eudysbiome User Manual

Xiaoyuan Zhou, Christine Nardini
zhouxiaoyuan@picb.ac.cn

October 27, 2020

Introduction

Large amounts of data for metagenomics, especially the earliest studies on 16S ribosomal RNA gene, are produced by high-throughput screening methods. These are processed in the form of quantitative comparisons (between two microbiomes' conditions) of reads' counts. Reads' counts are interpreted as a taxon's `abundance` in a microbial community under given conditions, such as a medical treatments or environmental changes. The comparative analysis of such microbiomes with a baseline condition permits to identify a list of microbes (classified in species, genus or higher-order taxa) that are differential in abundance among the conditions. The taxonomic classification of 16S rRNA sequences is generally dependent on two strategies to assign sequences into populations: one is the phylotype-based method that assigns sequences to taxa based on the similarity to a reference database; the other is the operational taxonomic unit (OTU)-based method which does not rely on the association to known taxa, but consists of the definition of clusters of sequences sharing high similarity among them (typically a 97% identity threshold). These clusters are identified by an OTU_ID, and where possible associated to a species. The phylotype-based method limits the taxonomic classification of novel sequences from previously unknown taxa and is less sensitive to the sequencing errors[1]. The OTU-based method overcomes the limitations of phylotype-based method and permits the classification of OTUs down to the species level, as typically an OTU is thought of as representing a species, and for this the OTU-based method necessarily relies on a reference data set including species.

`eudysbiome` applies the OTU-based classification by mapping OTUs' unclassified *representative* sequences (produced via microbiome analysis pipelines such as emphMothur[2] or *QIIME*[3] during the generation of the OTUs) to a selection of classified representative sequences of OTUs, obtained from clustering the truncated *SILVA*[4] Small subunit (16S/18S, SSU) ribosomal RNA (rRNA) reference dataset at 97% similarity. This representative dataset is built within QIIME and included in our package. In this mapping, each entry in the representative dataset is associated to a taxonomic path containing the taxonomy rank designations from kingdom to species.

In order to translate the GI microbiome variations into potential clinical interpretation, it is helpful to assess whether the variation leads towards a microbiome that is more or less synergistic with the host, i.e. more or less pathogenic. To achieve this goal, in addition to the standard taxonomic annotation (`classification`, from now on) described above, this package annotates species as `harmful/harmless` based on their ability to contribute to mammals' host diseases (as indicated in literature) and hence based on their pathogenic potential (`annotation` from now on). Several assumptions are needed to achieve this goal, due to the quickly growing,

but still limited amount of information currently available on microbes in our GI. First, the number of unknown (uncultured, ambiguous, unclassified, etc.) species resulting from the classification process leads to the practical impossibility to assess the pathogenic potential directly on the annotated species. For this reason, the analysis is run at the genus level, annotated based on the annotated species. The rationale used throughout the annotation process relies on the assumption that larger and successful efforts have been devoted to the study of pathogens, and we have observed that classified species that are not explicitly (literature based) annotated as `harmful`, can be safely annotated as `harmless`. Genera are then further annotated as harmful, in a conservative fashion, if at least one species in the genus is pathogenic or less pathogenic otherwise. This outputs the reference table `harmGenera` offered in the package, and allows for direct annotation of classified genera. However, to achieve a more precise annotation, it is recommended to provide (or compute via our package `tableSpecies` function) the data frame genus-species, to control for the very stringent harmful annotation. In fact, if, in the data to be analyzed, species are provided in addition to the genus, the package checks whether the genera defined as harmful in the annotation table actually include the harmful species. In this case the genus is coherently annotated as harmful. However, if the harmful species are not present in the data to be analyzed, the genus provides in fact a harmless contribution, and it is annotated as such. Similarly, if the genus turns out to include none known species, we discard the one listed in the harmGenera table from the analysis while annotate the others as harmless.

Finally, the package statistically measures the `eubiotic` (harmless genera increase or harmful genera decrease) or `dysbiotic` (harmless genera decrease or harmful genera increase) relevance of a given treatment or environmental change in terms of its ability to modify the harmful/harmless frequencies.

- The package requires as inputs:
  - a FASTA-formatted 16S rRNA sequence file to be classified to the species level;
  - the list of microbe genera, for example, a list of differential genera identified by the comparative analysis to be annotated as `harmful/harmless`;
  - the microbial abundance variations, a simple difference of the differential genera abundance (defined as Δg) in the two conditions to be compared, as defined above in Introduction;
  - a table qualifying the genera as `harmful`/`harmless`, as defined by literature. Such a table, manually curated, is included in this package, but is by no means exhaustive: continuous advances in microbiology make this input incomplete and flexible; we encourage users to share expansions of this table.
- The package outputs:
  - Species-classification results, a `*.taxonomy` file which contains a taxonomic path for each 16S rRNA sequence and a `*.tax.summary` file which contains a taxonomic outline indicating the number of sequences that were found at each level (kingdom to species);
  - optional, a two-column Genus-Species data frame extracted from the assigned taxonomic paths in `*.taxonomy` file, which includes only the differential genera (as *inputs*);

Comparison	EI	DI	*Row Total*
C1	a	b	$a+b$
C2	c	d	$c+d$
Column Total	$a+c$	$b+d$	$a+b+c+d(=n)$

Table 1: Contingency Table

- a graphical output of the genus abundance difference-Δg across the tested conditions (y-axis) and their harmful/harmless nature (negative/positive x-axis);
- a contingency table showing as frequencies the cumulated contributions to an eubiotic/dystbiotic microbiome (see Table 1, columns, namely `EI` and `DI`) under different conditions (comparisons between a condition and a reference, listed in rows, namely `C1` and `C2`). The eubiotic impact (EI) is quantified by the $|\Delta g|$ cumulation of increasing harmless genera and decreasing harmful genera, while the dysbiotic impact (DI) is quantified by the reverse, i.e. $|\Delta g|$ accumulation of decreasing harmless genera and increasing harmful genera;
- the results (probability) of testing the null hypothesis that there is no difference in the proportions of frequencies of `EI` between `C1` and `C2` using Fisher exact test (two sided) or Chi-squared test[5], computed as the probability that the proportion of frequencies in `EI` under `C1` ($\frac{a}{a+b}$) is different from that in `DI` under `C2` ($\frac{c}{c+d}$). The results of the one-sided Fisher's exact test[5] assess whether `C1` is more likely to be associated to a eubiotic microbiome than `C2`, and is computed as the probability that the proportion of EI under `C1` is higher than `C2`.

Methods

1 Representative Sequences

To achieve the Species-level classification, we recommend classifying the unknown 16S rRNA sequences to a well-curated representative dataset of 16S rRNA reference sequences, such as Greengene and SILVA representative sets (as recommended by *Mothur* with very stringent 99% similarity and *QIIME* with 97% similarity). We here use the SILVA representative set created by clustering at 97% sequence identity, to guarantee a fast Species-level classification and also require less computational resources when assigning sequences to a reference dataset, both requirements are crucial to allow automation of this classification step, as we offer in this package. The representative dataset "Silva_119_rep_set97.fna" is downloaded in latest version SILVA119 provided by QIIME team from (https://www.arbsilva.de/no_cache/download/archive/qiime). A taxonomic mapping file "Silva_119_rep_set97_taxonomy.txt", mapping each entry in the representative dataset to a taxonomy rank designation, was also downloaded and prepared into the input format to *Mothur*, which is included in our package for the usage of further sequence classification.

2 Taxonomy Assignment

The assignment requires a FASTA-formatted input of unclassified sequence, a representative sequence file and a taxonomic mapping file for the representative sequences(Section 1). Given a set of unclassified 16S rRNA sequences, e.g. a set of OTU representative sequences, we assign the taxonomic paths to these sequences by calling the `classify.seqs` command in *Mothur* (http://www.mothur.org/). Of the two alternative methods (`Wang` and `k-Nearest Neighbor (knn)`) provided in the `classify.seqs` command for the taxonomic assignment, we use ***`Wang`***'s, implemented by the RDP classifier. This method queries both the unclassified and reference sequences k-mer by k-mer (subsequences of length k) and assigns the unclassified sequences to the appropriate taxa based on the highest matching probability. To calculate the confidence of the assignments, bootstrapping by random replacement of 1/8 (k = 8) of the k-mers in the unclassified sequence is used.

```
> library("eudysbiome")
> input.fasta = "Unclassified.fasta"
> # using the extracted fasta and taxonomy as template
> assignTax(fasta = input.fasta,ksize = 8, iters = 100,
+                         cutoff = 80, processors=1, dir.out = "assignTax_out")
```

The parameters, `k-size` (length k), `iterates` (bootstrap iterations) and `processors` (number of central processing units) are used as defaults in *Mothur*. We set a cutoff of bootstrap confidence score to 80, which means a minimum 80% sequences were assigned to the same taxonomy, a higher value gives a more strict and accurate taxonomy assignment. A `*.taxonomy` file and a `*.tax.summary` file of the classification results are outputd into the `assginTax` directory.

3 Genus-Species Table Construction

To identify only the species under certain genera from the `*.taxonomy` file, the `tableSpecies` function constructs a two-column Genus-Species data frame, where one column refers to the provided genera while the other refers to the species included in these genera.

```
> genera = c("Lactobacillus","Bacteroides")
> #species = tableSpecies(tax.file = "*.taxonomy", microbe = genera)
```

4 Microbe Annotation

A differential genera list (input) can be annotated as `harmless` or `harmful` by the function `microAnnotate` based on our manually curated table named `harmGenera` in this package. The table lists the harmful genera based on the pathogenic or opportunistic pathogenic species included in the genera, using a stringent approach: one armful species is sufficient to define the genus harmful (this is what indeed matters to the eubiotic/dysbiotic trend). Although a genus list is acceptable and can be processed with this genera annotation table, we recommend inputting for the data to be analyzed the Genus-Species data frame, as in the `diffGenera` table below to gain a more accurate annotation. In fact, if the species abundances are known, it is possible to discard a genus in case none of the taxonomically annotated species is present in the dataset (only unknown ones), or mark as harmless a genus that would be harmful by annotation table, in case the harmful species is not present in the dataset under study. For example, genus1 will be annotated as harmful if any of the three species (1, 2 and 3) under this genus is annotated as `harmful`, otherwise, genus1 will be annotated as `harmless`. For the data lacking of Species-level classifications, we suggest to do the classification and construct such a Genus-Species table for better annotation by functions described above (Section 3).

```
> library("eudysbiome")
> data(diffGenera)
> head(diffGenera)
```

```
   Genus  Species
1 genus1 species1
2 genus1 species2
3 genus1 species3
4 genus2 species1
5 genus2 species2
6 genus3 species1
```

```
> data(harmGenera)
> annotation = microAnnotate(diffGenera, annotated.micro = harmGenera)
```

5 Cartesian Plane Plot

The function `Cartesian` accepts either a data frame or a numeric matrix of Δg, whose rows represent differential genera and columns represent condition comparisons, these are the argument to produce the cartesian plane (4 quadrants (see details below and in Figure 1 below).

The Δgs are log-2 converted and redundantly represented by the height on the y-axis and the dots diameter. Because of its definition, the increase of harmless (1st cartesian quadrant) and/or the decrease of harmful (3rd cartesian quadrant) define microbiome variation that are eubiotic (beneficial) and highlighted by a blue box, and the decrease of harmless (2nd quadrant) and/or the increase of harmful (4th quadrant) as dysbiotic (non-beneficial) and highlighted by a yellow box. The unknown genera are removed from the plot.

For example below, a data frame `data` is constructed from the `microDiff` dataset with Δg of ten differential genera among comparisons `A vs C`, `B vs C` and `D vs C`, where `A`, `B` and `D` are three conditions and `C` is a control. The genera are annotated as `harmless`, `harmful` or `unknown` in `micro.ano` based on the output by the `microAnnotate` function, and comparisons are defined as `A-C` (A vs C), `B-C` (B vs C), and `D-C` (D vs C) in `comp.ano` and indicated by the column names of the input data if no other `comp.anno` is specified. Eubiotic changes associated to conditions `A, B, D` compared to control `C` are plotted in the up-utmost right and bottom-utmost left quadrants (increase of harmless and decrease of harmful genera) and dysbiotic variations are plotted on the bottom-utmost right and up-utmost left quadrants (increase of harmful and decrease of harmless genera) in Figure 1.

```
> data(microDiff)
> microDiff

$data
        A vs C B vs C D vs C
genus1      99    551      0
genus2       0     57   -290
genus3     441   -303     41
genus4     300  -1624  -1138
genus5     -77    200  -1240
genus6      15      0   -190
genus7       0      5      0
genus8    -106      0    206
genus9    -145     10      0
genus10   1277     90    -58

$micro.anno
 [1] "harmless" "unknown"  "harmless" "harmful"  "unknown"  "harmful"
 [7] "harmless" "harmful"  "harmful"  "harmless"

$comp.anno
[1] "A-C" "B-C" "D-C"

> attach(microDiff)
> par(mar = c(5,4.1,5.1,5))
> Cartesian(data ,micro.anno = micro.anno,comp.anno= comp.anno,
+                 unknown=TRUE,point.col = c("blue","purple","orange"))
```

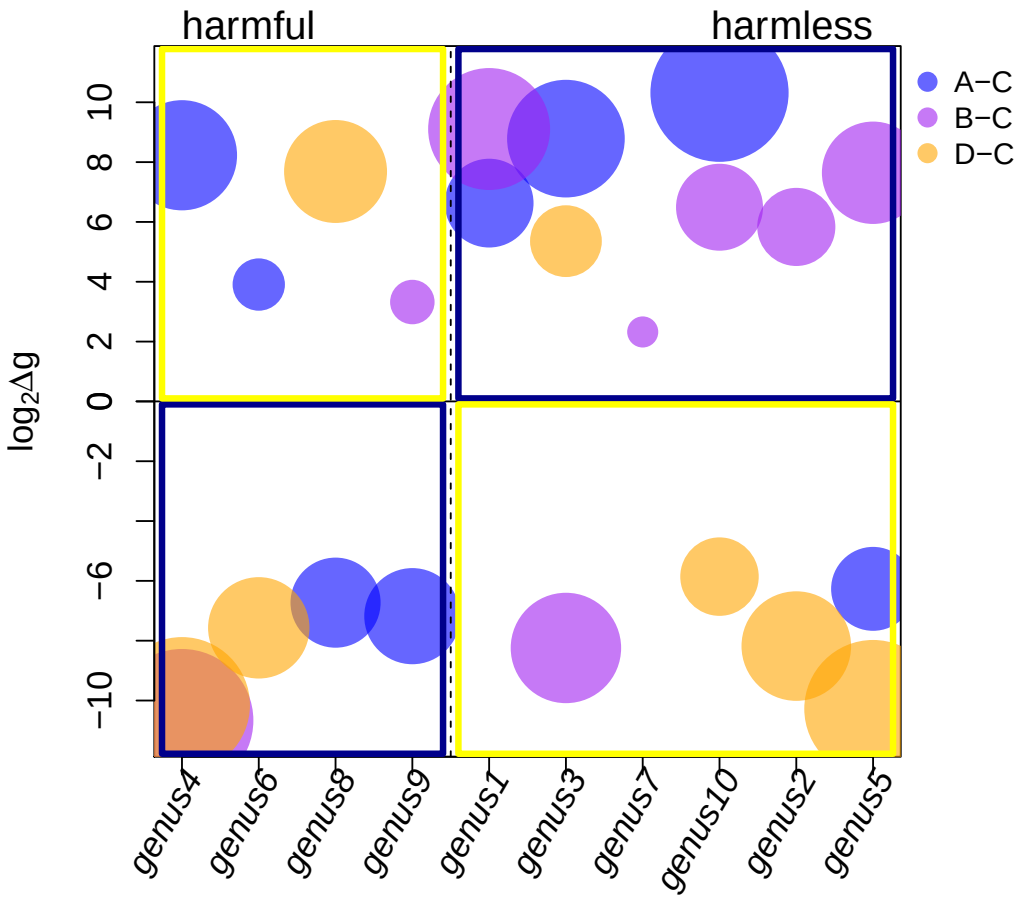

Figure 1: Cartesian plane of the harmful/harmless annotated genera (on the x-axis) and their abundance variations among the condition comparisons (log2 (Δg), y-axis). The eubiotic microbiome impact is highlighted by a dark blue box while the dysbiotic one is highlighted by a yellow box.

Condition	Eubiotic Impact	Dysbiotic Impact
A-C	2068	315
B-C	2270	313
D-C	1369	264

Table 2: Condition-impact contingency table of microbial frequencies

6 Contingency Table Construction

This function computes the frequencies of the contingency table as the cumulated $|\Delta g|$ classified by each couple formed by a condition and an impact (eubiotic/dysbiotic, see Table 1). This outputs the significance of the association (contingency) between conditions and impacts by `contingencyTest`. For example, the benefits of conditions `A, B, D` are measured by the increase Δg of harmless genera and the decrease Δg of harmful genera in the comparisons to `C`, while the non-beneficial impact is evaluated in reverse by the decrease Δg of harmless genera and the increase Δg of harmful genera. Absolute values of Δg are cumulated as frequencies and used into the contingency table (Table 2).

```
> microCount = contingencyCount(data ,micro.anno = micro.anno,
+                               comp.anno= comp.anno)
```

7 Contingency test for count data

To elaborate the significance of the association between conditions and eubiotic/dysbiotic impacts, Chi-squared test and Fisher's exact test (one- and two- sided) are performed on the frequencies from `contingencyCount` for testing the null hypothesis that conditions are equally likely to lead to a more eubiotic microbiome when compared to the control while the alternative hypothesis is that this probability is not equal or one condition is more likely to be associated to an eubiotic microbiomes than the other (only with Fisher test, one-sided). Taking Table 2 as an example, we hypothesize that the proportion of eubiotic frequencies are different (Chi-squared and two-sided Fisher test) between condition comparisons `A-C`, `B-C` and `D-C` or even higher (one-sided Fisher test) in one comparison than the other, and we want to test whether this difference is negligible or refers to a significant association between the condition and the (GI) microbiome composition modification. Both Fisher and Chi-squared tests are performed by the `contingencyTest` function and significance values are output in tables.

```
> microTest = contingencyTest(microCount,alternative ="greater")
> microTest["Chisq.p"]

$Chisq.p
        Chisq.Pvalue
A-C:B-C  0.261245444
A-C:D-C  0.010267809
B-C:D-C  0.000233087

> microTest["Fisher.p"]
```

```
$Fisher.p
        Fisher.Pvalue_greater
A-C:B-C          0.8866246202
A-C:D-C          0.0052786178
B-C:D-C          0.0001289438
```

References

[1] Patrick D. Schloss, et al. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Applied and Environmental Microbiology* 2011; 77(10): p. 3219-3226

[2] Patrick D. Schloss, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology* 2009; 75(23): 7537-7541.

[3] J Gregory Caporaso, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 2010; 7(5):335-336.

[4] Quast C.,et al. The SILVA ribosomal RNA gene database project:improved data processing and web-based tools. *Nucleic Acids Research* 2013; 41(D1):D590-D596

[5] Rice, John A., Mathematical statistics and data analysis, Belmont, CA, Thomson/Brooks/Cole, Duxbury advanced series, 3rd, 2007.

Session Information

The session information records the versions of all the packages used in the generation of the present document.

- R version 4.0.3 (2020-10-10), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Ubuntu 18.04.5 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.12-bioc/R/lib/libRblas.so`
- LAPACK: `/home/biocbuild/bbs-3.12-bioc/R/lib/libRlapack.so`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: eudysbiome 1.20.0
- Loaded via a namespace (and not attached): BiocGenerics 0.36.0, BiocParallel 1.24.0, Biostrings 2.58.0, GenomeInfoDb 1.26.0, GenomeInfoDbData 1.2.4, GenomicRanges 1.42.0, IRanges 2.24.0, R.methodsS3 1.8.1, R.oo 1.24.0, R.utils 2.10.1, RCurl 1.98-1.2, Rcpp 1.0.5, Rsamtools 2.6.0, S4Vectors 0.28.0, XVector 0.30.0, bitops 1.0-6, compiler 4.0.3, crayon 1.3.4, parallel 4.0.3, plyr 1.8.6, stats4 4.0.3, tools 4.0.3, zlibbioc 1.36.0