MetaComp User's Guide

Peng Zhai

November 14, 2017

Contents

1 Introduction 2
2 Contact information 2
3 Citing MetaComp 2
4 Prerequisites and installation 2
4.1 Prerequisite 2
4.2 Source code 2
4.3 Installation 2
5 Input data 3
5.1 Abundance profile matrix data 3
5.2 Obtain profile from BLAST 4
5.3 Obtain profile from HMMER 5
5.4 Obtain profile from Kraken 6
5.5 Obtain profile from MG-RAST 7
5.6 Obtain profile from PhymmBL 8
5.7 Obtain profile from MZmine 9
6 Multivariate statistics 10
6.1 Cluster analysis 10
6.2 Principal component analysis 11
7 Hypothesis testing 12
7.1 Two samples test 12
7.2 Multiple samples test 13
7.3 Two groups of samples test 14
8 Environmental factors analysis 15

1 Introduction

MetaComp is a graphical software for analyzing meta-omic (i.e. metagenomics, metatranscriptomics, metaproteomics and metabolomics) profiles with related environmental information, such as phylogenetic profiles indicating the number of marker genes assigned to different taxonomic units or functional profiles indicating the number of sequences assigned to different subsystems or pathways. The aim of this document provide an easy but comprehensive introduction to MetaComp and show how it can be used to analyze meta-omic data. MetaComp is applicable to any meta-omics data by accepting abundance profile matrices (APM) saved as txt or BIOM format files [1]. Moreover, MetaComp can autmatically converts the output of several widely used platform into MetaComp-compatible input file.

2 Contact information

MetaComp is in active development. We encourage you to send any suggestions, comments and bug reports to hqzhu@pku.edu.cn. If reporting a bug, please provide as much information as possible and the related data which causes the bug. This will allow us to quickly resolve the issue.

3 Citing MetaComp

4 Prerequisites and installation

4.1 Prerequisite

- Windows 7 or higher version.
- Install Microsoft Office Excel 2010 or higher version.
- Install the required R packages using the following commands in the R console:

  install.packages("pheatmap")

4.2 Source code

https://github.com/pzhaipku/MetaComp

4.3 Installation

- Windows: Download file "" setup from our website: `http://cqb.pku.edu.cn/ZhuLab/MetaComp/download.html`.

- Linux: Download file “” setup from our website: `http://cqb.pku.edu.cn/ZhuLab/MetaComp/download.html`. File Annotation.RData is the annotation information for Linux. Please put it in the same folder with file MetaComp.R. Finally, please input the following commands:

  source(“.//MetaComp.R”)

5 Input data

5.1 Abundance profile matrix data

MetaComp reads input file in text format, and the values in the file should be separated by tab. The first row of the file shows the name of samples, while the first column represents the selected statistical feature. The cell of the table indicates the hit number of one sample to the given feature. Users must select **Abundance profile matrix (.txt or .biom)** radio button from **Profile** dialog box in **Load Data** option within **File** menu before choose the input profile. Moreover, the .biom format input must be convert to biom.table format before loading.(Figure 1-3)

This format of input is the only format that can be load in Linux version. The command line is as follow:

input_data = readFeature(file pathway, featureType = “pfam” or “cog”, format = “txt” or “biom”)

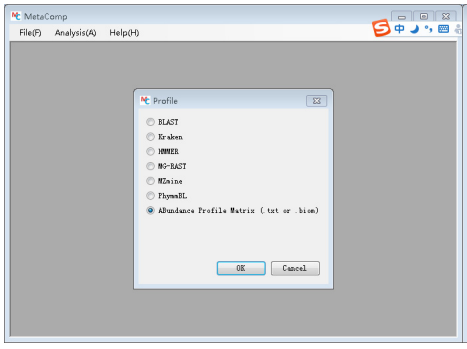

Figure 1:

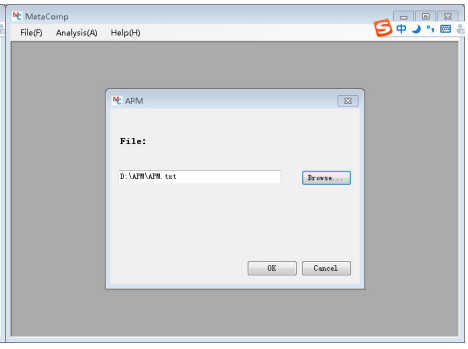

Figure 2:

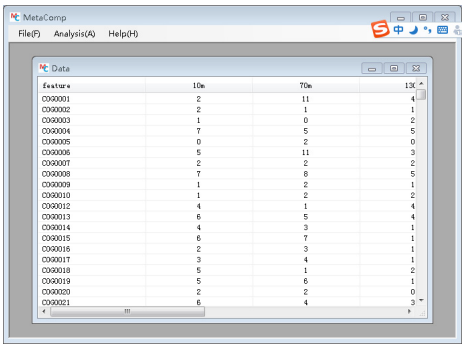

Figure 3:

5.2 Obtain profile from BLAST

MetaComp also accepts meta-omics profiles obtained from BLAST ([2], `https://blast.ncbi.nlm.nih.gov/Blast.cgi`) result. MetaComp works directly with BLAST result obtained by clicking on download in result web page, followed by selecting Hit Table(text) output type choice. Moreover, the BLAST result file can be obtained from table format (-outfmt 7). MetaComp can convert these BLAST results to standard Abundance profile matrices (APM) data through selecting **BLAST** radio button from **Profile** dialog box in **Load Data** option within **File** menu. After opening up the **BLAST** dialog box, you can select the BLAST result files you wish to input. (Figure 4-6)

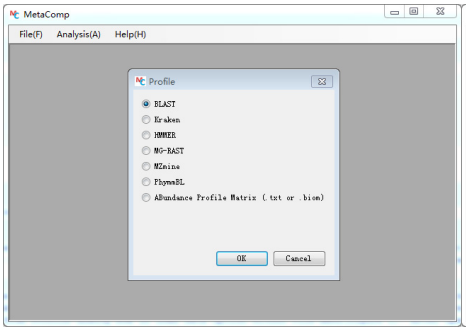

Figure 4:

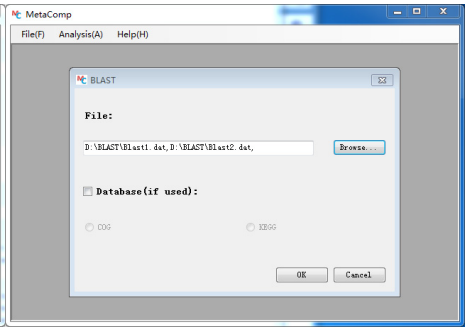

Figure 5:

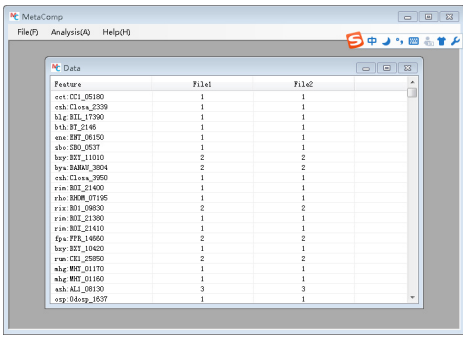

Figure 6:

5.3 Obtain profile from HMMER

The input profile can also be acquired from HMMER ([3], http://hmmer.org/). After downloading **hmmer-3.1b2.tar.gz** from `http://hmmer.org/` and unpacking it, you can get the desired results from **hmmsearch** command. MetaComp can convert these file into MetaComp-compatible profiles through selecting **HMMER** radio button convert these BLAST results to standard Abundance profile matrices (APM) data through selecting **BLAST** radio button from **Profile** dialog box in **Load Data** option within **File** menu. Click on the **OK** button after selecting the result file you wish to convert. (Figure 7-9)

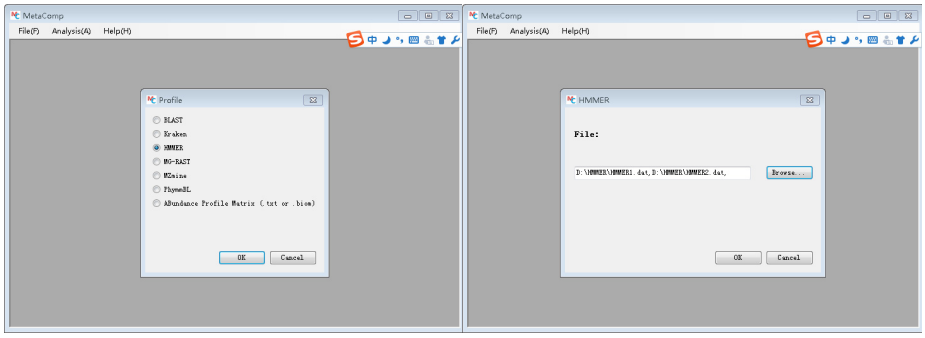

Figure 7: Figure 8:

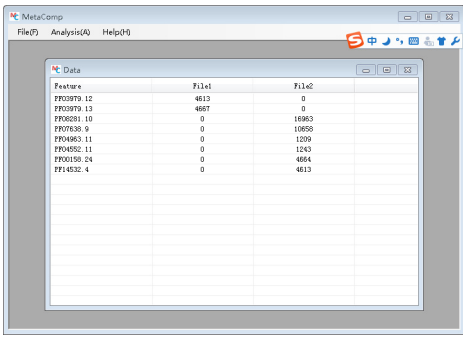

Figure 9:

5.4 Obtain profile from Kraken

Kraken ([4], `http://ccb.jhu.edu/software/kraken/`) result files are achieved from kraken-translate command. The selection of Kraken result file can initiate after choosing **Kraken** radio button from **Profile** dialog box in **Load Data** option within **File** menu. (Figure 10-12)

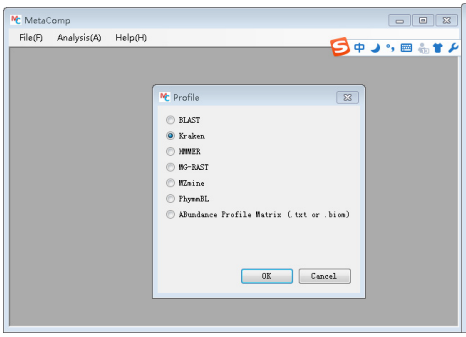

Figure 10:

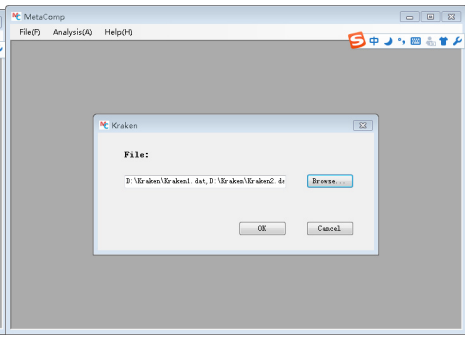

Figure 11:

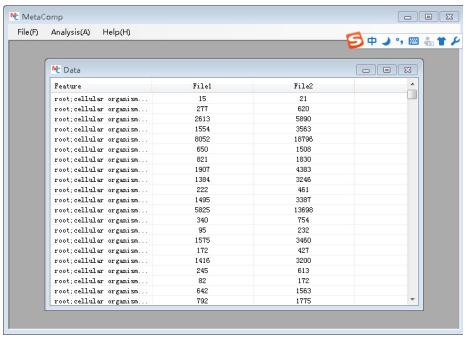

Figure 12:

5.5 Obtain profile from MG-RAST

MetaComp provides support for analyzing MG-RAST taxonomic or functional profiles. Visit the MG-RAST website ([5], `http://metagenomics.anl.gov/`) and browse the list of pubic metagenomes. Profiles for multiple samples can be obtained and downloaded as tab-separated values (tsv) file using the table data visualization. To work with MG-RAST profiles, they must be converted into a MetaComp-compatible profile. From within MetaComp, select the **MG-RAST** radio button from **Profile** dialog box in **Load Data** option within **File** menu. This opens up the **MG-RAST** dialog box. Click on the **OK** button after selecting the MG-RAST profile you wish to convert. (Figure 13-15)

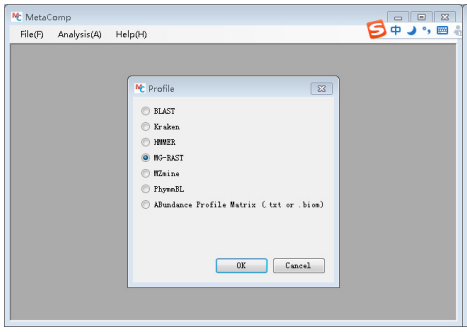

Figure 13:

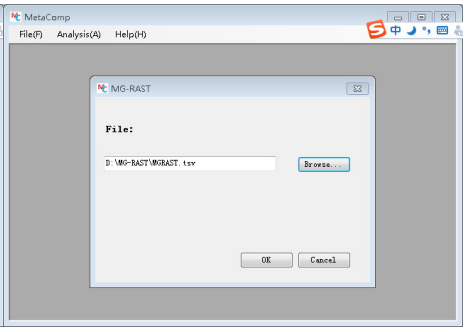

Figure 14:

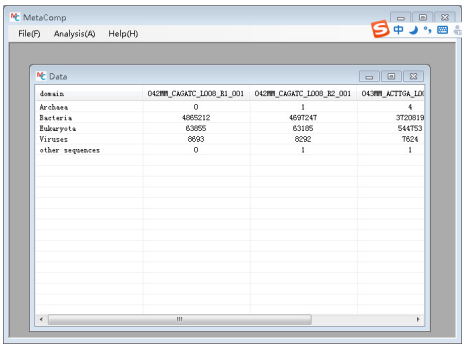

Figure 15:

5.6 Obtain profile from PhymmBL

PhymmBL([6], `http://www.cbcb.umd.edu/software/phymm/`) result files are achieved from **scoreReads.pl** command. The selection of PhymmBL result file can initiate after choosing **PhymmBL** radio button from **Profile** dialog box in **Load Data** option within **File** menu. (Figure 16-18)

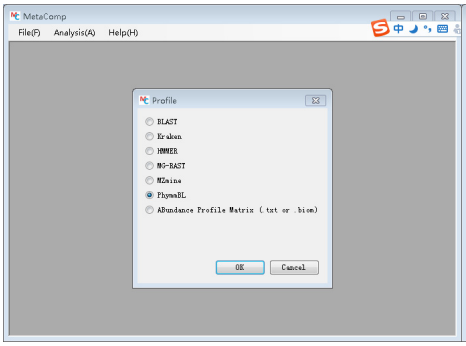

Figure 16:

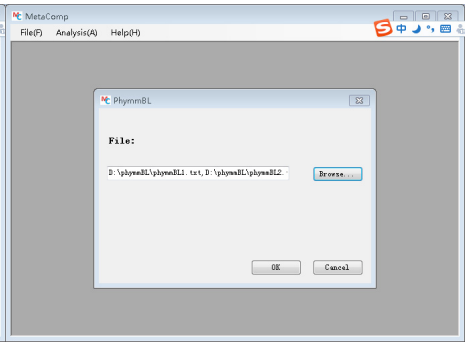

Figure 17:

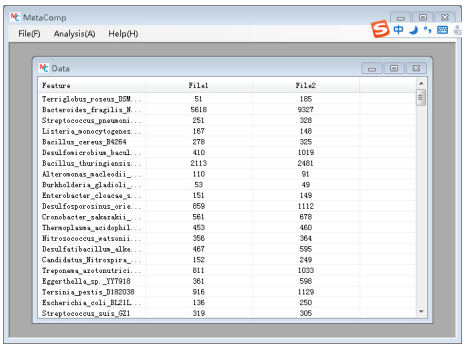

Figure 18:

5.7 Obtain profile from MZmine

MZmine([7], `http://mzmine.github.io/`) result files are achieved as Figure 19. The selection of MZmine result file can initiate after choosing **MZmine** radio button from **Profile** dialog box in **Load Data** option within **File** menu. (Figure 20-22)

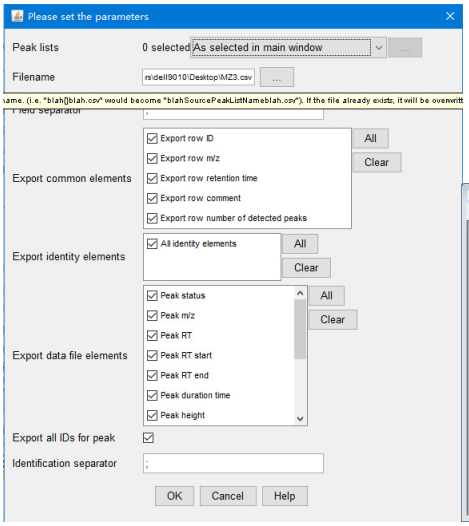

Figure 19:

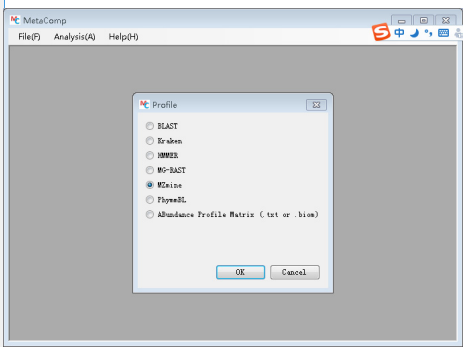

Figure 20:

All these input format example mentioned above can be download from:

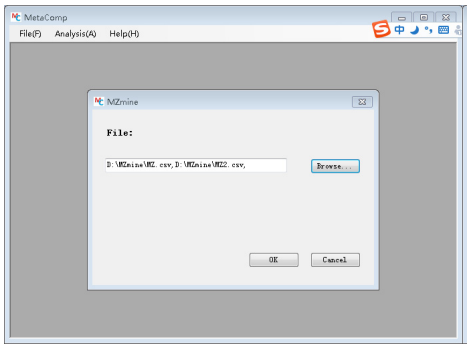

Figure 21:

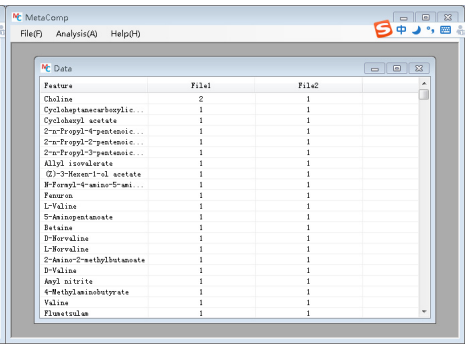

Figure 22:

6 Multivariate statistics

6.1 Cluster analysis

Cluster analysis can be perform in two model: K-means clustering and hierarchical clustering. K-means clustering model requires users to input the cluster number. Cluster analysis can be operate through the **Clustering analysis** dialog in **Analysis** menu.(Figure 23-26)

Linux commands line:

Hcluster(input_data)(for hierarchical clustering)

KMeans(input_data,cluster number)(for k-means clustering)

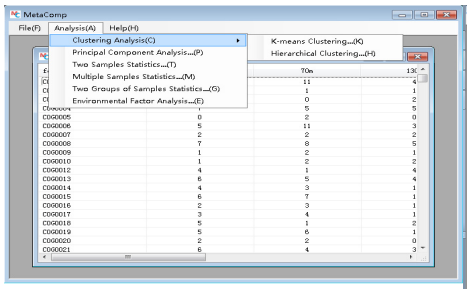

Figure 23:

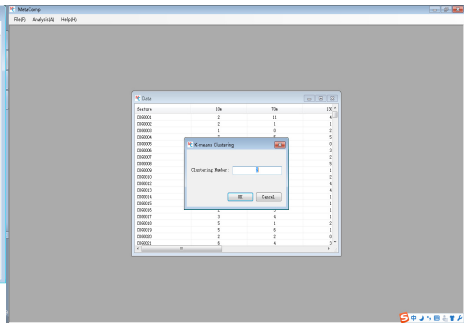

Figure 24:

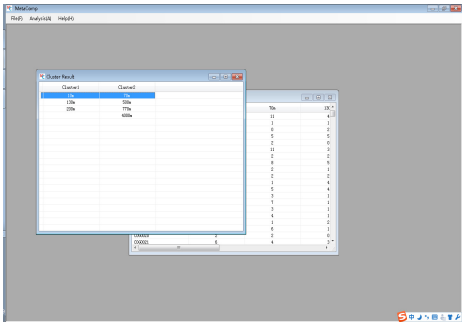

Figure 25: Result of k-means cluster.

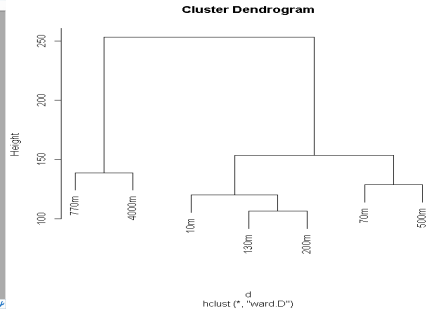

Figure 26: Result of hierarchical clustering.

6.2 Principal component analysis

Principal component analysis (PCA) is applied in two model: whole data analysis model and clustering analysis model. Whole data analysis model is the model we common used. Also, clustering analysis model can apply PCA within the clustering information.The example of clustering information can be download from PCA can be applied through the **Principal component analysis** dialog in **Analysis** menu.(Figure 27-29)

Linux commands line:

PCA(input_data,ShowsampleName="text" or "NA")

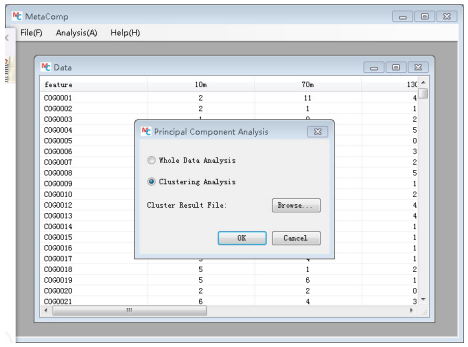

Figure 27:

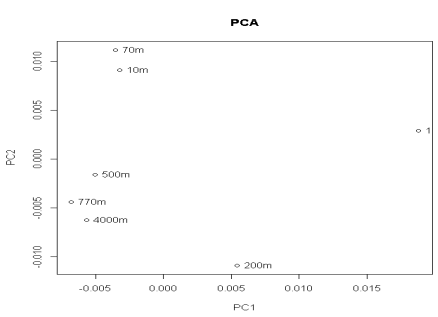

Figure 28: Result of whole data analysis model.

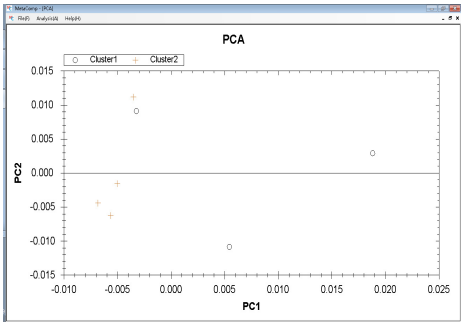

Figure 29: Result of clustering analysis model.

7 Hypothesis testing

7.1 Two samples test

To analyze a pair of samples, click on the **Two samples Statistic** dialog in **Analysis** menu. In this dialog, you can choose a favorable statistical test, p-value and data type. Moreover, you can choose the database you require if the feature in your profile is Pfam or COG database.(Figure 30-32)

Linux commands line:

result=twoSamplesComp(input_data)

plotTopVar(result)

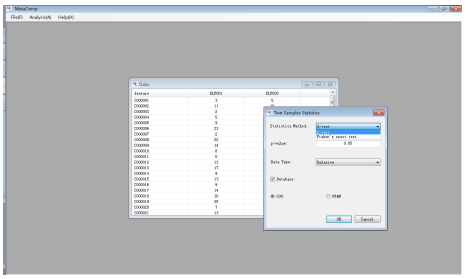

Figure 30:

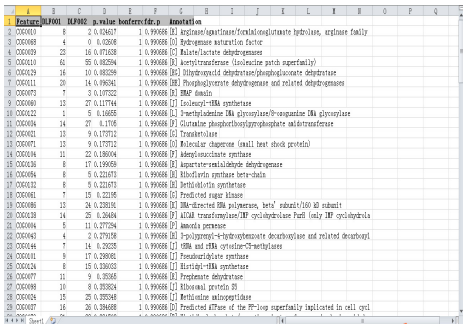

Figure 31: Analysis result(excel).

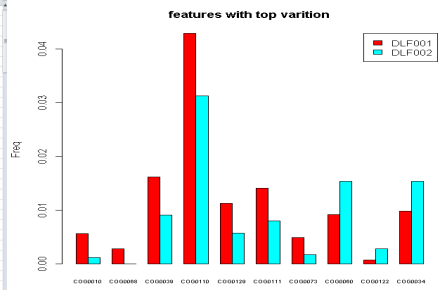

Figure 32: Analysis result(figure).

7.2 Multiple samples test

To analyze multiple samples, click on the **Multiple samples Statistic** dialog in **Analysis** menu. In this dialog, you can choose a favorable statistical test, p-value and data type. Just like Two samples test, you can choose the database you require if the feature in your profile is Pfam or COG database. Also you can select the most favorable visualizations you demanded.(Figure 33-35)

Linux commands line:

result=twoSamplesComp(input_data)

plotTopVar(result)

plotClust(result)

plotMDS(result, ShowsampleName = "legend","text","both"or "NA")

plotHeatMap(input_data, show_rownames = T or F, cluster_rows = T or F)

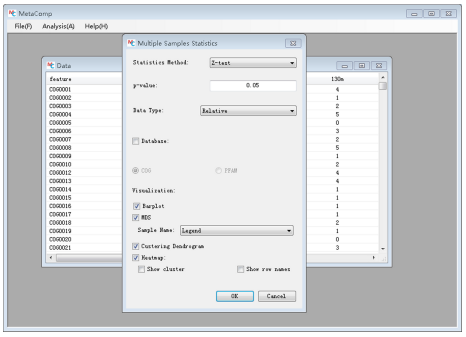

Figure 33:

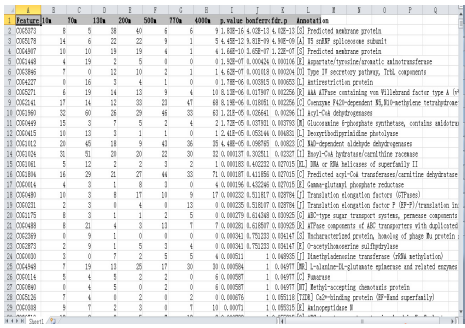

Figure 34: Analysis result(excel).

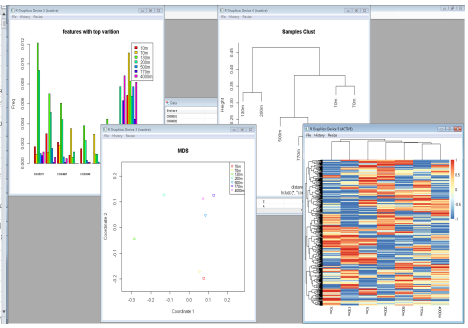

Figure 35: Analysis result(figure).

7.3 Two groups of samples test

To analyze two group of samples, click on the **Two group of samples Statistic** dialog in **Analysis** menu. In this dialog, you can choose a favorable statistical test, group number and p-value. Meanwhile, you can choose **I don't know** option in **Statistics Method** combo box while you don't know which method is the most suitable statistics test according to you input profile. Also, you can choose the database you require if the feature in your profile is Pfam or COG database. Also you can select the most favorable visualizations you demanded.(Figure 36-38)

Linux commands line:

result=twoSamplesComp(input_data,groupsep)(*groupsep* represents the sample number in first group.)

plotTopVar(result)

plotClust(result)

plotMDS(result, ShowsampleName = "legend","text","both"or "NA", isgroupData = T, groupsep)

plotHeatMap(input_data, show_rownames = T or F, cluster_rows = T or F)

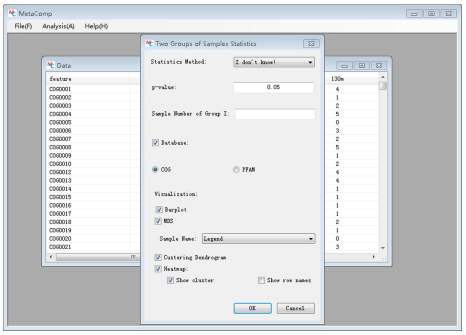

Figure 36:

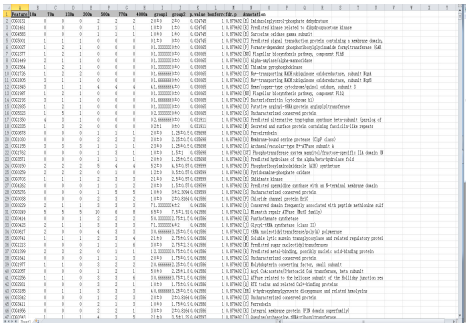

Figure 37: Analysis result(excel).

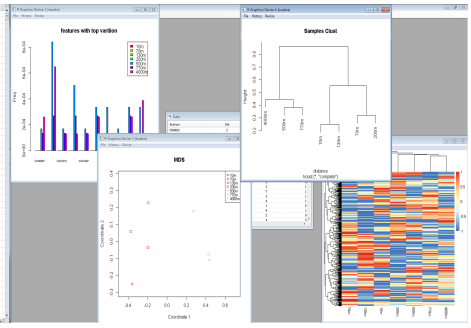

Figure 38: Analysis result(figure).

8 Environmental factors analysis

To operate environmental factors analysis, click on the **Environmental factors analysis** dialog in **Analysis** menu. In this dialog, you need to load the environmental factors information, input the p-value, choose whether you require to include the cross term of environmental factors and load Pfam or COG database while analysing. The example of environmental factors information can be download from.(Figure 39-40)

Linux commands line:

EnvironmentFactor(input_data,environment factor file pathway,Feature number)

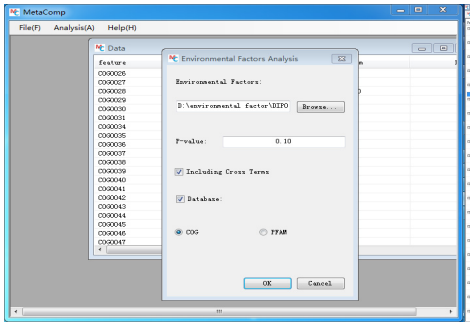

Figure 39: Analysis result(excel).

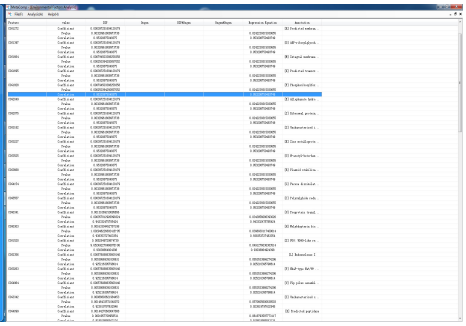

Figure 40: Analysis result(figure).

Bibliography

[1] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Ram Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, et al. The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):1, 2012.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[3] Jaina Mistry, Robert D Finn, Sean R Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41(12):e121–e121, 2013.

[4] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):1, 2014.

[5] Elizabeth M Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010(1):pdb–prot5368, 2010.

[6] Arthur Brady and Steven L Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673–676, 2009.

[7] Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, 2010.