# Script R12: Functional Analysis

**HANNIGAN GD, GRICE EA, ET AL.**

## Abstract

This protocol outlines the analysis used to plot KEGG functional annotation.

**Citation:** HANNIGAN GD, GRICE EA, ET AL. Script R12: Functional Analysis. **protocols.io** dx.doi.org/10.17504/protocols.io.ejibcke
**Published:** 10 Mar 2016

## Guidelines

sessionInfo()

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.4 (Yosemite)
## ## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5    formatR_1.2    tools_3.2.0    htmltools_0.2.6
## [5] yaml_2.1.13    stringi_0.4-1    rmarkdown_0.7    knitr_1.10.5
## [9] stringr_1.0.0    digest_0.6.8    evaluate_0.7
```

## Before start

Supplemental information available at:

https://figshare.com/articles/The_Human_Skin_dsDNA_Virome_Topographical_and_Temporal_Diversity_Genetic_Enrichment_and_Dynamic_Associations_with_the_Host_Microbiome/1281248

## Protocol

### Step 1.

Load the required libraries.

cmd COMMAND

```
library("plyr")
packageVersion("plyr")
```

```
library("gplots")

packageVersion("gplots")
```

EXPECTED RESULTS

```
## [1] '1.8.2'
```

```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
## lowess
```

```
## [1] '2.17.0'
```

**Step 2.**
Read in the metadata.

cmd COMMAND

```
metadata<-
read.delim("../../IntermediateOutput/Mapping_files/SkinMet_and_Virome_001_metadata.tsv")
metadata$NexteraXT_SampleID<-as.character(metadata$NexteraXT_SampleID)
metadata$NexteraXT_Virome_SampleID<-as.character(metadata$NexteraXT_Virome_SampleID)
metadata<-
metadata[,c("NexteraXT_SampleID", "NexteraXT_Virome_SampleID", "SubjectID", "TimePoint", "S
ite_Symbol")]
metadata<-subset(metadata, metadata$NexteraXT_SampleID != "NA")
metadata<-subset(metadata, metadata$NexteraXT_Virome_SampleID != "NA")
metadata<-subset(metadata, metadata$TimePoint != 1)
metadata<-subset(metadata, !(metadata$SubjectID %in% c(2,3,9,11)))
metadata<-subset(metadata, !(metadata$Site_Symbol %in% c("Neg", "Vf", "Ba", "Ph")))
```

**Step 3.**
Import the tab delimited file containing the output from HUMAnN. Read in the data for the whole metagenome and standardize sample ID's.

cmd COMMAND

```
mpm.s<-read.delim("../../IntermediateOutput/KEGG_humann/04b-hit-keg-mpm-cop-nul-nve-nve-
skinmet.txt")
colnames(mpm.s)<-
gsub(colnames(mpm.s),pattern="_R1_trimmed_subsampled_blastx",replacement="")
colnames(mpm.s)<-
gsub(colnames(mpm.s),pattern=".hit.keg.mpm.cop.nul.nve.nve",replacement="")
```

NOTES

**Geoffrey Hannigan** 16 Feb 2016
This file was generated using the shell script <>.

**Step 4.**
Read in data for the virome and standardize sample ID's.

cmd COMMAND

```
mpm.v<-read.delim("../../IntermediateOutput/KEGG_humann/04b-hit-keg-mpm-cop-nul-nve-nve-
virome.txt")
colnames(mpm.v)<-gsub(colnames(mpm.v),pattern="_R1_blastx",replacement="")
```

```
colnames(mpm.v)<-
gsub(colnames(mpm.v),pattern=".hit.keg.mpm.cop.nul.nve.nve",replacement="")
```

**Step 5.**

The data has information in the first 17 rows that we are not going to use. It needs to be formatted for downstream analysis. The last column is also a mock community sample, which we need to remove.

cmd COMMAND

```
#remove unnecessary information
mpm.s<-mpm.s[-c(1:17),-ncol(mpm.s)]
mpm.v<-mpm.v[-c(1:17),-ncol(mpm.v)]
```

**Step 6.**

Keep only paired samples.

cmd COMMAND

```
metadata<-subset(metadata, metadata$NexteraXT_SampleID %in% colnames(mpm.s))
metadata<-subset(metadata, metadata$NexteraXT_Virome_SampleID %in% colnames(mpm.v))
mpm.s<-mpm.s[, colnames(mpm.s) %in% c(metadata$NexteraXT_SampleID, "ID","NAME")]
mpm.v<-mpm.v[, colnames(mpm.v) %in% c(metadata$NexteraXT_Virome_SampleID, "ID","NAME")]
```

**Step 7.**

Merge whole metagenome and virome output.

cmd COMMAND

```
mpm<-merge(mpm.s, mpm.v, by=c("ID","NAME"))
```

**Step 8.**

Looking at the KEGG modules, we see they are very specific. We want to categorize them at a higher level.

**Step 9.**

Read in the level information.

cmd COMMAND

```
factor(mpm$NAME[1:5])

mpm_levelC <-
 read.delim("~/Club_Grice/reference/mpm_levelC.txt",header=F,colClasses=c("character", "cha
racter"))
colnames(mpm_levelC)=c('C','ID')
mpm_labs<-mpm[,c(1,2)]
mpm_levels<-merge(mpm_levelC,mpm_labs)
```

EXPECTED RESULTS

```
## [1] M00001: Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate
## [2] M00002: Glycolysis, core module involving three-carbon compounds
## [3] M00003: Gluconeogenesis, oxaloacetate => fructose-6P
## [4] M00004: Pentose phosphate pathway (Pentose phosphate cycle)
## [5] M00006: Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P
## 5 Levels: M00001: Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate ...
```

NOTES

**Geoffrey Hannigan** 16 Feb 2016

I originally downloaed this information from http://www.genome.jp/kegg-bin/ and clicked "Download htext". This gave me the file ko00002.keg which I parsed to get level C information.

**Step 10.**

Aggregate modules that are in the same higher level.

cmd COMMAND

```
mpm_levs<-merge(mpm_levels, mpm, by=c("NAME","ID"))
mpm_levs$ID<-NULL
```

```
mpm_levs$NAME<-NULL
mpm_levs[]<-lapply(mpm_levs,function(x) type.convert(as.character(x)))
mpm_agg<-aggregate(mpm_levs[,-1], by=list(mpm_levs$C), FUN=sum)
row.names(mpm_agg)<-mpm_agg$Group.1
mpm_agg$Group.1<-NULL
```

**Step 11.**

Transpose matrix so that the rows are samples and columns are KEGG module levels.

cmd COMMAND

```
mpmheat<-t(mpm_agg)
```

**Step 12.**

Now that we have categorized the KEGG modules, we want to compare them between the whole metagenome and the virome. We need to format our data frames so that the metagenome and virome samples are in the same order and then do pairwise comparisons. Finally, we want to calculate the log fold change between the metagenome and virome for each KEGG category that is significantly different.

**Step 13.**

Order the skin metagenome samples.

cmd COMMAND

```
mpmheat.s<-
merge(metadata[,c("NexteraXT_SampleID","SubjectID","TimePoint","Site_Symbol")],mpmheat, by.
y="row.names", by.x="NexteraXT_SampleID")
mpmheat.s<-mpmheat.s[order(mpmheat.s[c("SubjectID", "TimePoint","Site_Symbol"),]),]
colnames(mpmheat.s)[1]<-c("SampleID")
row.names(mpmheat.s)<-mpmheat.s$SampleID
mpmheat.s$SampleID<-NULL
mpmheat.s$TimePoint<-NULL
mpmheat.s$SubjectID<-NULL
mpmheat.s$Site_Symbol<-NULL
```

**Step 14.**

Order the virome samples.

cmd COMMAND

```
mpmheat.v<-
merge(metadata[,c("NexteraXT_Virome_SampleID","SubjectID","TimePoint","Site_Symbol")],mpmhe
at, by.y="row.names", by.x="NexteraXT_Virome_SampleID")
mpmheat.v<-mpmheat.v[order(mpmheat.v[c("SubjectID", "TimePoint","Site_Symbol"),]),]
colnames(mpmheat.v)[1]<-c("SampleID")
row.names(mpmheat.v)<-mpmheat.v$SampleID
mpmheat.v$SampleID<-NULL
mpmheat.v$TimePoint<-NULL
mpmheat.v$SubjectID<-NULL
mpmheat.v$Site_Symbol<-NULL
```

**Step 15.**

Do a wilcoxon to determine which modules are significantly different between skinmet and virome samples.

cmd COMMAND

```
wil.test.out<-vector()
for( i in 1:ncol(mpmheat.s) ) {
  wil<-
wilcox.test(x=as.numeric(mpmheat.s[,i]),y=as.numeric(mpmheat.v[,i]),p.adj=c("fdr"), paired=
TRUE)
  wil.test.out<-c(wil.test.out, wil$p.value)
}
wil.test.out.mat<-as.matrix(wil.test.out)
row.names(wil.test.out.mat)<-colnames(mpmheat.s)
```

```
keep<-subset(wil.test.out.mat, wil.test.out.mat[,1] < 0.05)
```

NOTES

**Geoffrey Hannigan** 16 Feb 2016

Note this will throw warnings-- some of the modules are almost all 0's which is not optimal when running a wilcoxon. However this does seem to pick up the major differences.

**Step 16.**

Add a small number to the relative abundances so that you can log transform them.

cmd COMMAND

```
c<-1e-25
met<-mpmheat.s+c
vir<-mpmheat.v+c
```

**Step 17.**

Calculate the log fold change.

cmd COMMAND

```
delta<-log2(met/vir)
delta.k<-delta[,colnames(delta) %in% row.names(keep)]
```

**Step 18.**

Finally, lets plot a heatmap to visualize the differences. Green indicates enrichment in the whole metagenome while red indicates enrichment in the virome.

cmd COMMAND

```
heatmap.2(as.matrix(t(delta.k)),notecol="black", density.info="none", trace="none",dendrogram="row",col=redgreen(1275),margins=c(2,30))
```

EXPECTED RESULTS