MaAsLin User Guide v1.0

February 28, 2012

Timothy Tickle

Curtis Huttenhower**Table of Contents**

A. Introduction to MaAsLin

B. Related Projects

C. MaAsLin Inputs

D. Process Flow Overview

E. Process Flow Detail

F. Expected Output Files

G. Description of Output Files

H. Troubleshooting**A. Introduction to MaAsLin**

MaAsLin is a multivariate statistical framework that finds associations between clinical metadata and some data. In our context we use it to assess associations between clinical metadata and microbial community abundance or microbial community percent function. Although, this context is the reason MaAsLin was developed it is not required for it's use. Essentially MaAsLin performs boosted additive general linear models between one group of data (metadata/the predictors) and another group (in our case taxanomic abundances/the response).

Given that metagenomic data is sparse, the boosting is used to select metadata that show some potential to be useful in a linear model between the metadata and abundances. This is similar to any model selection or regularization technique. In the context of metadata and community abundance, a sample's metadata is boosted for one otu. The metadata data that is selected for use by boosting is then used in a general linear model using metadata as predictors and otu abundance as the response. This occurs for every otu and sample. Given we work with proportional data, the Yi (abundances) are arcsin(sqrt(Yi)) transformed. A final formula for the association is as follows arcsin(sqrt(Yi)) = bo sum BiXi + ei . If you are not working with proportional data or have a different transformation you prefer, this can be updated in the .R file described later in section C. MaAsLin inputs.

**B. Related Projects**

Two other projects exist at www.bitbucket.com which may help in your analysis are MetaCheck and QiimeToMaAsLin. MetaCheck is a utility project still in development which is targeted at diagnosis and visualizing the metadata in a study. The intent is to create visualizations to comment on the balance, correlational structure, and other properties of your data to better use MaAsLin. QiimeToMaAsLin is a project that reformats abundance files from Qiime for MaAsLin. Only a current format of Qiime is supported for this project, as more formats are encountered, more will be supported. If interested in using either script, these projects are upstream in analysis from MaAsLin.

**C. MaAsLin Inputs**

There are 3 required files for each project, the "\*.read.config" file, the "\*.pcl" file, and the "\*.R" script. Although the "\*" in the file names can be anything, it needs to be identical for all three files. All three files need to be in the ../sfle/input/maasalin/input/ folder. Details of each file follow:

1. "\*.pcl"

A PCL file is the file that contains all the data. This file is metadata/otu as rows and samples as columns. All metadata rows should come first before any abundance data. This is a tab delimited text file with the extension ".pcl" .

2. "\*.read.config"

A read config file allows one to indicate what data is read from a PCL file without having to change the pcl file or change code. This means one can have a pcl file which is a superset of metadata and abundances which includes data MaAsLin you are not interested in for the run. This file is a text file with ".read.config" as an extension. Inside the file the first line should be ###

3. "\*.R"

The R script file is using a call back programming pattern which allows one to add/modify specific code to customize analysis without touching the main MaAsLin engine. This file also allows one to set certain constants like lower threshold for abundance measurement filtering in quality control. A generic R script is provided and can be used for any study. The R script can be modified to add quality control or formatting of data, add ecological measurements, affect the MFA visualization, or other changes. A default script is given which can be generically used or modified.

**D. Process Flow Overview**

1. Obtain your abundance table.

2. Obtain your metadata.

2. Format and combine your abundance table and metadata as a pcl file for MaAsLin.

3. Create your read.config file.

4. Create your R script or use the default.

5. Place .pcl, .read.config, .R files in ../sfle/input/maasalin/input/

6. Run.

7. Discover amazing associations in your results!

**E. Process Flow Detail**

1. Obtain your abundance table.

Abundance tables are normally derived from sequence data using Mothur or Qiime. Please refer to thier documentation for further details.

2. Obtain your metadata.

Metadata would be information about the samples in the study. For instance, one may analyze a case / control study. In this study, you may have a disease and healthy group (disease state), the sex of the patents (patient demographics), medication use (chemical treatment), smoking (patient lifestyle) or other types of data. All would be study metadata. This section can have any type of data (factor, ordered factor, continuous, integer, or logical variables). If a particular data is missing for a sample for a metadata you may leave it blank or write NA. It is preferable to write NA so that, when looking at the data, it is understood the metadata is missing and it's absence is intentional and not a mistake.

3. Format and combine your abundance table and metadata as a pcl file for MaAsLin.

i. Row 1 is expected to be #ID\_indicator and then sample ids in each following column separated by tabs.

ii. Rows of metadata. Each row is one metadata, the first column entry being the name of the metadata and each following column being the metadata value for that each sample.

iii. Row of taxa/otu abundance. Each row is one taxa/otu, the first column entry being the name of the taxa/otu followed by abundances of the taxa/otu per sample.

iv. Abundances should be normalized by dividing each abundance by the sum of the column (sample) abundances.

v. Here is an example of the contents of an extremely small pcl file.

#SampleID Sample1 Sample2 Sample3 Sample4

metadata1 True True False False

metadata2 1.23 2.34 3.22 3.44

metadata3 Male Female Male Female

taxa1 0.022 0.014 0.333 0.125

taxa2 0.406 0.029 0.166 0.300

taxa3 0.571 0.955 0.500 0.575

4. Create your read.config file.

A \*.read.config file is a structured text file used to indicate which data in a \*.pcl file should be read into MaAsLin and used for analysis. Allows one to keep their \*.pcl file intact while varying analysis. Hopefully, this avoids errors which may occur while manipulating the pcl files.

Here is an example of the contents of a \*.read.config file.

Matrix: Metadata

Delimiter: TAB

Read\_Columns: 1-4,6,8,10-12

Read\_Rows: 2-15

DT\_Numeric: 6

DT\_Factor: 1-4

DT\_Logical: 11

DT\_Integer: 8,10,12

Matrix: Abundance

Delimiter: TAB

Read\_Columns: 18-100

Read\_Rows: 2-15

DT\_Numeric: 18-100

The minimal requirements for a MaAsLin .read.config file are Matrix:, Read\_Columns:, and Read\_Rows: entries should be specified. Metadata needs to be named "Matadata" for the metadata section and "Abundance" for the abundance section.

A minimal example is shown here:

Matrix: Metadata

Read\_Columns: 4,6,7,109,111,113,114,116,117,282,283,284

Read\_Rows: 2-232

Matrix: Abundance

Read\_Columns: 286-737

Read\_Rows: 2-232

With this minimal example, the delimiter of the file is assumed to be a tab and the data types of each metadata and abundance entry are inferred by the system. If you would like more explicit control on the interpretation of the type of data you enter, use the DT\_Factor:, DT\_Integer:, DT\_Logical:, DT\_Numeric:, and DT\_Ordered\_Factor: Commands to force data typing.

5. Create your R script or use the default.

The R script is used to set default thresholds, to add code which manipulates your data before analysis, and for manipulating the multifactoral analysis figure. A default “\*.R” script is available with the default MaAsLin project. In practice, “\*.R” script can be completely blank (in this case defaults will be used). The following details parts of the default “\*.R” script which may be modified at your discretion:

########

6. Place .pcl, .read.config, .R files in ../sfle/input/maasalin/input/

All studies should have these 3 input files in the maaslin/input folder.

7. Run.

Go to ../sfle and type the following: scons output/the\_Name\_of\_your\_pcl\_file\_without\_the\_extension

8. Discover amazing associations in your results!

**F. Expected Output Files**

The following files will be generated per MaAsLin run. In the following listing the term *projectname* refers to what you named your “\*.pcl” file without the extension.

**Troubleshooting:**

*projectname*.txt

Contains the detail for the statistical engine. Is useful for detailed troubleshooting.

data.tsv

The data matrix that was read in (transposed). Useful for making sure the correct data was read in.

data.read.config

Can be used to read in the data.tsv .

metadata.tsv

The metadata that was read in (transposed). Useful for making sure the correct metadata was read in.

metadata.read.config

Can be used to read in the data.tsv .

read\_merged.tsv

The data and metadata merged (transposed). Useful for making sure the merging occurred correctly.

read\_merged.read.config

Can be used to read in the read\_merged.tsv .

read\_cleaned.tsv

The data read in, merged, and then cleaned. After this process the data is written to this file for reference if needed.

read\_cleaned.read.config

Can be used to read in read\_cleaned.tsv .

ProcessQC.txt

Contains quality control for the MaAsLin analysis. This includes information on the magnitude of outlier removal.

**Analysis:**

*projectname*-metadata.txt

Each metadata will have a file of associations. Any associations indicated to be performed after initial boosting is recorded here. Included are the information from the final general linear model (performed after the boosting) and the FDR corrected p-value (q-value). Can be opened as a text file or spreadsheet.

*projectname*-metadata.pdf

Any association that had a q-value less than or equal to the threshold given in the “\*.R” file (default is 0.25) will be plotted here. If this file does not exist, the projectname-metadata.txt should not have an entry that is less than or equal to the threshold. Factor data is plotted as knotched box plots; continuous data is plotted as a scatter plot with a line of best fit.

*projectname*.pdf

Contains the multifactoral analysis visualization. This visualization is presented as a build and can be affected by modifications in the R.script

*projectname* \_Summary.txt

Any entry in the projectname-metadata.pdf are collected together here. Can be opened as a text file or spreadsheet.

**G. Description of Output Files**

**H. Troubleshooting**

1. When using the command "scons output/maaslin/..." to run my projects I get the message:

ImportError: No module named sfle:

File "/home/user/sfle/SConstruct", line 2:

import sfle

Solution: You need to update your path. On a linux terminal in the sfle directory type the following.

export PATH=/usr/local/bin:`pwd`/src:$PATH

export PYTHONPATH=$PATH

2. When trying to run a script I am told I do not have permission even though file permissions have been set for myself.

Solution: Most likely, you have not set the script you wrote as an executable.

If you right click on your file, you should be able to change the properties of the file to be executable.