MUSiCC User Manual

MUSiCC is a toolkit for correcting biases in gene abundance measurements derived from shotgun metagenomic sequencing, and is available as an online tool and as a Python module. MUSiCC is developed by the Borenstein group at the University of Washington and is available online at:

http://elbo.gs.washington.edu/software_musicc.html.

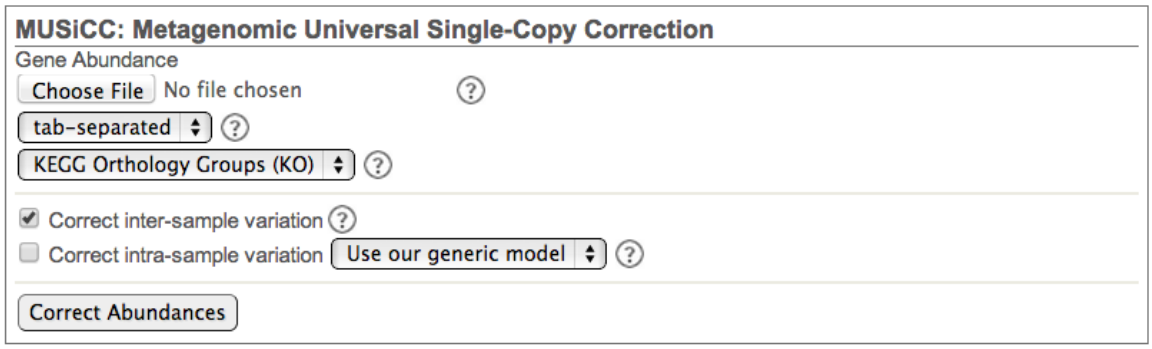
MUSiCC>Web

MUSiCC>Web allows researchers to normalize and correct their metagenomic gene abundance measurements online, and requires only a web browser. The MUSiCC>Web application is built on top of the MUSiCC>Python module.

Overview of use

To normalize and correct the abundance measurements of genes using MUSiCC>Web, the user simply uploads their gene abundance file to the MUSiCC server by filling out the form shown in Figure 1 a). After checking the integrity of the file, MUSiCC>Web normalizes and corrects the measured abundance. A link to the file containing the corrected abundances is displayed along with statistics of the data and correction procedures, as shown in Figure 1 b).

a)



b)

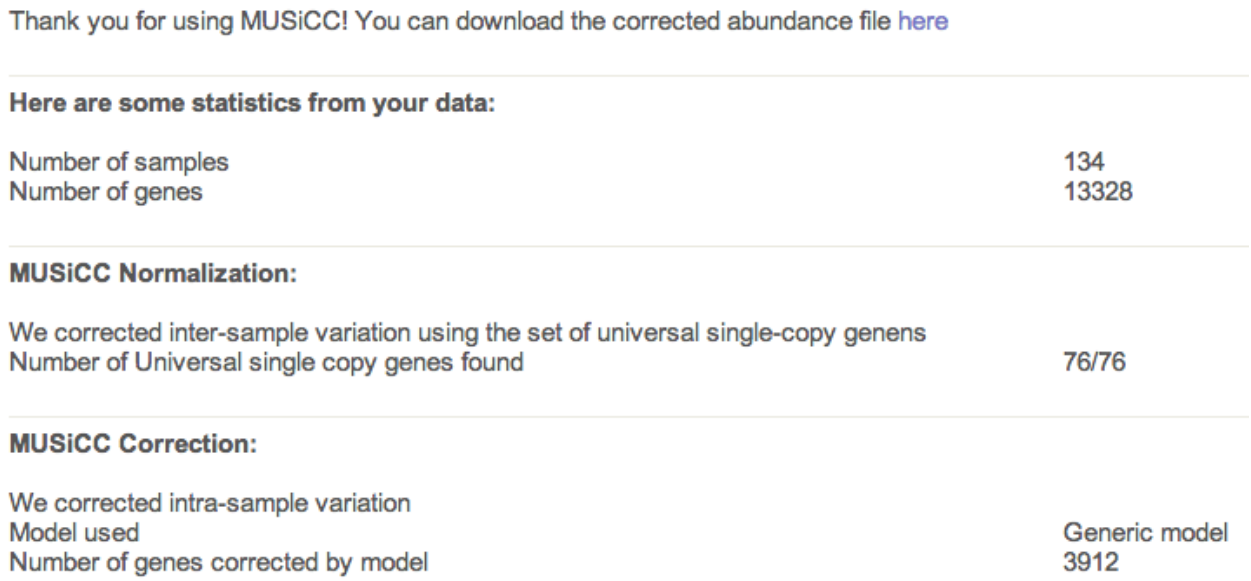


Fig. 1: The MUSiCC>Web interface. A) The user is first presented with a form to upload the abundance file and select the analysis options. B) The gene abundance file is corrected and available for download, along with statistics collected during the correction.

Setting up the analysis

Selecting the abundance file to correct

Choose File / Browse: Select the abundance file on your local drive. In the drop-down box below, choose the appropriate file type.

MUSiCC>Web supports abundance files formatted as plain text with various delimiters. Gene abundance files have one line per gene, with the gene ID appearing as a row header, followed by all the abundance measurements corresponding to different samples separated by a space, comma or tab. Each line must use the same separating character. The first line of the file must contain a column header with sample names.

Analysis Options

Once you have specified the abundance file, the next step is to choose what types of corrections should be applied to the data:

Correct inter-sample variation: Selecting this will perform the normalization step of MUSiCC, correcting inter-sample variation.

Correct intra-sample variation: Selecting this will perform the intra-sample correction, with the generic model learned from the Human microbiome project stool samples.

MUSiCC>Python

MUSiCC>Python is a stand-alone Python module that implements the MUSiCC functionality. It is distributed under the GPL and can be readily incorporated into custom analysis tools.

Installation Instructions

Prerequisites for installing:

In order for MUSiCC to run successfully, the following Python modules should be pre-installed on your system:

- Numpy (http://www.numpy.org/)
- Scipy (http://www.scipy.org/)
- Sklearn (http://scikit-learn.org/stable/)

To install MUSiCC>Python, simply download the package from http://depts.washington.edu/elbogs/MUSiCC/MUSiCC_Python.zip. This is zip archive containing the following files/directories:

- *MUSiCC.py*: The MUSiCC Python module.
- *Data/*: A directory containing several data files MUSiCC requires to run properly.
- *Example/*: A directory containing examples of input and output files.

- *COPYING*: A copy of the GNU General Public License. This is required to be distributed with the MUSiCC>Python package.

Interface

The MUSiCC>Python module handles all calculations internally. MUSiCC>Python offers an interface to the MUSiCC functionality via the command line:

usage:

MUSiCC.py [-h] [-o OUTPUT_FILE] [-if {tab,csv,biom}]

[-of {tab,csv,biom}] [-n] [-c {use_generic,learn_model}]

[-perf] [-v]

input_file

positional arguments:

input_file Input abundance file to correct

optional arguments:

-h, --help show this help message and exit

-o OUTPUT_FILE, --out OUTPUT_FILE

Output destination for corrected abundance (default:

MUSiCC.tab)

-if {tab,csv,biom}, --input_format {tab,csv,biom}

Option indicating the format of the input file

(default: tab)

-of {tab,csv,biom}, --output_format {tab,csv,biom}

Option indicating the format of the output file

(default: tab)

-n, --normalize Apply MUSiCC normalization (default: false)

-c {use_generic, learn_model}, --correct {use_generic,learn_model}

Correct abundance per-sample using MUSiCC (default:

false)

-perf, --performance Calculate model performance on various gene sets (may add to running time) (default: false)

-v, --verbose Increase verbosity of module (default: false)

Example

In the *Example* directory, the file *Gene_vs_Sample_HMP_STOOL.tab* contains gene abundance measurements of 5 stool samples downloaded from the Human Microbiome Project (HMP): ftp://public-ftp.hmpdacc.org/HMMRC/kegg_kos.pcl.gz. Using this file as input for MUSiCC results in the normalized and corrected abundance file: *Example/MUSiCC.tab*.

The command used is the following:

Python MUSiCC.py Examples/Gene_vs_Sample_HMP_STOOL.tab -o Examples/MUSiCC.tab -n -c learn_model

MUSiCC>MATLAB

MUSiCC>MATLAB is a MATLAB function that implements the MUSiCC functionality. It is distributed under the GPL and can be readily incorporated into custom analysis tools.

Installation Instructions

Prerequisites for installing:

In order for MUSiCC to run successfully, MATLAB needs to be installed on your system.

To install MUSiCC>MATLAB, simply download the package from http://depts.washington.edu/elbogs/MUSiCC/MUSiCC_MATLAB.zip. This is zip archive containing the following files/directories:

- *MUSiCC.m*: The MUSiCC MATLAB function.
- *Data/*: A directory containing several data files MUSiCC requires to run properly.

- *Example/*: A directory containing examples of input and output files.
- *COPYING*: A copy of the GNU General Public License. This is required to be distributed with the MUSiCC>MATLAB package.

Interface

The MUSiCC>MATLAB module handles all calculations internally. MUSiCC>MATLAB offers an interface to the MUSiCC functionality via the MATLAB environment:

usage:

>> CorrectedAbundance = MUSiCC(abundance_file, varargin)

required input:

abundance_file Input abundance file to correct

optional input:

output_file <str>: name for the output file (default: MUSiCC.tab)

normalize <'true'/'false'>: Determines if MUSiCC normalization is performed (default 'true')

correct_method <'use_generic'/'learn_model'>: Determines if MUSiCC correction is performed (default 'none')

show_scores <'true'/'false'>: Determines if MUSiCC reports scores for variuos gene sets (default 'true')

verbose <'true'/'false'>: Determines if MUSiCC increases verbosity (default 'true')

Output:

CorrectedAbundance: the corrected gene abundance matrix

Example

In the *Example* directory, the file *Gene_vs_Sample_HMP_STOOL.tab* contains gene abundance measurements of 5 stool samples downloaded from the Human Microbiome Project (HMP): ftp://public-ftp.hmpdacc.org/HMMRC/kegg_kos.pcl.gz. Using this file

as input for MUSiCC results in the normalized and corrected abundance file: *Example/MUSiCC.tab*.

The command used is the following:

```
>> MUSiCC('Examples/Gene_vs_Sample_HMP_STOOL.tab', 'output_file',
'Examples/MUSiCC.tab' ,'correct_method', 'learn_model')
```