Meta-Storms 2 Users' Manual

Version: Std 2.2.1
Release date: Jan 4, 2019

Introduction

Meta-Storms 2 is the standalone implementation of the Microbiome Search Engine (MSE; *http://mse.single-cell.cn*). MSE is a search engine designed to efficiently search a database of microbiome samples and identify similar samples based on phylogenetic or functional relatedness. Meta-Storms 2 consists of the following steps: (*i*) creating a database composed of reference microbiome samples, and (*ii*) searching for similar samples in the database with given query microbiome sample(s) via phylogenetic similarities. Meta-Storms 2 relies on an advanced indexing algorithm, providing a fast, and constant, search speed in very large databases.

Download

The latest release is available at:

http://mse.single-cell.cn/

Packages

At present, Meta-Storms 2 provides two alternative packages for installation.

Prebuilt binary package (recommended)

Meta-Storms 2 prebuilt binary package, with all the tools integrated, is available for Linux (64 bit) and Mac OS X.

Source code package

Meta-Storms 2 source code package is also available for building and installation for other Unix/Linux/Mac OS X based operating systems.

System requirement and dependency

Hardware Requirements

Meta-Storms 2 only requires a standard computer with sufficient RAM to support the operations defined by a user. For typical users, this would be a computer with about 2 GB of RAM. For optimal performance, we recommend a computer with the following specs:

RAM: 8+ GB
CPU: 4+ cores, 3.3+ GHz/core

Software Requirements

OpenMP library is the C/C++ parallel computing library. Most Linux releases have OpenMP already been installed in the system. In Mac OS X, to install the compiler that supports OpenMP, we recommend using the Homebrew package manager:

brew install gcc --without-multilib

Installation guide

Automatic Installation (recommended)

At present, Meta-Storms 2 provides a fully automatic installer for easy installation.

a. Extract the package:

tar -xzvf meta-storms-2-std-bin.tar.gz

b. Install by installer:

cd meta-storms-2-std
source install.sh

The package should take less than 1 minute to install on a computer with the specifications recommended above.

Tips for Automatic Installation

1. Please "cd meta-storms-2-std" directory, before running the automatic installer.
2. The automatic installer configures the environment variables to the default configuration specified in the file of "~/.bashrc" or "~/.bash_profile". If you prefer to configure the environment variables to other configuration file, please choose the option of manual installation below.
3. If the environment variables are not activated automatically, please enable them manually by running the command "source ~/.bashrc".
4. If the automatic installer fails, Meta-Storms can still be installed manually by the following options.

Manual Installation

If the automatic installer fails, Meta-Storms 2 can still be installed manually.

a. Extract the package:

tar –xzvf meta-storms-2-std-src.tar.gz

b. Configure the environment variables (the default environment variable configuration file is "~/.bashrc"):

export MetaStorms=Path to Meta-Storms 2
export PATH="$PATH:$MetaStorms/bin/"
source ~/.bashrc

c. Compile the source code (this is required **only** when installing the source code package):

cd meta-storms-2-std
make

Notice before use

1. For source code package based installation, please make sure proper versions of compilers have been installed: gcc 4.4 or higher for Linux / gcc-8 or higher for Mac OS X (refer to Software Requirements).
2. Meta-Storms 2 optionally accepts microbiome sample(s) that are pre-processed by Parallel-META 3 (version 3.2 or hihger; *http://bioinfo.single-cell.cn/parallel-meta.html*) or QIIME (version 1.9.1; http://qiime.org). Meta-Storms 2 software can also accept OTU tables (refer to **File format**). However, if starting from DNA sequences, Parallel-META 3 and QIIME are the recommended software for converting amplicon sequences to OTU tables (refer to **Pre-computing**).
3. Make sure that Meta-Storms 2 has the write permission in the output path.
4. We strongly recommend reading this manual carefully before using Meta-Storms 2.

Pre-computing

To use Meta-Storms, all sequences of microbiome samples must be pre-computed and profiled against the Greengenes database (version 13-8) by Parallel-META 3 (version 3.2 or higher; *http://bioinfo.single-cell.cn/software.html*) or QIIME (version 1.9.1; *http://qiime.org*). Then the profiling results will be used as input to Meta-Storms 2.

Pre-computing by Parallel-META 3

For a give sequence file (FASTA or FASTQ format, eg. sample1.fa), to convert the sequences to OTUs by Parallel-META 3:

PM-parallel-meta -f F -r sample1.fa -o sample1.out

Then the output file *sample1.out/classification.txt* is qualified as the input for Meta-Storms 2 (refer to **Single sample**). For multiple samples as input, samples should be listed in the sample list (refer to **Sample list**).

Pre-computing by QIIME

For a give sequence file (FASTA format, eg. sample1.fa), to convert the sequences to OTUs by QIIME

pick_otus.py -m uclust_ref --suppress_new_clusters -i sample1.fa -o sample1.out

MetaDB-parse-qiime-otu -i sample1.out/sample1_otus.txt -o sample1.out/classification.txt

Then the output file *sample1.out/classification.txt* is qualified as the input for Meta-Storms 2 (refer to **Single sample)**. For multiple samples as input, samples should be listed in the sample list (refer to **Sample list**).

Example dataset

Here we provide a demo dataset with 20 human oral microbiome samples in two different healthy statuses from *Huang, et al., 2014**. The pre-computing result (in the OTU table format and derived from Parallel-META 3) and the meta-data are in the "**example**" folder in the installation package. We use this dataset to demonstrate all the following example commands. Please change your work directory to the "**example**" folder by

cd example

* Huang, S., et al., *Predictive modeling of gingivitis severity and susceptibility via oral microbiota*. ISME J, 2014. 8(9): p. 1768-80.

Build a MSE database

Build a MSE database

The command of **MetaDB-make** builds a new MSE database for Meta-Storms 2 based search from the given samples. Samples are listed in either *(i)* single sample list (for Parallel-META 3 format, by -i or -l with optional –p,), or *(ii)* OTU table (OTU table format, by -T). It outputs a database file (*.*mdb*).

Usage:

MetaDB-make [-option] value

[Input options]

-i or **-l** Input filename list (for multiple samples in a sample list, refer to Sample list)

-p List file path prefix for '-i' or '-l' [Optional for -i or -l] (for sample list, refer to **Sample list**)

or

-T (upper) Input OTU table (*.Count) (for OTU table format , refer to **OTU table**)

or

-d (*.mdb) Make the HDD mode data files for a database

[Output options]

-o Output database name, default is "database.mdb"

-H (upper) If enable the HDD mode (low RAM usage but slower, refer to the **HDD mode**), T(rue) or F(alse), default is F

[Other options]

-h Help

Example (make sure you are in "**example**" path):

MetaDB-make -T taxa.OTU.Count -o database

HDD mode

The HDD (Hard Drive Disk) mode uses the re-encoding technique to minimize the RAM usage for database search (although the mode is slower). When the HDD mode is enabled via –H t, **MetaDB-make** will generate accessory data named as *.*mdb.hdd* under the same directory of the output database (*.*mdb*). For extremely large databases (e.g., sample number > 10,000), we strongly recommend users to enable the HDD mode to minimize the RAM consumption.

For an existing database (*.*mdb*), HDD mode can also be enabled by making its HDD files via the command below. Then the *.*mdb.hdd* would be generated and stored under the same directory as the database

Example (make sure you are in "**example**" path):

MetaDB-make -d database.mdb

Merge MSE databases

The command of **MetaDB-merge** merges two existing databases (*.*mdb*) into one.

Usage:

MetaDB-merge [-option] value

[Input and Output options]

-1 The 1st database name [Required]

-2 The 2nd database name [Required]

-o Merged output database name, default is "database_merge.mdb"

[Other options]

-h Help

Example: Here you can make another database named as "*database_2.mdb*"

MetaDB-merge -1 database.mdb -2 database_2.mdb -o database_merged

Search the MSE database

Search via Meta-Storms 2

Query sample(s) should also be pre-computed by Parallel-META 3 or QIIME using the Greengenes database as reference (refer to **Pre-computing**). The database is built by **MetaDB-make** (**.mdb*). Meta-Storms 2 supports the index-based query, which features an extremely fast and constant search speed against very large microbiome databases.

The query sample(s) can be provided via either (*i*) single sample (for a single sample in Parallel-META 3 format, by -i), or (*ii*) single sample list (for multiple samples in Parallel-META 3 format, by -l with optional -p), or (*iii*) OTU table (for OTU table format, by -T).

We also recommend users to enable the HDD mode for large databases to minimize the RAM consumption (e.g., sample number > 10,000) (See **HDD mode**).

Usage:

MetaDB-search [-option] value

[Database options]

-**d** Database file (*.mdb) [Required]

-**H** (upper) Whether to enable the HDD mode (low RAM usage but slower), T(rue) or F(alse), default is F

-**P** (upper) Path for the HDD mode data files [Optional for '-H T']

[Input options]

-**i** Single input file name [Conflicts with -l and -T] (for a single sample, refer to **Single sample**)

or

-**l** Input filename list (for multiple samples in a sample list, refer to **Sample list**)

-**p** List file path prefix for '-l' [Optional] (**Sample list**)

or

-**T** (upper) Input OTU table (*.Count) (for OTU table format, refer to **OTU table**)

[Output options]

-**o** Output file, default is "query.out"

[Advanced options]

-**n** Number of the matched sample(s), default is 10

-**m** Minimum similarity of the matched sample(s), range (0.0 ~ 1.0], default is 0

-**w** Abundance weighted or unweighted, T(rue) or F(alse), default is T

[Other options]

-**t** CPU core number, default is auto (use all CPUs)

-**h** Help

Example (make sure you are in "**example**" path):

MetaDB-search -d database.mdb -T taxa.OTU.Count -o query.out

Search output

MetaDB-search generates a number of matches, each with its sample ID and its similarity score (always between 0 and 1) to the query. In the output, for each of the query samples, all of its matches are listed in tandem in a single line, e.g.

#	Query	Match	Similarity	Match	Similarity
Query:	q_id_0	ref_id_x	0.9823	ref_id_y	0.9758
Query:	q_id_1	Ref_id_m	0.9541	ref_id_n	0.9386

In the output above, the first query sample (q_id_0) matches against the reference sample (ref_id_x) with a similarity of 0.9823. In addition, q_id_0 also matches ref_id_v with a similarity of 0.9758. The number of matches is assigned by the parameter -n, and default is 10.

The similarity between query sample(s) and matched sample(s) is a phylogeny-based similarity that is computed using the Meta-Storms scoring function. This algorithm takes the relative abundance of OTUs and their binary phylogeny between two samples as input, and output their quantitative similarities (always between 0 and 1). For high performance and parallel computing, this algorithm is optimized by non-recursive transformation, memory recycling and variable reallocation. Please also refer to "Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data, *Bioinformatics*, 2012" for details.

Multiple classification based on search result

Meta-data prediction

Important features of the query sample, such as the meta-data of habitat, status, etc., can potentially be predicted based on the meta-data of its matches. From a search output generated by **MetaDB-search**, the meta-data of the query sample can be predicted by:

Usage:

MetaDB-parse-meta [Options] Value

[Input and output options]

-i Input file name (the output of MetaDB-search) [Required]

-m Input meta-data file name (meta-data of the database in MetaDB-search) [Required]

-l Meta-data column, default is 1 (exclude the ID column)

-o Output file name, default is "query.out.meta"

[Advanced options]

-r Number of predicted meta-data, default is 1 (refer to the output for details)

-b Base of the similarity in the input file, default is 0

-s Number of skipped matches in the input file (usage example see below)

[Other options]

-h Help

Usage for the -s:

When the query sample has already been included in the database, the search result must contain the query samples itself as the top hit since they have the 100% similarity, which causes the bias in meta-data prediction. Here we can use the parameter –s 1 to exclude this top hit in the meta-data prediction to avoid such bias.

Example (make sure you are in "**example**" path):

MetaDB-parse-meta -i query.out -m meta.txt -o query.out.meta

Multiple classification output

MetaDB-parse-meta generates the predicted meta-data with the assigned scores (always between 0 and 1). In the output, for each of the query samples, all of its predicted meta-data are listed in tandem in a single line, e.g.

#ID	Meta-data	Score	Meta-data	Score
q_id_0	Healthy	0.75	Disease	0.25
q_id_1	Disease	0.72	Healthy	0.28

In the output above, the first query sample (q_id_0) is predicted as "Healthy" with a score of 0.75, and "Disease" with a score of 0.25. The predicted meta-data are sorted by their scores.

The number of predicted meta-data is assigned by parameter -r, and default is 1 (i.e., only reporting the predicted meta-data with the highest score).

Microbiome Novelty Score (MNS) based on search results

Calculate the Microbiome Novelty Score (MNS)

With the search output generated by **MetaDB-search**, the Microbiome Novelty Score (MNS) of each sample can be calculated by:

Usage:

MetaDB-parse-mns [Options] Value

[Input and output options]

-i Input file name (the output of MetaDB-search) [Required]

-o Output file name, default is "query.out.mns"

[Advanced options]

-b Base of the similarity in the input file, default is 0

-s Number of skipped matches in the input file (usage example see below)

[Other options]

-h Help

Usage for the -s:

When the query sample has already been included in the database, the search result must contain the query samples itself as the top hit since they have the 100% similarity, which causes bias in calculating the MNS. Here we can use the parameter –s 1 to exclude this top hit in calculating the MNS.

Example (make sure you are in "**example**" path):

MetaDB-parse-mns -i query.out -o query.out.mns

Microbiome Novelty Score (MNS) output

MetaDB-parse-mns generates the MNS (always between 0 and 1) of each query sample in a single line, e.g.

#ID	MNS
q_id_0	0.06
q_id_1	0.12

In the output above, the first query sample (q_id_0) reports a MNS of 0.06.

Other tools

Parse the QIIME OTUs for Meta-Storms 2

MetaDB-parse-qiime-otu can parse the OTU map generated by QIIME from sequences into a format that is required by Meta-Storms 2. OTUs should be picked by reference based methods against GreenGenes 13-8 with 97% similarity threshold (refer to **Pre-computing by QIIME** for details).

Usage:

MetaDB-parse-qiime-otu [Options] Value

[Input and output options]

-i Input file name (the OTU map by QIIME, eg. pick_otu.py) [Required]

-o Output file name, default is "classification.txt"

[Other options]

-h Help

File format (supplementary)

Meta-Storms 2 accepts the alternative two formats as input.

Single sample file and sample list

A single sample is the OTUs and taxonomy information of a single microbiome sample profiled by Parallel-META 3 or QIIME from the amplicon sequences (refer to **Pre-computing** for details). It is a plain-text file, normally named as "*classification.txt*". An example of the single sample is below:

#Database_OTU	Count
OTU_1	10
OTU_2	17
OTU_3	38

A sample list is a plain-text file for listing multiple samples (by –l) as Meta-Storms 2 input, which consists of two columns: the sample IDs and the directories of samples' "*classification.txt*" files, e.g.

Sample_1	/home/data/single_sample/Sample_1/classification.txt
Sample_2	/home/data/single_sample/Sample_2/classification.txt

The directory can be either absolute directory or relative directory. Meta-Storms 2 also provides an optional parameter –p to add a prefix for the all the directories in the sample list in case of a relative directory is preferred.

OTU table

An OTU table is a plain-text file that contains the OTUs and their sequence numbers for each of multiple samples. An example of the OTU table is bellow

#Sample_ID	OTU_1	OTU_2	OTU_3	OTU_4	OTU_5
Sample_1	10	17	38	2	2
Sample_2	0	5	57	0	0
Sample_3	2	35	7	0	0
Sample_4	58	30	23	3	0
Sample_5	95	5	5	4	0

Contact

Any problem please contact MSE development team:

JING Gongchao E-mail: jinggc@qibebt.ac.cn