Taxonomic biomarker discovery with LEfSe

Curtis Huttenhower

Abstract

In this protocol we show how to perform the taxonomic biomarker discovery operation using LEfSe.

Citation: Curtis Huttenhower Taxonomic biomarker discovery with LEfSe. **protocols.io** dx.doi.org/10.17504/protocols.io.d9699d
Published: 14 Jan 2016

Guidelines

We first need to specify the conditions (or classes) used for the biomarker discovery. Examples could be host disease states from which a gut microbiome was sampled, or environmental conditions (e.g. pH, environmental contaminant) from which a microbial community was sampled. In this tutorial, we will use body sites (contrasting the tongue dorsum microbiome with the buccal mucosa microbiome).

The commands reported in this protocol can be retrieved as a bash script.

Before start

REQUIREMENTS: LEfSe installed (and in the system path), and the matplotlib python library installed. Alternatively, the users are welcome to use the Galaxy interface for LEfSe. LEfSe can be downloaded using using Mercurial: `hg clone ssh://hg@bitbucket.org/nsegata/lefse` or using the direct links to the zip, gz, or bz2 archives. Notice that LEfSe has some additional requirements.

Protocol

Step 1.

Before beginning, we need to convert the sample names into consistent class names. This can easily be done by manually editing the `output/merged_abundance_table.txt` generated in the previous protocols or using the following "sed" based Unix command:

cmd COMMAND

```
$ sed 's/\([A-Z][A-
Z]\)_\w*/\1/g' output/merged_abundance_table.txt > tmp/merged_abundance_table.4lefse.txt
```

Renaming sample names into consistent class names.

NOTES

Bahar Sayoldin 07 Dec 2015
For this specific image we increased the LDA threshold to 4.7 in order to display a more compact image with fewer biomarkers.
Bahar Sayoldin 07 Dec 2015

Notice that the cladograms produced by LEfSe can be less graphically appealing and detailed than those built using GraPhlAn. Using the `tmp/merged_abundance_table.lefse.out` file and, a combination of scripting or manual editing, it is possible to obtain an improved graphical output for LEfSe.

Step 2.

The first LEfSe step consists of formatting the input table, making sure the class information is in the first row and scaling the values in [0,1M] which is useful for numerical computational reasons.

cmd COMMAND

```
$ format_input.py tmp/merged_abundance_table.4lefse.txt tmp/merged_abundance_table.lefse -
c 1 -o 1000000
```

Step 3.

Now, the LEfSe biomarker discovery tool can be used with default statistical options. Here we change one default parameter to increase the threshold on the LDA effect size from 2 (default) to 4.

cmd COMMAND

```
$ run_lefse.py tmp/merged_abundance_table.lefse tmp/merged_abundance_table.lefse.out -l 4
```

Step 4.

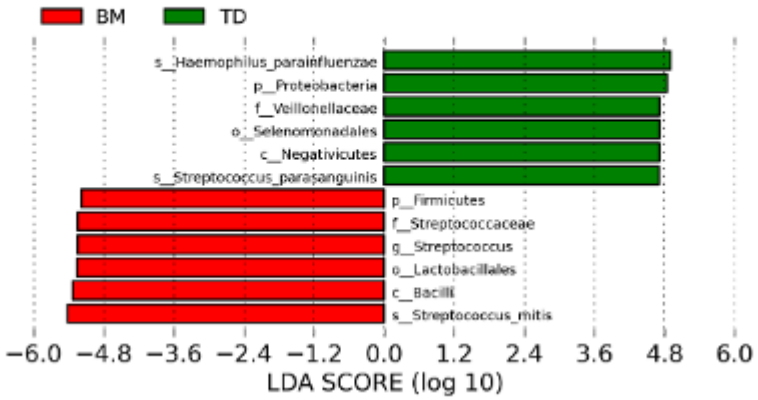
The results of the operation can now be displayed plotting the resulting list of biomarkers with corresponding effect sizes.

cmd COMMAND

```
$ plot_res.py --
dpi 300 tmp/merged_abundance_table.lefse.out output_images/lefse_biomarkers.png
```

Step 5.

The resulting image (`output_images/lefse_biomarkers.png`) is shown below:



NOTES

Bahar Sayoldin 07 Dec 2015

For this specific image we increased the LDA threshold to 4.7 in order to display a more compact image with fewer biomarkers.

Step 6.

Another complementary visualization focuses on showing the biomarkers on the underlying taxonomic tree.

cmd COMMAND

```
$ plot_cladogram.py --dpi 300 --
format png tmp/merged_abundance_table.lefse.out output_images/lefse_biomarkers_cladogram.pn
g
```

Step 7.
The default plotting output is shown here:

NOTES

Bahar Sayoldin 07 Dec 2015

Notice that the cladograms produced by LEfSe can be less graphically appealing and detailed than those built using GraPhlAn. Using the `tmp/merged_abundance_table.lefse.out` file and, a combination of scripting or manual editing, it is possible to obtain an improved graphical output for LEfSe.

Step 8.
Single features can also be plotted as barplots with `plot_features.py`. For example, to plot the distribution of abundances for Firmicutes the command is:

cmd COMMAND

```
$ plot_features.py -f one --
feature_name "k__Bacteria.p__Firmicutes" tmp/merged_abundance_table.lefse tmp/merged_abunda
nce_table.lefse.out output/Firmicutes.png
```

Step 9.
All features (or all biomarkers) can also be exported in one compressed archive:

cmd COMMAND

```
$ plot_features.py -f diff --
archive zip tmp/merged_abundance_table.lefse tmp/merged_abundance_table.lefse.out biomarker
s.zip
```

NOTES

Bahar Sayoldin 07 Dec 2015

Notice that with `-f diff` we are exporting only features that are biomarkers according to LEfSe, but with can export all features with `-f all`.

Step 10.
Here are the resulting figures for Firmicutes and for Veillonellaceae (extracted from biomarkers.zip).