Mining EBI Metagenomics R-Analysis

Hubert Denis, European Bioinformatics Institute
Bioplatforms Australia
CSIRO, Australia

Canberra, ACTU July 2014

General information

The following standard icons are used in the hands-on exercises to help you locating:

Information information.

Instructions to perform.

General information/notes.

Warning - PLEASE take care and read carefully.

Questions.

An optional bonus section for those progressing rapidly.

An optional advanced section for those progressing very rapidly or to be used for future reference.

Resources used

Tools Used

EBI Metagenomics Portal
https://www.ebi.ac.uk/metagenomics

rRNASelector
http://www.ncbi.nlm.nih.gov/pubmed/21887657

RDP database
http://rdp.cme.msu.edu/taxomatic/main.spr

RStudio interface
https://www.rstudio.com/

HeatMap_in_R
http://sebastianraschka.com/Articles/heatmaps_in_r.htm

Sources of Data

HOT station, Central North Pacific Gyre, ALOHA
https://www.ebi.ac.uk/metagenomics/project/SRP000110

Tutorial objectives

This tutorial provides an introduction to understand EBI Metagenomics (EMG) resource results files and a guide for mining them. You will learn about the format of the EMG results files and how to mine them using R After completing this course, you should: Understand the format of the EMG result files Be able to extract information from the result files available on the EMG website using open source tools (The R statistical environment)

An introduction to the EMG result files

The EBI Metagenomics resource (EMG) generates taxonomy and functional analysis of metagenomic data sets. In addition of displaying the results on the EMG website, the results are available to download in order for the user to mine these according to their needs. What are the downloadable results files of the EMG resource? In addition to raw and filtered data, EMG provides 3 types of downloadable result files.

The first category is constituted of 5 different sequence files in FASTA format:

- Processed reads with predicted CDS

- Processed reads with InterPro matches

- Processed reads without InterPro match

- Predicted CDS

- Predicted CDS with InterPro matches

The second group comprise results of the functional annotation of the sequences in tab-separated (TSV) and comma-separated (CSV) format:

- InterPro matches (TSV)

- Complete GO annotation (CSV)

- GO slim annotation (CSV)

The last set corresponds to the taxonomic analysis output in

- Reads encoding 5S rRNA (FASTA)

- Reads encoding 16S rRNA (FASTA)

- Reads encoding 23S rRNA (FASTA)

- OTUs and taxonomic assignments (BIOM)

- Phylogenetic tree (Newick format)

- OTUs and taxonomic assignments (TSV)

The BIOM format is used to represent the number of 'Operational taxonomic unit' (OTU) observed in the sample. However this formart is not easily human-readable so these data are also available as a tab-separated file (OTUs and taxonomic assignments.tsv).

The Newick format is a computer-readable form to describe phylogenetic trees and is recognized by most viewers such as TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html), FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL (interactive Tree Of Life, http://itol.embl.de/) for example.

Downloading result files from the EMG website

All results files are available via the Download tab on the EMG website (www.ebi.ac.uk/metagenomics/) The purpose of this exercise is to illustrate how users can mine EMG result output files. We will produce a HeatMap graph to compare taxonomy annotations from 4 metagenomic datasets obtained by sequencing genomic DNA obtained from filtered Pacific Ocean samples collected at different depths.

The EMG pipeline performs taxonomy annotation from 16S rRNA-containing sequences identified using rRNASelector (http://www.ncbi.nlm.nih.gov/pubmed/21887657). Using Qiime, the 16S sequences are clustered by similarity (97%) and a representative sequence of each cluster, called Operational Taxonomy Unit (OTU), is annotated up to the kingdom, phylum, class, order or family level depending on the models present in the RDP database (http://rdp.cme.msu.edu/taxomatic/main.spr). These are these OTUs annotations that we will investigate in this practical.

I- Accessing the taxonomy annotation output file

The result of Qiime annotation performed by the EMG pipeline can be accessed on the EMG pipeline website.

Open the Metagenomics Portal homepage in a web browser. From the Project page (click the 'Project' tab), find the project 'HOT station, Central North Pacific Gyre, ALOHA' and click on the project name. You should now have a Project Overview page open describing the project, related publications, and links to the 4 sample that this project contains.

Click on the sample 'HOT Station ALOHA, 25m'. You should now been on the Sample Overview page. Click on the 'download' tab and then select the 'OTUs and taxonomic assignments (TSV)' link in the 'Taxonomic Analysis' group. In the menu window, select open it. You should be able to see that this is a tab-separated file with column 1 containing an OTU number and the name of the representative sequence, column 2 containing the predicted annotation and column 3 containing the e-value for this prediction.

Note that several OTUs have the same taxonomy annotations (for example line 3 and 8 for example).

Question 1: How could you explain that different OTUs could have the same taxonomy prediction ?

II- Investigation of the taxonomy annotations from the 'HOT Station ALOHA, 25m' sample set

To investigate the taxonomy annotations further, we will use the statistical environment R. We will use the graphical console RStudio to perform our analysis. There is no need of prior knowledge in R as all steps are decomposed step by step below as indicated by the step pictogram Type or paste the code lines in Rstudio as shown in figure 4. You can investigate further any function using the help tab as indicated in figure 4?

Together, we will answer this question: How could you obtain the unique OTUs and their occurrence in the taxonomy annotation ?

Open the R environment by starting Rstudio (Applications/Other/RStudio).

A screen similar to the following will start ADD A Figure 4 PNG

First we have to change the work space to the folder containing the datafiles (Hubert-Tutorial/Assembly). Type the following line in the console pane:

```
setwd("~/Desktop/HubertTutorial/Output_mining")
```

Load the tab-separated value HOT_Station_ALOHA,_25m_depth_otu output file.

```
Data25 <- read.delim("HOT_Station_ALOHA_25m_depth_otu.tsv", \
    header=FALSE)
```

To only extract the taxonomy information from this file, we only need data from the second column. This can be achieved simply in R:

```
Taxo25 = Data25[2]
```

The number of OTU annotations present in Taxo25 can be obtained by using the function 'dim':

```
dim(Taxo25)
168 1
```

As we saw earlier, more than one OTU could have the same annotation therefore some taxonomy annotations will be present multiple times in Taxo25. To obtain the unique annotations and their occurrence, R provides the function table which will compute a list of unique data and their occurrences:

```
UniqTaxo25 = table(Taxo25)
```

Visualize the table by typing in your console:

```
print(UniqTaxo25)
```

In order to transform this list in a table containing the unique OTUs and their occurrence, we will change the list as a table format called data frame in R. Put simply, a data frame is a table, in which each column contains measurements on one variable, and each row contains one case. We can re-cycle the variable UniqTaxo25 :

```
UniqTaxo25 = as.data.frame(UniqTaxo25)
```

The number of unique OTU annotations present in Taxo25 can be obtained by using the function 'dim':

```
dim(UniqTaxo25)
   36 2
```

You can also visualise the data frame by typing in your console:

```
print(UniqTaxo25)
```

This indicates that among the 168 OTU annotations, there is 36 unique annotations.

III- Obtaining the taxonomy annotations from the 'HOT Station ALOHA, 500m' sample set

Using the same step than described above, extract the taxonomy information for the sample 'HOT Station ALOHA 500m' creating the variables Data500, Taxo500 ad UniqTaxo500. You should obtain that among the 192 OTU annotations for the 500 depth file, there is 49 unique annotations.

IV- Generating a CSV file containing occurences annotations from both files

In order to compare these two annotation sets, we first need to merge them in order to obtain a table containing the unique OTUs and occurrence in both sets.

Again, R provides a very handy function for this:

```
TaxoComp = merge (UniqTaxo25, UniqTaxo500, by.x = "Taxo25", by.y= \
    "Taxo500", all = TRUE)
```

we can rename the columns for clarity:

```
 colnames(TaxoComp)=c("OTU", "HOT\_25", "HOT\_500")
```

If you visualise TaxoComp data frame, you should obtain: INSERT picture

Before drawing the heatmap, we need to replace the 'NA' by the value "0"

```
TaxoComp[is.na(TaxoComp)] = 0
```

Save this file as CSV (comma-separated) file:

```
write.csv(TaxoComp, file = "Comp25_500.csv", row.names = FALSE)
```

V- Generating a heatmap from taxonomy files for the 4 depth files

In interest of time, precomputed CSV files have been generated containing the taxonomy occurences for the 4 depth files. In addition, to increase heatmap clarity, the files at the class level of taxonomy have been collapsed. The resulting file has been saved as Taxo25_c.csv in folder HubertTutorial/Output_mining.

To collapse the UniqTaxo25 dataset at a chosen level, a python script is provided (Choose_level_RDP_count.py) which calculates the occurrence of the OTU after collapsing them at the chosen level. The script can be found in HubertTutorial/Output_mining.

DO NOT RUN THIS SCRIPT AS IT HAS ALREADY BEEN DONE FOR YOU

If you have a local installation of python, it can be run by typing:

```
1 python Choose_level_RDP_count.py UniqTaxo25 "level" where level can \
      be k (kindom), p (phylum), c (class), o (order) or f (family).
```

Taxo25_"level" was saved as text file:

```
1 write.table(UniqTaxo25_"level", file = "filename", row.names = \
      FALSE,col.names = FALSE)
```

To generate the heatmap, we will use a pre-written R script. To do so, go to File, OpenFile, navigate to the folder HubertTutorial/Output_mining and choose HeatMap_in_R.

Check that the 22nd line points to the correct UniqTaxo25_c.csv file. If you wish, you can modify the name of the heatmap image file at line 40. Then run the script by clicking on "Source" on the top right hand corner of the scripting window.

The heatmap will be saved into the HubertTutorial/Output_mining folder under the name indicated in line 40.

Comments: the HeatMap_in_R script is fully annotated and yours to re-use. Additional information about it can be read at http://sebastianraschka.com/Articles/heatmaps_in_r.html . The same approach could be used to generate an heatmap representation of the IPR or GO terms annotations.

Question 2: What can you observe by looking at the heatmap ? Comment on the (1) colour distribution (2) appearance/disappearance of OTUs and discuss the biological significance of your observations.