eudysbiome User Manual

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

October 14, 2015

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. Overall, the comparative analysis of such microbiomes with a baseline condition permits to identify a list of microbes (classified in species, genus or higher taxa) that are differential among the conditions in `differential abundance`.

`eudysbiome` is a package that permits to annotate the differential genera of a (gut-intestinal, GI) microbiome as `harmful/harmless` based on their ability to contribute to mammals' host diseases (as indicated in literature) or `unknown` based on their ambiguous genus classification. Further, the package statistically measures the `eubiotic` (harmless genera increase or harmful genera decrease) or `dysbiotic` (harmless genera decrease or harmful genera increase) impact of a given treatment or environmental change on the microbiome in comparison to the microbiome of the reference condition.

The package requires as inputs:

- the microbial abundance variations, a simple difference of the differential genera abundance (Δg) in the two conditions to be compared, as defined above;
- a table qualifying the differential genera as harmful/harmless/unknown, 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:

- a graphical output of the genus abundance difference-Δg across the tested conditions (y-axis) and their harmful/harmless nature (negative/positive x-axis). Since a number of microbes have unknown genus classifications as a result of unknown genus annotations, the x-axis is broken into a positive (harmless), negative (harmful) and "neutral" (unknown) segments (pseudo-cartesian plane);
- the contingency table showing as frequencies the cumulated contributions to an eubiotic/dystbiotic microbiome impacts (see Table 1, columns, namely EI and DI) under

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

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 Chi-squared test[1], 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[1] 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`.

1 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 and the harmful species included in the genera. Although a genus list is acceptable and can be processed by the package, we recommend inputting a Genus-Species data frame, as in the `diffGenera` table below, which represents the differential genera and the included corresponding species to gain a more accurate annotation. 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`.

```
> 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)
```

2 Pseudo-Cartesian Plane Plot

The function `pseudoCartesian` 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 pseudo-cartesian plane (6 sub-areas -pseudo quadrants- instead of 4 quadrants where the 2 central are called neutral areas (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 pseudo-cartesian quadrant) and/or the decrease of harmful (3rd pseudo-cartesian quadrant) define microbiome variation that are eubiotic (beneficial) and highlighted by a green shade, and the decrease of harmless (2nd pseudo-quadrant) and/or the increase of harmful (4th pseudo-quadrant) as dysbiotic (non-beneficial) and highlighted by a red shade. The unknown genera can be optionally shown in the two central neutral areas.

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"
```

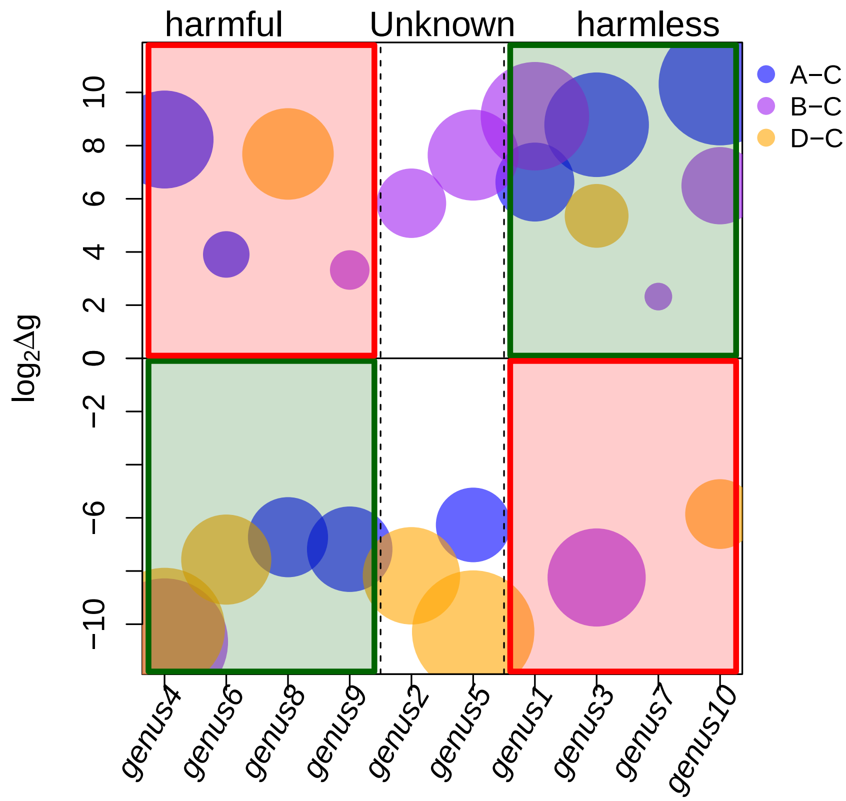

Figure 1: Pseudo-cartesian plane of the harmful/unknown/harmless annotated genera (on the x-axis) and their abundance variations among the condition comparisons (log2 (Δg), y-axis). The eubiotic microbiome impact is highlighted by a green shade while the dysbiotic one is highlighted by a red shade.

```
> attach(microDiff)
> par(mar = c(6,5.1,4.1,6))
> pseudoCartesian(data ,micro.anno = micro.anno,comp.anno= comp.anno,
+                 unknown=TRUE,point.col = c("blue","purple","orange"))
```

3 Contingency Table Construction

This function computes the frequencies of the contingency table as the cumulated |Δg| 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`,

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

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)
```

4 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] 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 3.2.2 (2015-08-14), `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`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: eudysbiome 1.0.0
- Loaded via a namespace (and not attached): Rcpp 0.12.1, plyr 1.8.3, tools 3.2.2