**A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes**

This is a brief introduction to using the proposed method for detection of significantly differentially abundant features between different metagenomic communities/conditions.

**1.1 Input feature count matrix and phenotype matrix for comparison of two different metagenomic conditions:**

The feature count matrix contains the number of observations of each feature within each subject. The element in the *i*th row and the *j*th column corresponds to the number of reads (or relative abundance) of feature *i* in subject *j*. The entire matrix is tab-delimited and contains labels for each feature and each subject in the following format:

\tsubject1\tsubject2\tsubject3\t ....\tsubjectN\n

feature1\t391\t729\t...

feature2\t668\t1978\t...

feature3\t174\t12\t...

feature4\t0\t58\t...

The phenotype matrix is also tab-delimited and contains the phenotype condition of each subject in the following format:

\tSample\Phenotype \n

Subject1\tDiseased

Subject2\tDiseased

…

Subject11\tNormal

Note the tab at the beginning of the first row. Sample matrices for feature count matrix and phenotype matrix are included on the website: *abundance.csv* and *phenotype.csv*.

**R commands**

Once you have R up and running:

1. Input the source file *TwoStage\_Package.2.0.R*

> source(“D:/dataset/TwoStage\_Package2.0.R”)

2. Load a feature count matrix

> count <- read.csv(file = "abundance.csv")

> phenotype <- read.csv(file = "phenotype.csv")

3. Analyze the loaded matrix

> TwoStage\_Package(count, phenotype, "sig.csv", 1)

In this example dataset, “1” means the normalization option of using the mean of the effective library sizes as a reference library size in TMM normalization; option of "2" represents an approach to regenerating counts with a common dispersion. “sig.csv” is the filename of the output containing information of the significantly differentially abundant features. Each row represents a significantly differentially abundant feature with its corresponding statistics.

Example:

The output file is a tab-delimited file containing 7 columns in the following order:

1. Annotation (name of feature)

2. mean\_group1 (the average feature abundance of population1)

4. sd\_group1 (standard deviation of feature abundance of population1)

3. mean\_group2 (the average feature abundance of population2)

5. sd\_group2 (standard deviation of feature abundance of population2)

6. p.val (p-values)

7. p.adj (p-adjusted values)

**1.2 Input feature count matrix and phenotype matrix for comparison of more than two different metagenomic conditions:**

The format of feature count matrix and the phenotype matrix are similar to those of two condition comparison (Section 1.1). The entire matrices are tab-delimited. Sample matrices for feature count matrix and phenotype matrix are also included on the website: *abundance\_multi.csv* and *phenotype\_multi.csv*. The samples matrices are an example of comparison among three different metagenomic conditions.

**R commands**

Once you have R up and running:

1. Input the source file *TwoStage\_Package2.0.R*

> source(“D:/dataset/TwoStage\_Package2.0.R”)

2. Load a feature count matrix

> count2 <- read.csv(file = "abundance\_multi.csv")

> phenotype2 <- read.csv(file = "phenotype\_multi.csv")

3. Analyze the loaded matrix

> TwoStage\_Package(count2, phenotype2, "sig\_multi.csv", 1)

In this example dataset, “sig\_multi.csv” is the filename of the output containing information of the significantly differentially abundant features. Each row represents a significantly differentially abundant feature with its overall adjusted p-values and corresponding pairwise adjusted p-values, as well as the mean and standard deviation of each group.

Example:

The output file is a tab-delimited file containing 11 columns in the following order (Note that the number of columns in the output file depends on the number of different metagenomic conditions to be compared.):

1. Annotation (name of feature)

2. p.overall (the overall adjusted p-values)

4. NormalvsStage1 (the pairwise adjusted p-values between normal group and stage 1 group)

3. NormalvsStage2 (the pairwise adjusted p-values between normal group and stage 2 group)

5. Stage1vsStage2 (the pairwise adjusted p-values between stage 1 group and stage 2 group)

6 -11 are the summary information of mean and standard deviation for each group.