MetaPhlAn profiling of 20 HMP samples

Curtis Huttenhower

Abstract

In this protocol we will show how to taxonomically profile 20 metagenomic shotgun samples Human Microbiome Project (HMP). Specifically, we will look at 10 samples from the buccal mucosa and 10 from the tongue dorsum. Please note these 20 samples are not necessarily an unbiased selection of microbiomes from the over 200 buccal mucosa and tongue dorsum samples generated by the HMP.

REQUIREMENTS: BowTie2 installed and in the system path, Mercurial, basic unix commands (wget, tar).

The size of the 20 samples (60GBs) can make their download quite slow. For this reason we are also providing the overall output of this protocol (few KBs) available for download so that one can start the tutorial from MetaPhlAn output merge and visualization if needed.

Citation: Curtis Huttenhower MetaPhlAn profiling of 20 HMP samples. **protocols.io**
dx.doi.org/10.17504/protocols.io.d9e93d
Published: 14 Jan 2016

Guidelines

MetaPhlAn2 is now ready at:

http://segatalab.cibio.unitn.it/tools/metaphlan2/

https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2

https://bitbucket.org/biobakery/metaphlan2

This tutorial is also a step-by-step description of the metagenomic pipeline we used in our review paper about computational meta'omics (specifically Figure 4). The references for each tool are reported at the end of this page. If you find these tools useful, we kindly ask you to use these references for citing. Additionally, links to the corresponding user groups are provided should you have any comments or suggestions you would like to share with us.

We assume you have a basic knowledge of shotgun metagenomics and of Unix-based operating systems (although you should be able to run the pipeline on MacOS and Windows, some commands may require modification). Also, Python >2.6 (but not Python 3) is required to be installed (Python 2.7 for Step 5).

Please do not hesitate to contact us if you have any comments, suggestions, or clarification requests regarding the tutorial or if you would like to contribute to this resource. We are also **looking for**

motivated students and postdocs for developing new metagenomic tools, check here if you are interested.

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

Before start

We assume you have a basic knowledge of shotgun metagenomics and of Unix-based operating systems (although you should be able to run the pipeline on MacOS and Windows, some commands may require modification). Also, Python >2.6 (but not Python 3) is required to be installed (Python 2.7 for Step 5).

Protocol

Step 1.

The first operation consists in obtaining the lastest version of MetaPhlAn downloadable as zip, gz, or bz2 compressed archives. In a Unix environment, you can obtain and uncompress it from the command line

SOFTWARE PACKAGE (Unix)

MetaPhlAn, 1.7.8

Curtis Huttenhower

cmd COMMAND

```
$ wget https://bitbucket.org/nsegata/metaphlan/get/default.tar.bz2
$ tar xjvf default.tar.bz2
$ mv *-metaphlan-* metaphlan
```

Step 2.

Alternatively, you can use Mercurial to obtain the package from the Bitbucket repository and keep it updated in the future:

cmd COMMAND

```
$ hg clone ssh://hg@bitbucket.org/nsegata/metaphlan
```

Step 3.

We then navigate to the MetaPhlAn folder and create a subfolder for storing the 20 metagenomes:

cmd COMMAND

```
$ cd metaphlan
$ mkdir input
```

Step 4.

Now, we can download the 20 samples to profile (additional information about the samples is available at the HMP DACC):

cmd COMMAND

```
$ buccal_mucosa_samples="SRS013506 SRS015374 SRS015646 SRS017687 SRS019221 SRS019329 SRS020
336 SRS022145 SRS022532 SRS045049"
$ for s in ${buccal_mucosa_samples}
$ do
$     wget http://downloads.hmpdacc.org/data/Illumina/buccal_mucosa/${s}.tar.bz2 -O input/$
{s}.tar.bz2
$ done
$ tongue_dorsum_samples="SRS011243 SRS013234 SRS014888 SRS015941 SRS016086 SRS016342 SRS017
713 SRS019219 SRS019327 SRS043663"
```

```
$ for s in ${tongue_dorsum_samples}
$ do
$     wget http://downloads.hmpdacc.org/data/Illumina/tongue_dorsum/${s}.tar.bz2 -O input/$
{s}.tar.bz2
$ done
```

NOTES

Bahar Sayoldin 02 Dec 2015

As discussed above, this operation will likely require several hours.

Step 5.

Next, let's create a folder for storing the MetaPhlAn output.

COMMAND

```
$ mkdir profiled_samples
```

Step 6.

We can now profile the 10 buccal mucosa samples using MetaPhlAn. We are running MetaPhlAn using the BowTie2 engine (this step requires BowTie2 to be installed and in the system path).

COMMAND

```
$ BM_samples="SRS013506 SRS015374 SRS015646 SRS017687 SRS019221 SRS019329 SRS020336 SRS0221
45 SRS022532 SRS045049"
$ for s in ${BM_samples}
$ do
$     tar xjf input/${s}.tar.bz2 --to-stdout | ./metaphlan.py --bowtie2db bowtie2db/mpa --
bt2_ps very-sensitive --input_type multifastq --
bowtie2out BM_${s}.bt2out > profiled_samples/BM_${s}.txt
$ done
```

NOTES

Bahar Sayoldin 02 Dec 2015

Notice we are piping the fastq reads directly from the compressed archive to MetaPhlAn. When piping MetaPhlAn's internal parallelization option (--nproc option) can not be used. This is available when the input is a an uncompressed file. Notice, however, that if your machine has multiple CPUs you can run multiple MetaPhlAn profiling operations in parallel.

Step 7.

Please refer to the MetaPhlAn help (see command below) or to the MetaPhlAn wiki for specific information about other strategies and additional MetaPhlAn options.

COMMAND

```
$  ./metaphlan.py -h
```

Accessing MetaPhlAn help page

Step 8.

Similarly, we can apply MetaPhlAn to the 10 tongue dorsum samples.

COMMAND

```
$ TD_samples="SRS011243 SRS013234 SRS014888 SRS015941 SRS016086 SRS016342 SRS017713 SRS0192
19 SRS019327 SRS043663"
$ for s in ${TD_samples}
$ do
$     tar xjf input/${s}.tar.bz2 --to-stdout | ./metaphlan.py --bowtie2db bowtie2db/mpa --
bt2_ps very-sensitive --input_type multifastq --
bowtie2out TD_${s}.bt2out > profiled_samples/TD_${s}.txt
$ done
```

Step 9.

The "profiled_samples" folder now contains 20 profiled metagenomes. Here is an example of first few lines of the BM_SRS013506 sample output.

COMMAND

```
$ cat profiled_samples/BM_SRS013506.txt
$ k__Bacteria     100.0
$ k__Bacteria|p__Firmicutes       80.97874
$ k__Bacteria|p__Proteobacteria    17.17081
$ k__Bacteria|p__Fusobacteria      0.34233
$ k__Bacteria|p__Bacteroidetes     0.17203
$ k__Bacteria|p__Firmicutes|c__Bacilli    80.62406
[truncated output]
```