overdispersion.

$$Derived \quad from \quad f(x; r, p) = \binom{x+r-1}{x} p^r(1-p)^x, \quad where \quad r = success, \quad x = failures.$$

The NB can also be expressed as a gamma mixture of Poisson:

$$Y \sim Poisson(\mu\epsilon), where \quad \epsilon \sim Gamma(\theta, \theta)$$

$$P(Y = y; \mu\epsilon)g(\epsilon) = \frac{e^{-\mu\epsilon}(\mu\epsilon)^y}{y!}\frac{\theta^\theta}{\Gamma(\theta)}(\mu\epsilon)^{\theta-1}e^{-\theta\mu\epsilon}$$

$$P(Y = y) = \int_0^\infty P(Y = y; \mu\epsilon)g(\epsilon)d\epsilon = \binom{y+\theta-1}{y}(\frac{\theta}{\mu+\theta})^\theta(\frac{\mu}{\mu+\theta})^y$$

$$E(Y) = \mu, \quad Var(Y) = \mu + \frac{\mu^2}{\theta}$$

$Var(Y) \geq E(Y)$. Shape parameter θ controls the amount of over-dispersion.
When $\theta = +\infty, Var(Y) = \mu$, NB converges to Poisson that cannot deal with over-dispersion.

Negative binomial mixed models(NBMMs):

$$log(\mu_i) = log(T_i) + X_i\beta + Z_i b$$

$$Often \quad assume \quad b \sim N_k(0, \tau^2 I)$$

Where T_i is total sequence read(depths of coverage, or library size). X_i, host factors, represent host clinical/environment or genetic variables. Z_i, sample variables, introduce hierarchical, spatial, and temporal dependence of microbiome counts.
The goal is to detect associations between microbiome features C_{ij} and host factors X_i.

3.4 ZINB

$$P(Y=y)=\begin{cases} p+(1-p)(1+\frac{\lambda}{\mu})^{-\mu} & y=0 \\ (1-p)\frac{\Gamma(y+\mu)}{y!\Gamma(\mu)}(1+\frac{\lambda}{\mu})^{-\mu}(1+\frac{\mu}{\lambda})^{-y} & y_i=1,2,\dots \end{cases}$$

$$E(Y)=(1-p)\lambda, Var(Y)=(1-p)\lambda(1-p\lambda+\frac{\lambda}{\mu})$$

Note that this distribution
approaches ZIP as $\mu \to 0$,
approaches NB as $p \to 0$.
If both $\frac{1}{\mu} \to 0$ and $p \to 0$, then ZINB reduces to Poisson.
The ZINB regression model relates p and λ to covariate matrix $\boldsymbol{X}$ and $\boldsymbol{Z}$ with regression parameters β and γ:

$$\lambda = e^{\boldsymbol{X\beta}}, \quad p=\frac{e^{\boldsymbol{Z\gamma}}}{1+e^{\boldsymbol{Z\gamma}}}$$

3.5 Hurdle regression model

It may be define as a two-part model where the first part is a binary outcome model, and the second part is a truncated count model. A dataset is split into zero and non-zero values to fit two different model with associated covariates in regression.

$$P(y=0)=f_1(0)$$
$$P(y=j)=\frac{1-f_1(0)}{1-f_2(0)}f_2(y), \quad j>0$$

3.6 Dirichlet-multinomial model[6]

Motivation: View the sequence data generated by sampling from the community. If assume sampling with replacement, the likelihood of an observed sample is a multinomial distribution with a parameter vector where a given entry represents the probability that a read is from a given taxa. The probabilities in the limit of very large community sizes will become the relative frequencies of the taxa.

$$p_1,\dots,p_k \sim Dirichlet(\alpha_1,\dots,\alpha_k)$$
$$y_1,\dots,y_k \sim Multinomial(p_1,\dots,p_k)$$

Multinomial distribution: models the probability of counts for rolling a k-sided die n times.
pmf: $f(y_1, \ldots, y_k; n, \ldots, p_1, p_k) = Pr(Y_1 = y_1 \text{ and } \ldots \text{ and } Y_k = y_k)$

$$= \begin{cases} \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \times \ldots p_k^{y_k} & when \sum_{i=1}^k y_i = n \\ 0 & otherwise \end{cases}$$

$$also \quad f(x|n,p) = \frac{\Gamma(n+1)}{\prod_{i=1}^k \Gamma(y_i+1)} \prod_{i=1}^k p_i^{y_i}$$

Dirichlet distribution: a continuous multivariate probability distributions parameterized by a vector α_i of positive reals. Multivariate generalization of the beta distribution. ($\sum_{i=1}^k p_i = 1$)

$$f(p_1, \ldots, p_k; \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1} = \frac{1}{\mathbf{B}(\alpha)} \prod_{i=1}^k p_i^{\alpha_i - 1}$$

Conjugate prior: the posterior distributions $P(\theta|x)$ are in the same family as the prior distribution $P(\theta)$, the prior and the posterior are conjugate distributions.

Distribution	**Conjugate prior**
Bernoulli	Beta
Multinomial	Dirichlet
Gaussian, σ^2	Gaussian
Gaussian, μ	Gamma
Gaussian, μ, σ^2	Gaussian-Gamma

$$P(\theta|x)P(x) = P(X = x|\theta)P(\theta)$$

Introduce $f(\theta)$ and $f(\theta|x)$, $P(X)$ is determined given observations, then

$$f(\theta|x) \propto P(X = x|\theta) f(\theta)$$

$$\begin{aligned} & f(p, Y|\alpha) = P(Y|p) f(p|\alpha) \\ & = \prod_{y_i \in Y} f(y_i|p_1, \ldots, p_k) \times \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} f(p_1, \ldots, p_k|\alpha_1, \ldots, \alpha_k) \\ & = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{y_i \in Y} \prod_{j=1}^k p_j^{y_i} \times \prod_{j=1}^k p_j^{\alpha_j - 1} \\ & = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{j=1}^k p_j^{\alpha_j - 1 + \sum_{y_i \in Y} y_i^{(j)}} \qquad (where \quad y_i^{(j)} = \#(y_i) = n_j) \\ & = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{j=1}^k p_j^{\alpha_j - 1 + n_j} \sim Dirichlet(\alpha_j + n_j) \end{aligned}$$

Since $\int \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1} dp = 1$, then $\int \prod_{i=1}^k p_i^{\alpha_i - 1} dp = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$

$$\begin{aligned} f(y|\alpha) &= \int \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{j=1}^k p_j^{\alpha_j - 1 + n_j} dp \\ &= \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\Gamma(\sum_{i=1}^k \alpha_i + n_j)}{\prod_{i=1}^k \Gamma(\alpha_i + n_j)} \end{aligned}$$

4 Statistical tests

4.1 Classical ones[1]

Compare α diversity between two groups, population abundance between two sets of relative abundance data: **standard t-test**, **Wilcoxon rank-sum test**.
Compare intergroup and intragroup β diversity, propotional abundance among more than two group: **ANOVA**, **Kruskal-Wallis test**.
Comparing categorical microbiome data, e.g. test a single *a priori* specified taxon is present at fidderent rates across groups: **chi-square test**.

4.2 Multivariate statistical tests

Why multivariate non-parametric test? High-dimensional, non-normality, multiple responses (associations of microbiome composition, e.g. OTUs, directly with potential environmental factors)

Permutational Multivariate Analysis of Variance(PERMANOVA)[4]

Goal: Test the significance of individual terms in a multi-factor ANOVA like framework for multiple response variables. (fit multivariate models to microbiome data, based on distance matrices and permutation)

Assumption: the observations units are exchangeable under a true null hypothesis. (Exchangeability can be ensured through the random allocation of treatments to units in experiments)

Null hypothesis: there are NO differences in the position and/or spread, in a multivariate space, of the compared groups attributes.

Let Y be a matrix of N rows (sampling units) by p columns (variables), let $D = \{d_ij\}, i = 1, \ldots, N, j = 1, \ldots, p$ consist of the distances between every pair (i, j) of sampling units.

Matrix $A = \{a_{ij}\} = \{-\frac{1}{2}d_{ij}^2\}$, Gower's matrix $G = g_{ij} = a_{ij} - \bar{a_{i.}} - \bar{a_{.j}} + \bar{a_{..}}$

The total sum-of-squares, $SS_T = \frac{1}{Np}\sum_{i=1}^{Np-1}\sum_{j=i+1}^{Np} d_{ij}^2$

The within-groups sum-of-squares, $SS_W = \frac{1}{N}\sum_{i=1}^{Np-1}\sum_{j=i+1}^{Np} d_{ij\in ij}^2$

The between-groups sum-of-squares, $SS_A = SS_T - SS_W$

Build a pseudo F-test: $F = \frac{SS_A/(p-1)}{SS_W/(N-p)}$ (Between groups/ Within groups)

Mantel test[5]

Test of the correlation between two matrices. In applications, they are matrices of interrelations between the same vectors of objects.

$$Z_m = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}d_{ij}$$

where g_ij and d_{ij} are the genetic and geographic distances between populations i and j, considering n populations.

If there is a relationship between matrices G and D, the sum of product Z_m will be relatively high. Randomization of rows and columns will destroy this relationship, and $Z_m^{'}$ will be lower. p-value = $1/(1 +$ times of $Z_m^{'} > Z_m)$

Standardized version of the Mantel's test (Z_N):

$$Z_N = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(g_{ij} - \bar{G})(d_{ij} \times \bar{D}))}{\sqrt{Var(G)}\sqrt{Var(D)}}$$

It is Pearson correlation r between the standardized elements of the matrices G and D.

4.3 Differential expression analysis within GLM

For RNA-seq data, we propose two models:

$$Y_{gij} \sim Poisson(C_{ij}\mu_{gij})$$
$$Y_{gij} \sim NegBin(C_{ij}\mu_{gij}, \phi_g)$$

where Y_{gij} denote the read count mapped to treatment i, replicate j and gene g. C_{ij} is the normalizing factor, estimated for each library ij based on the whole library.

Deviance and GLM

In GLM context, deviance amounts to

$$-2\ln\Lambda$$

where Λ is the likelihoods between two nested models (l_0,smaller)/(l_1).

Test DE with GLM: likelihood ratio test

Consider fitting Poisson to each gene:

$$Y_{ij} \sim Poisson(C_{ij}\mu_i)$$

$$\log(\mu_i) = \tau_i, \quad or\, \log(C_{ij}\mu_i) = \log(C_{ij}) + \tau_i$$

where C_{ij} are offset.

$H_0 : \tau_1 = \tau_2$
(here τ_i denote the treatment). Consider *reduced* model ($\tau_1 = \tau_2 = \tau$) vs. *saturated* model (τ_1, τ_2)

$$T.S. = -2\ln\frac{L(\hat{\tau}|X_1, X_2, \ldots, X_n)}{L(\hat{\tau_1}, \hat{\tau_2}|X_1, X_2, \ldots, X_n)} \sim \chi^2(df_{full} - df_{reduced})$$

Accounting for overdispersion: Quasi-likelihood and F test

$$E(Y) = \lambda, \quad Var(Y) = \phi\lambda$$

$$\hat{\phi} = \frac{-2\ln\Lambda}{\#observations - \#parameters} = \frac{-2\ln\Lambda}{n-p} = \frac{residual \quad deviance}{df.residual}$$

when $\phi > 1$, the data are over-dispersed.

$$T.S. = \frac{T}{\hat{\phi}df_T} \overset{a}{\sim} F(df_T, n-p)$$

where T is a T.S. that has an approximate χ^2 with df_T under null when there is no overdispersion (i.e. the LRT T.S.)

4.4 Multiple testing for high-dimensional gene data

When performing many hypothesis tests, each test has a probability of producing type I error, performing a large number of hypothesis tests guarantees the presence of type I error among findings.

$$
\begin{aligned}
&P(Making\ an\ error) = \alpha \\
&P(Not\ making\ an\ error) = 1 - \alpha \\
&P(Not\ making\ an\ error\ in\ m\ tests) = (1-\alpha)^m \\
&P(Making\ at\ least\ 1\ error\ in\ m\ tests) = 1 - (1-\alpha)^m
\end{aligned}
$$

The key goal of multiple testing method is to control the flood of type I errors that arise when many hypothesis tests are performed simultaneously.
Let $H_{01}, \ldots, H_{0m}$ denote the null hypotheses, suppose m_0 are true, m_1 are false.
If we set level 5% for each test, then the number of type I error is expected to be **5%** of m_0, which will be very big due to the high-dimensionality.
So, multiple testing is: let $p_1, p_2, \ldots, p_n$ denote the p-values corresponding to the m tests. Let c denote a value between 0 and 1 that will serve as a cutoff for significance.

$$
\begin{aligned}
&Reject \quad H_{0i}, \quad if \quad p_i \leq c \\
&Accept \quad H_{0i}, \quad if \quad p_i > c
\end{aligned}
$$

For each test,

	Accpet H_0	Reject H_0	
True H_0	U	V	m_0
False H_0	T	S	m_1
Total	W	R	m

Note that only **W, R, m** are observable.
U, V, T, S, W, R are ramdom variables.

Error rate

Family-wise error rate (FWER) = P($V > 0$)(The probability of at least 1 type I error.)
False discovery rate (FDR) = E(Q)

$$
Q = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}
$$

pFDR = E(V/R|$R > 0$)=FDR/P($R > 0$)

Controlling FWER

Meaning: control FWER to choose the significance cutoff c so that FWER is less than or equal to some desired level α
Bonferroni method:

$$
c = \frac{\alpha}{m}
$$

Too strong, the general null is all the null hypotheses are true.
Holm's method: $c = p_{(k)}$, where k is the largest integer k so that

$$
p_{(i)} \leq \frac{\alpha}{m - i + 1}, for \quad all \quad i = 1, 2, \ldots, k
$$

Holm's is less conservative for resulting in more rejected null hypotheses.

Controlling FDR: more powerful

Meaning: FWER is appropriate when you want to guard against ANY false positives. But in many cases, we can live with a certain number of false positives.

FDR is to control the proportion of false positives among the set of rejected hypotheses (R).
Benjamini and Hochberg procedure: $c = p_{(k)}$, where k is the largest integer k so that

$$p_{(k)} \leq \frac{k\alpha}{m}$$

Distribution of p-value: improve B&H
Could help estimate m_0, it is a mixture of p-values from true null and that from false null.
When H_0 is true,

$$P(p \leq t) = t \sim Uniform(0, 1)$$

Here, type I error is t if we reject null whenever p-value is less than or equal to t.
Rather than finding the largest integer k such that

$$p_{(k)} \leq \frac{k\alpha}{m}$$

consider finding the largest integer k such that

$$\frac{p_{(k)}\hat{m_0}}{k} \leq \alpha$$

5 Compositional analysis[7]

5.1 When thinking as *Compositional...* [12]

For the fact that the sequencing instruments can only deliver reads only up to the capacity of the instrument, it is proper to assume the total read count observed in a high throughput sequencing run is a fixed size, random sample of the relative abundance of the taxons.

No more traditional methods:

1. Normalization: subsampling (will lose information and precision). Trimmed mean of M valuses, median method (the number of counts observed by the instrument cannot contain information on the actual number of taxons)
2. Calculation of distance or dissimilarity: Not in an euclidian space.
3. Correlation: Constant sum constraint.
4. Differentail (relative) abundance measures: Sensitive to sparsity(some are true, some are from the insufficient sequecing depth and screening from high-abundance taxons).

5.2 Definition of compositional data:[8]

1. A positive real vector. Often sum to 1. If the total amount is fixed and known, reduce one degree of freedom.
2. Compositions only carry relative information/ratio. Proporional positive vectors and vectors multiplied by positive constant are equivalent.
3. Representatives, a poisitive vectors adding to a given constant k. $Closure[x_1, \ldots, x_D] = [\frac{x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i}]$
4. The sample space is called the *simplex*: the set of vectors of positive (or zero) components and constant sum:

$$S^D = \{x = [x_1, \ldots, x_D] | x_i \geq 0, \sum_{j=1}^{D} x_j = k\}$$

Relative abundance changing from absolute abundance:
violate the assumption of sample independence, creating inevitable errors in covariance estimates, leading to bias inference.
Motivation in microbiome sequencing: After gathering data, a correction must be made for different samples having different numbers of sequences, while the total absolute abundance of all bacteria in each sample is unknown.

5.3 Vital transformations

Rarefy
To convert the number of sequences for each taxon within each sample to relative abundance .

N = total number of items, K = total number of groups, N_i = the number of items in group i, M_j = number of groups consisting in j elements

$$\sum_{i=1}^{k} N_i = N, \quad \sum_{j=1}^{\infty} M_j = K, \quad \sum_{j=1}^{\infty} jM_j = N$$

Let X_n = the number of groups in the subsample which have n items. The rarefaction curve is,

$$f_n = E(X_n) = K - \binom{N}{n}^{-1} \sum_{i=1}^{K} \binom{N - N_i}{n}$$

Assumptions: individuals are randomly distributed, sample size is sufficiently large, samples are taxonomically similiar, all of the samples are performed in the same manner.

However, rarefying does not correct for compositionality. The lack of correction could lead to spurious inference.

log-ratio transformation
Make the constant sum constraint does not distort the underlying covariance or correlation structure. The simplest transformation is to choose one componenet as a reference. (May not be obvious to see which to choose, and results may vary dependent on the choice of reference)

additive log-ratio transformation

$$alr(x) = [\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}] = \ln(x) \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 \end{bmatrix}$$

centered log-ratio transformation

$$g(x) = \sqrt[n]{x_1 \ldots x_D}$$

$$clr(x) = [\ln \frac{x_1}{g(x)}, \ldots, \ln \frac{x_D}{g(x)}] = \frac{\ln(x)}{D} \begin{bmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & D-1 \end{bmatrix}$$

isometric log-ration transformation

$$irl(x) = clr(x) \cdot V = \ln(x) \cdot V$$

for a given matrix V of D rows and $D-1$ columns such that $V \cdot V' = I_D + a\mathbf{1}$ where a may be any value, and $\mathbf{1}$ is a matrix full of 1's.

Sparsity
Parametric models must make accurate estimates of variance for meaningful inference, but such estimates are essentially impossible on samples with excess 0's.
"Rounded zeros": result from undersampling.
Pseudo-count addition, or imputation: replace rounded zeros with a small, nonzero value which is below the detection limit. However, hard to make robust analysis if the degree of sparsity changes dramatically.
"Structural or essential zeros": truly represent the absence of taxa from a particular sample.
Binomial condional logistic normal model for continuous data, Poisson-log normal distribution for discrete data.

6 Some simulations and test cases

6.1 Case study: grass root microbiome[16]

1. How bacterial $\alpha-$diversity was affected by sample type, time point, species and watering treatment:
   Calculate Shannon's diversity index, ANOVA table.

2. $\alpha-$diversity of nine most abundant bacterial types:
   Boxplot, reveal the trend of sample type and $\alpha-$diversity.

3. Difference of Shannon's diversity between treatments:
   Kernel density estimation plot.
   ANOVA table.

4. Which factors most influenced differences in community structure:
   $\beta-$diversity, PCoA.

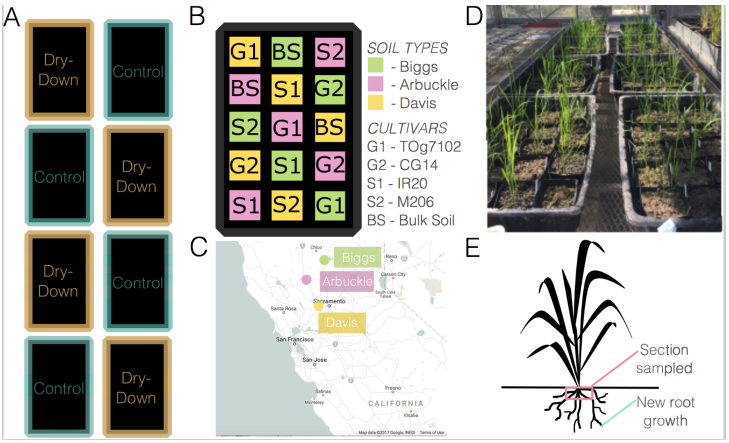
5. Amount of variance from each experimental variables:
   Canonical analysis of principal coordinates.

6. Relative abundance:
   Constitution histogram.
   Heatmap for fold enrichment.
   Phylogenetic tree of all significent Actinobacteria genera with fold enrichment in sample styles.

7. Positive cirrelation between phylogenetic distance and host species:
   (Preparation: phylogenetic tree for host species)
   Mantels tests to compare microbiome distances and host phylogeny & microbiome distance and physical field distance.

8. Common taxa:
   Heatmap: the percent of overlap between each host's taxa
   Histogram of host species and its corresponding #OTUs

6.2 Case study: Drought stress-rice root microbiome[17]



The data is available at https://github.com/cmsantosm/Drought-Root-Microbiome
According to the figure, it is a RCBD experiment design. We could set drought/watering, compartment, soil type, cultivar as covariates/predictors, and the 16S-sequenced-based OTUs as response.
After cleansing the dataset and expanding the simple symbols for treatments, I generate a dataset which is suitable for fitting OTU into different models.

```
> head(otu.fit.temp)
  Counts index Treatment Compartment Soil Cultivar
1      2     1        DS          RS    D       G1
2      2     2        DS          RS    D       G2
```

```
3       0      3          DS            RS     D         S1
4       6      4          DS            RS     D         S2
5       0      5          DS            RS     A         G1
6       0      6          DS            RS     A         G2
```

```
> summary(otu.fit.temp)
     Counts           index          Treatment  Compartment  Soil
 Min.   :0.000   Min.   :  1.00   DS:96      BS: 0        A:64
 1st Qu.:0.000   1st Qu.: 48.75   WC:96      RS:96        B:64
 Median :0.000   Median : 96.50              ES:96        D:64
 Mean   :0.474   Mean   : 96.50
 3rd Qu.:0.000   3rd Qu.:144.25
 Max.   :8.000   Max.   :192.00
 Cultivar
 G1:48
 G2:48
 S1:48
 S2:48
```

Start from simple situation: under the fact that all the treatments are balanced, fix the *Compartments*, *Soil*, *Cultivar*, only choose *Treatment* as predictor. Since for the experiment, we have 8 tubs. It can be treated as 8 replicates for one OTU, 4 for WS and 4 for DC.
Note that the subset of original OTU table may have several unexpected cases:

1. For specific OTUs, all 8 replicates are 0's. These OTUs don't provide any information. So discard them.

2. For specific OTUs, all 8 replicates contain no 0. These are fine with Poisson and NegBin, but unfeasible for zero inflated models. So for the purpose of model comparasion, discard them. Otherwise, we could fit GLM models to these separately.

Treatment has 2 levels, 1 = "DS", 2 = "WC" , column names are the ID for OTUs.

```
> otu.ds.wc.t.no0.zi[,1:5]
          Treatment 4479944 185100 623634 460067
drght.1           1       2      0      0      0
drght.25          2       6      1      1      0
drght.49          2       0      0      3      0
drght.73          1       8     10      1      0
drght.97          1       2      1      4      0
drght.121         2       7      0      2      2
drght.145         2       0      0      1      0
drght.169         1       6      0      3      0
```

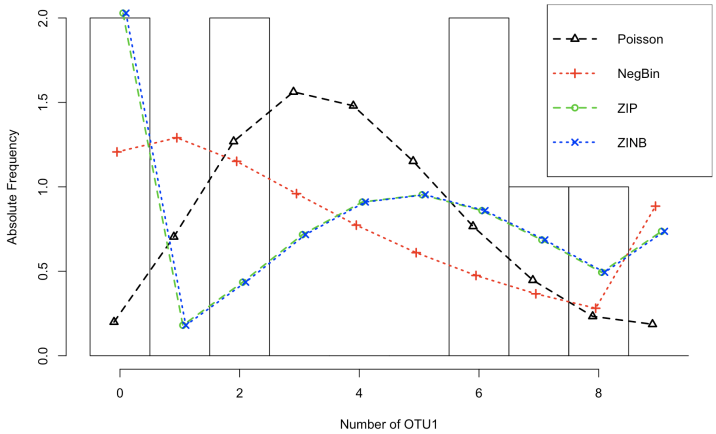
For GLM models, I use following to fit the data and generate raw p-values.

```
## Poisson
pvalue.pois<-function(df){
  plist<-c()
  for(i in 2:ncol(df)){
    fm.pois<- glm(df[,i] ~  as.factor(Treatment),
                  data = df, family = poisson)
    result <- anova(fm.pois,test = "Chisq")
    plist<- append(plist, result$`Pr(>Chi)`[2])
  }
  return(plist)
}
p.pois.no0.zi <- pvalue.pois(otu.ds.wc.t.no0.zi)

## NegBin
library(MASS)
pvalue.nb<-function(df){
  plist<-c()
  for(i in 2:ncol(df)){
    fm.pois<- glm.nb(df[,i] ~  as.factor(Treatment),
```

```
                    data = df)
    result <- anova(fm.pois)
    plist<- append(plist, result$`Pr(>Chi)`[2])
  }
  return(plist)
}
p.nb.no0.zi <- pvalue.nb(otu.ds.wc.t.no0.zi)
```

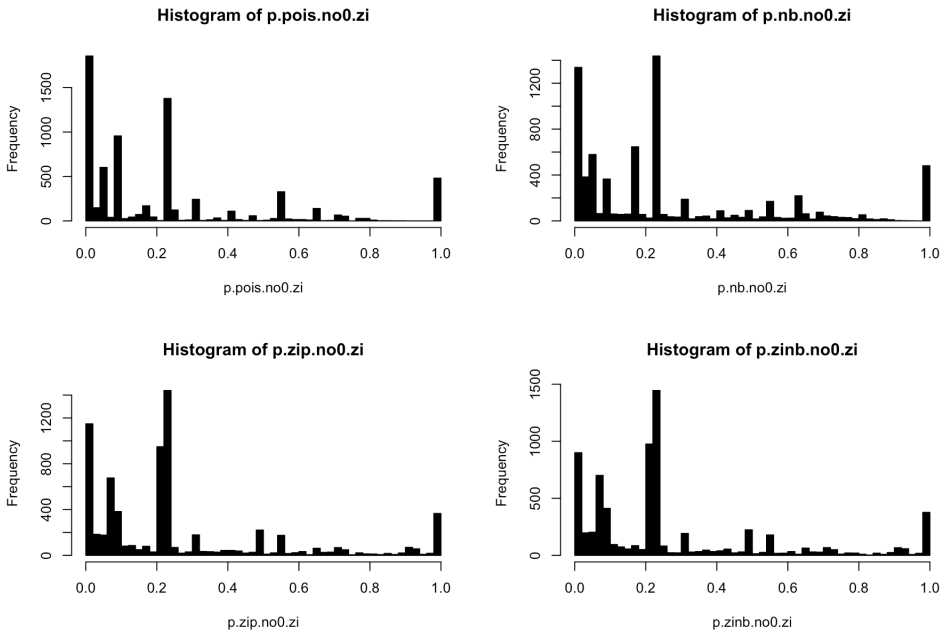
For zero inflated models, I use package *pscl* to do that.

```
library(pscl)
## For ZIP
pvalue.zip<-function(df){
  plist<-c()
  for (i in 2:ncol(df)){
    fm.zip  <- zeroinfl(df[,i] ~ as.factor(Treatment)|
                          1, dist="poisson", data = df)
    fm.zip.r <- zeroinfl(df[,i] ~ 1|
                           1, dist="poisson", data = df)
    p.temp <- 1-pchisq(-2*(fm.zip.r$loglik-fm.zip$loglik),1)
    plist<-append(plist,p.temp)
  }
  return(plist)
}
p.zip.no0.zi <- pvalue.zip(otu.ds.wc.t.no0.zi)

## For ZINB
pvalue.zinb<-function(df){
  plist<-c()
  for (i in 2:ncol(df)){
    fm.zinb  <- zeroinfl(df[,i] ~ as.factor(Treatment)|
                          1, dist="negbin", data = df)
    fm.zinb.r <- zeroinfl(df[,i] ~ 1|
                           1, dist="negbin", data = df)
    p.temp <- 1-pchisq(-2*(fm.zinb.r$loglik-fm.zinb$loglik),1)
    plist<-append(plist,p.temp)
  }
  return(plist)
}
p.zinb.no0.zi <- pvalue.zinb(otu.ds.wc.t.no0.zi)
```



Since all the test here are assumed independence, if we want to observe the significant DE OTUs under the two treatments, Benjamini-Hochberg method could apply here to adjust p-values.
P-value hisogram comes as following:

```
> a.p.pois.no0.zi<-p.adjust(p.pois.no0.zi,method = "BH")
```




```
> a.p.nb.no0.zi<-p.adjust(p.nb.no0.zi,method = "BH")
> a.p.zip.no0.zi<-p.adjust(p.zip.no0.zi,method = "BH")
> a.p.zinb.no0.zi<-p.adjust(p.zinb.no0.zi,method = "BH")
> p.value[,1:5]
                      [,1]        [,2]      [,3]      [,4]      [,5]
a.p.pois.no0.zi 0.4605390 0.01299827 0.8549889 0.1921817 0.3892230
a.p.nb.no0.zi   0.7261128 0.31397899 0.8681926 0.3377536 0.4130435
a.p.zip.no0.zi  0.4271010 0.09927679 0.8666337 0.3262009 0.4093378
a.p.zinb.no0.zi 0.4427613 0.33186585 0.8680520 0.3318659 0.4146156

> otu.sig.for.all<-which(apply(p.value,2,max)<=0.05)
> otu.sig.for.pois <- which(p.value[1,]<=0.05)
> otu.sig.for.nb <- which(p.value[2,]<=0.05)
> otu.sig.for.zip <- which(p.value[3,]<=0.05)
> otu.sig.for.zinb <- which(p.value[4,]<=0.05)
> length(otu.sig.for.all)
[1] 170
> length(otu.sig.for.pois)
[1] 1503
> length(otu.sig.for.nb)
[1] 875
> length(otu.sig.for.zip)
[1] 603
> length(otu.sig.for.zinb)
[1] 173
```

In order to explore and understand more about the models fitted, generation of OTU count data manually is applied. First, consider the GLM-Poission model. The case here is simple, let watering treatment denote the covariate, where $X = 1$ is watering and $X = 0$ is drought.

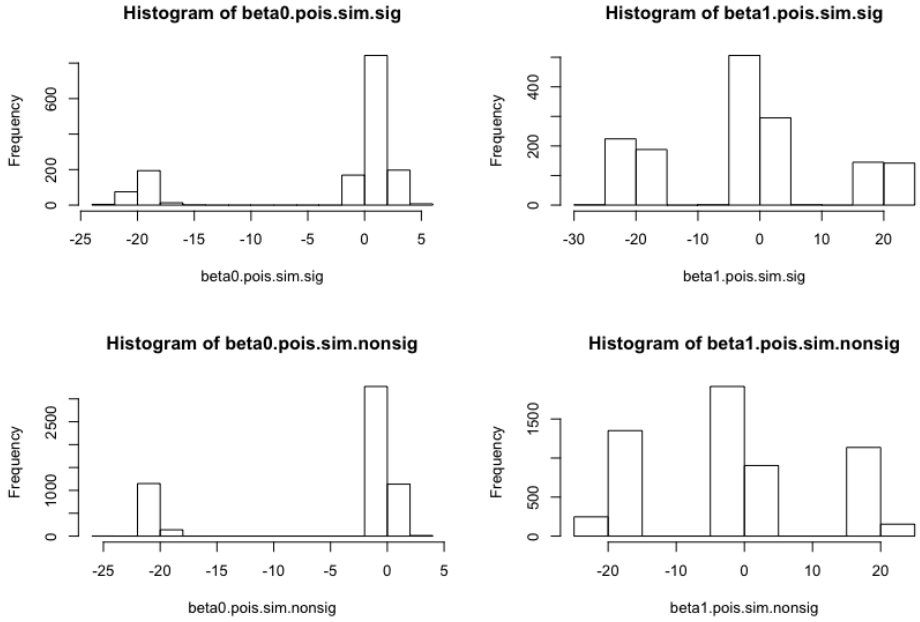
$$\log \lambda = \beta_0 + \beta_1 \quad for\ X = 1$$
$$\log \lambda = \beta_0 \quad for\ X = 0$$

These two λ could be plugged into the link function of GLM model. To match the sample size of orginal data, I simulate 4 replicate response OTU count for $X = 0$ and another 4 for $X = 1$.
With the help of previous adjusted p-values for poisson, I know that 1503/7211 are grouped into "DE". I create such a "DE" group of these 1503 OTUs, getting a list of β_0 and β_1 after fitting poisson model

respectively. For each round of generation, I sample 1 pair of β_0 and β_1.

```
index <- sample(ncol(beta),1)
b0 <- beta[1,index]
b1 <- beta[2,index]
mu <- c(exp(b0), exp(max(b0+b1,0)) )
otu <- c(rpois(4, mu[1]), rpois(4,mu[2]))
```

Another common method here is to find the prior distribution of β_0 and β_1, then randomly generate β_0 and β_1 for each round of simulation. The histogram shows as following:




```
> sim1<-sim.otu.glmpois(2800,beta.pois.sig)
> head(sim1)
  trt1 trt1 trt1 trt1 trt2 trt2 trt2 trt2          beta0      beta1 index of beta
1    1    3    0    4    3    3    7    4  5.233041e-15   1.446919             14
2    0    0    0    0    1    3    4    1 -2.030259e+01  20.862201            807
3    0    0    0    0    1    3    2    1 -1.930259e+01  19.862201           1086
4    1    1    5    3    0    1    4    1  9.162907e-01  -2.302585             23
5    6    8    1    5    0    0    2    1  9.162907e-01 -20.218876            733
6    0    2    0    1    2    0    2    3  5.596158e-01 -19.862201            279
```

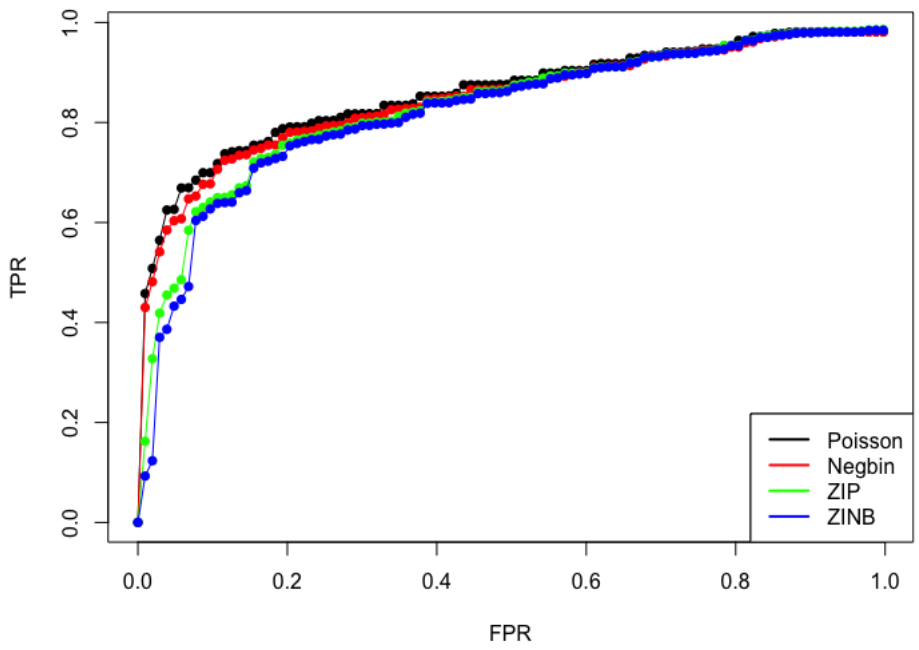
```
> sim2<-sim.otu.glmpois(11000,beta.pois.nonsig)
> head(sim2)
  trt1 trt1 trt1 trt1 trt2 trt2 trt2 trt2        beta0          beta1 index of beta
1    0    0    0    0    0    1    2    2 -20.3025851  1.891629e+01            602
2    0    0    0    0    2    2    1    1 -20.3025851  1.891629e+01           3477
3    0    1    3    0    1    1    2    1  -1.3862944 -1.891629e+01           1710
4    3    7    8    8    0    2    4    1   1.3217558 -9.162907e-01           2639
5    2    1    0    1    0    2    2    2  -0.6931472 -6.931472e-01           4305
6    1    0    0    1    0    1    0    2  -1.3862944 -3.139971e-16           2428
```

Here, I simulate 2293 OTUs for "DE" group and 10325 OTUs for "non-DE" group, where the proportion of "DE" is approximately 20%, which is similiar to the original dataset. In order to draw the ROC curves,

	Accpet H_0	Reject H_0	
True H_0	TN	FP	10325
False H_0	FN	TP	2293
Total	W	R	m

$$FPR = \frac{FP}{N} = \frac{FP}{10325}$$
$$TPR = \frac{TP}{P} = \frac{TP}{2293}$$

```
> for(i in 1:length(FP=seq(from=0,to=10325,by=100))){
>     FPR <- FP[i]/10325
>     sig <- which(a.p < FPR)
>     TP <- which(sig <=2293)
>     TPR <- length(TP)/2293
> }
```



Under H_0, test statistic T has the distribution $F_T(t)$.
T is a R.V that represents all of the possible values of a test statistic under H_0. Consider that small T indicate against H_0

$$p = \Pr_{H_0}(T \leq t) = F_T(t)$$
$$\text{Let } Y = F_T(t)$$
$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(F_T(t) \leq y) = \mathrm{P}(T \leq F_T^{-1}(y)) = F_T(F_T^{-1}(y)) = y$$

So $p - value \sim$ Uni(0,1) under H_0.
Here, our situation is that we have different test statistic to do the multiple testing. The conclusion holds since each separate part is Uni(0,1).

.

Algorithm 1 Simulation of microbiome count data: 1 factor, 2 levels

function GET.PVAL(raw.df) ▷ raw.df: id for row and trt for column, 4 replicates for each trt
  $pval \leftarrow NULL$
  $beta0 \leftarrow NULL$
  $beta1 \leftarrow NULL$
  for $i = 1$ to ncol($raw.df$) **do**
    $fit \leftarrow$ Fit GLM-Poisson model to $raw.df[i]$
    $beta0[i] \leftarrow fit$ $coefficient[1]
    $beta1[i] \leftarrow fit$ $coefficient[2]
    $pval[i] \leftarrow$ anova(fit, test="Chisq")
  return $adjust.pval \leftarrow$ p.adjust ($pval$, method="BH")
function SIM.POISSON(adjust.pval, n, raw.df)
  $de.index \leftarrow$ which ($adjust.pval < 0.05$)
  $de.para \leftarrow$ rbind ($beta[0][de.index], beta[1][de.index]$)
  $nonde.index \leftarrow$ which ($adjust.pval \geq 0.05$)
  $nonde.para \leftarrow beta[0][nonde.index]$
  for $i = 1$ to ($n *$ proportion of de) **do**
    $j \leftarrow$ sample(c(1:ncol($de.para$)),1)
    $mu \leftarrow$ c($de.para$[1,j], $de.para$[1,j]+$de.para$[2,j])
    $otu.de \leftarrow$ c(rpois(4, mu[1]), rpois(4, mu[2]))
  for $i = 1$ to (rest of n) **do**
    $mu \leftarrow$ sample($nonde.para$,1)
    $otu.nonde \leftarrow$ rpois(8, mu)
  return rbind(de.otu, nonde.otu)

function SIM.ZINB(n, m)[19] ▷ n: OTU index, m: sample
  for $i \leq n$ **do**
    for $j \leq m$ **do**
      $p \leftarrow$ sample(c(0.1, 0.2, 0.3, 0.4, 0.5), 1)
      $theta \leftarrow$ runif(1, 0.8, 1)
      $beta \leftarrow$ sample(c(0.5, 1, 1.5), 2)
      if $beta[1] == beta[2]$ **then**
        $mu \leftarrow \exp(beta[1])$
        $otu.nde \leftarrow$ rbinom(1, 1, p)
        Keep all 0's
        Replace all 1's with rnegbin(1, mu, $theta$)
      else
        $mu \leftarrow \exp(beta[1])$, $\exp(beta[2])$
        $otu.de[i, j] \leftarrow$ rbinom(1, 1, p)
        Keep all 0's
        Replace the first half 1's with rnegbin(1, mu[1], $theta$)
        Replace the second half 1's with rnegbin(1, mu[2], $theta$)
  return n by m OTU table

function SIM.NONPARA(df, n, p)[20] ▷ n: OTU's, p: 2-c nrom probability
  $otu \leftarrow$ Sample n from nrow(df) ▷ df: subframe with chosen interested factor
  Permutate each of n otu
  for x% of $otu[i]$ **do**
    $para \leftarrow \text{sample}(c((mu = -0.5, sig2 = 0.7^2), (mu = 0.5, sig2 = 0.7^2)), p)$
    $delta \leftarrow$ rnorm($para$)
    $scale1 \leftarrow \exp(-delta)$
    $scale2 \leftarrow \exp(delta)$
    first half of $otu[i]$ scaled by $scale1 \rightarrow otu.de.trt1$
    second half of $otu[i]$ scaled by $scale2 \rightarrow otu.de.trt2$
    $otu.de[i] \leftarrow$cbind($otu.de.trt1, otu.de, trt2$)
  $1 - x$% of $otu \rightarrow otu.nde$

Algorithm 2 Permutation test

function PERMUTATIONTEST(df, treatment)[21] ▷ Interaction effect: fix or do analysis at first
  Normalize the samples(columns) with upper-quantile method
  Scale the kth OTU(row) by $\{\sum\sum\sum\sum(Y_{ijk} - Y_{i'j'k})^2\}^{-1/2}$
  $D \leftarrow NULL$
  $Dm.bar \leftarrow NULL$
  Choose one treatment factor, separate *df* with corresponding levels
  for each level i **do**
    $D_i \leftarrow$ mean(Euclidean distances between pairs of column within level i)
  $D.bar \leftarrow$ mean(all D_i)
  for $j < 50000$ **do**
    $index \leftarrow$ sample(c(1:ncol(df)), ncol(df))
    Divide the new index into i levels
    for each level i **do**
      Repeat line 6
    Compute mean $\leftarrow Dm.bar[j]$
  $num \leftarrow$ length(which($D.bar \geq Dm.bar$))
  $p \leftarrow num/50000$
return p

References

[1] Noyes, N., Cho, K. C., Ravel, J., Forney, L., & Abdo, Z. (2017). Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis. bioRxiv, 211631.

[2] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

[3] Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. BMC bioinformatics, 18(1), 4.

[4] Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online.

[5] Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. D. C., ... & Bini, L. M. (2013). Mantel test in population genetics. Genetics and molecular biology, 36(4), 475-485.

[6] Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS one, 7(2), e30126.

[7] Tsilimigras, M. C., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of epidemiology, 26(5), 330-335.

[8] Aitchison J. (1986), The Statistical Analysis of Compositional Data, Chapman & Hall; reprinted in 2003, with additional material, by The Blackburn Press.

[9] Sneath, P. H., & Sokal, R. R. (1973). Numerical taxonomy. The principles and practice of numerical classification.

[10] Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology, 44(4), 846-849.

[11] Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., ... & Starr, M. P. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic and Evolutionary Microbiology, 37(4), 463-464.

[12] Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology, 8, 2224.

[13] Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nature methods, 10(12), 1200.

[14] Xia, F., Chen, J., Fung, W. K., & Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics, 69(4), 1053-1063.

[15] Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zeroinflated count data: Beyond zeroinflated Poisson regression. British Journal of Mathematical and Statistical Psychol-

ogy, 65(1), 163-180.
[16] Naylor, D., DeGraaf, S., Purdom, E., & Coleman-Derr, D. (2017). Drought and host selection influence bacterial community dynamics in the grass root microbiome. The ISME journal, 11(12), 2691.
[17] Santos-Medelln, C., Edwards, J., Liechty, Z., Nguyen, B., & Sundaresan, V. (2017). Drought stress results in a compartment-specific restructuring of the rice root-associated microbiomes. MBio, 8(4), e00764-17.
[18] Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLoS Comput Biol 8(12): e1002808. https://doi.org/10.1371/journal.pcbi.1002808
[19] Zhang, X., Mallick, H., & Yi, N. (2016). Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. Journal of Bioinformatics and Genomics, (2 (2)).
[20] Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNAseq data. American journal of botany, 99(2), 248-256.
[21] Nettleton, D., Recknor, J., & Reecy, J. M. (2007). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics, 24(2), 192-201.