Run MASH using iMicrobe

Alise Ponsero

Abstract

How to run MASH version (Ondov *et al.* 2016) through the iMicrobe plaform.

MASH is a kmer-based sample-comparison tool using the MinHash algorithm to reduce the sample dimensionality and calculating a distance betwwen sample.

More informations about MASH can be found here : https://mash.readthedocs.io/en/latest/

Citation: Alise Ponsero Run MASH using iMicrobe. **protocols.io**
dx.doi.org/10.17504/protocols.io.s9ueh6w
Published: 12 Sep 2018

Guidelines

More informations and details about MASH can be found there : https://mash.readthedocs.io/en/latest/

Several parameters are available to the user through the iMicrobe app.

- **kmer size :**
  As in any k-mer based method, larger k-mers will provide more specificity, while smaller k-mers will provide more sensitivity. Larger genomes will also require larger k-mers to avoid k-mers that are shared by chance. The k-mer size can be tuned by the iMicrobe user using the "kmer size" parameter. The default setting is 21bp.
- **Sketch size :**For sequences to be compared with MASH, they must first be *sketched*, which creates vastly reduced representations of them. Sketch size corresponds to the number of (non-redundant) min-hashes that are kept. Larger sketches will better represent the sequence, but at the cost of larger sketch files and longer comparison times. The sketch size can be tuned by the iMicrobe user using the "Sketch size" parameter. The default setting is 1000.

Before start

- You need a working Cyverse account to connect to iMicrobe.To explore how to log into iMicrobe, read the dedicated protocol.
- Your dataset of interest should be metagenomic reads, in a fasta or fastq format.
- In iMicrobe, there is several ways to run an app on a dataset (from the cart, from your personnal datastore and form an URL). If you need more information on how to run an app, read the protocol associated.

Protocol

Run MASH on iMicrobe

Step 1.

Note :This protocol uses mock communities available on iMicrobe. These mock communities are artificially generated 454 reads (10 million reads per file) using GemSim, from known composition profiles.

In the iMicrobe sample search page, select the mock communities to add them in your cart. In the 'tools' dropdown menu, select 'Apps'. You are presented the list of apps currently available on iMicrobe. Click on mash-all-vs-all-0.0.5u1

In the page app, provide the input files using the cart. Choose the following parameters :

- **kmer size :** by default set as 21bp
- **sketch size** : by default set as 1000.

Note : for more details on the app parameters, please read the 'Guidelines' section of this protocol.

After the job is effectively ran, you can access your results using the drop-down menu 'Tools' and selecting 'Jobs'. Select the job corresponding to your centrifuge run, and go to the section 'Outputs'.

The mash output files are now in your cyverse datastore. Click on 'Browse and view output files in the CyVerse Datastore'.

In the job folder created in the CyVerse datastore, the input fasta/fastq files are copied, along with the logs of the job (*.err and *.out). In order to retrieve your results go to the mash-out folder. It contains two folders :

- ***sketches***
  This folder contains the mash sketches index for the input files. These files have a *.msh extension.

- ***figures***
  This folder contains the MASH final outputs.
  - **distance.txt**
    This txt file contains the standard matrix output of MASH. This matrix contains the all versus all distance computation of MASH in a tabular table.
- In addition to this standard output, the iMicrobe app offers the user with some quick vizualization of their data:
  - **dendrogram.png**
    This data vizualization uses the hclust R function and a ward clustering method.
  - **dendrogram_fan.png**
    Fan dendrogram of a Newick tree.
  - **heatmap.png**
    Heatmap of the distance matrix. No clustering method is applied. The scale range from 0 (white) to 1 (dark blue).
  - **pcoa.pdf**
    PCOA applied on the dataset.
  - **tree.newick**
    Newick tree used for the dendrogram_fan vizualization.

EXPECTED RESULTS

	meta_mock1-2_c10M_single.fastq	meta_mock1_c10M_single.fastq	meta_mock2_c10M_single.fastq	meta_mock3_c10M_single.fastq	meta_mock4_c10M_single.fastq
meta_mock1-2_c10M_single.fastq	0	0.035505	0.0496454	0.0716227	0.0794837
meta_mock1_c10M_single.fastq	0.035505	0	0.0523149	0.0726524	0.0786628
meta_mock2_c10M_single.fastq	0.0496454	0.0523149	0	0.0523149	0.0747985
meta_mock3_c10M_single.fastq	0.0716227	0.0726524	0.0523149	0	0.0799006
meta_mock4_c10M_single.fastq	0.0794837	0.0786628	0.0747985	0.0799006	0