Run Centrifuge using iMicrobe

Alise Ponsero[1]

[1]University of Arizona

dx.doi.org/10.17504/protocols.io.spuednw

Oct 09, 2018

Working

iMicrobe , Metafunc course 2018

Alise Ponsero
University of Arizona

ABSTRACT

How to run Centrifuge version 1.0.4-beta (Kim *et al.* 2016) through the iMicrobe plaform.

Centrifuge is a microbial classification engine that enables rapid, accurate, and sensitive taxonomic labeling of metagenomic reads and quantification of species. The system uses an indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, optimized specifically for the metagenomic classification problem.

More informations about centrifuge can be found here : https://ccb.jhu.edu/software/centrifuge/manual.shtml

TAGS

metagenomics

imicrobe

Show tags

EXTERNAL LINK

https://www.imicrobe.us/#/apps/71

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

GUIDELINES

More informations and details about Centrifuge can be found here : https://ccb.jhu.edu/software/centrifuge/manual.shtml

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

- **Index** :
  Centrifuge indexes are the reference databases used for the taxonomic classification. Standard choices are all of the complete RefSeq prokaryotic, human and viral genomes, or using the sequences that are part of the NCBI nt database.
  The indexes provided in iMicrobe are provided by the tool authors, available here. The updataing dates for the indeces are the following :

Bacteria, Archea, Virus, Human : 12/06/2016
Bacteria, Archea : 4/15/2018
NCBI nt : 3/3/2018

- **Exclude** :
  A comma-separated list of taxonomic IDs that will be excluded in classification procedure. The descendants from these IDs will also be excluded. To find the taxonomic ID, one can use the NCBI taxonomic browser.
- **File format** :
  The user can provide a Fasta or a Fastq formated file as an input. Please use the drop down menu to select to correct format for your file.
- **Reads are paired** :
  If you are working on illumina paired reads, select that option to allow the reads to be paired before their taxonomic assignment.

- **Figure title :**
  The iMicrobe Centrifuge app provides a bubble-chart representation of the result. The title of the vizualization can be changed.

BEFORE STARTING

- 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.

Running Centrifuge on 454 datasets

1 *This protocol section 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 centrifuge-1.0.4u1.

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

- **Index :** 'Bacteria, Archea, Viruses and human (compressed)'
- **File type** : 'Read Fastq'

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 centrifuge 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 centrifuge-out folder. It contains two folders :

- ***figures***
  Contains a bubble plot vizualization of the results. The species name are displayed on the left side of the chart, and the abundance is represented by dots of increasing size. Any hits with a global abundance below 2% are removed from the chart. A csv file accompanies the chart, and displays the name and proportion of each species found in the submited samples.

- ***reports***
  This folder contains the centrifuge outputs. Two types of files will be produced :

- ***.sum**
  This is the classic centrifuge output. It display the taxonomic assigment found for each read. This output has 8 columns.

readID	seqID	taxID	score	2ndBestScore	hitLength	queryLength	numMatches
r1_from_NC_002...	cid|562	562	32360	0	234	235	1
r2_from_NC_002...	cid|562	562	62001	0	264	264	1
r3_from_NC_009...	cid|1063	1063	41756	0	422	443	1
r4_from_NC_004...	cid|1396	1396	59542	12886	460	482	1
r5_from_NC_009...	cid|1063	1063	133524	0	528	548	1
r6_from_NC_004...	cid|1282	1282	77761	0	453	455	1
r7_from_NC_009...	cid|1063	1063	32851	0	338	358	1
r8_from_NC_007...	cid|1280	1280	7056	0	99	99	1
r9_from_NC_002...	cid|562	562	148032	147456	438	438	1
r10_from_NC_00...	cid|1282	1282	4624	0	83	83	1
r11_from_NC_00...	cid|1309	1309	117821	0	460	524	1
r12_from_NC_00...	cid|562	562	128881	0	374	374	1

The first column is the read ID
The second column is the genomic sequence ID for which a hit was found
The third column is the taxonomic ID for this genomic sequence
The fourth column is th score for this classification, corresponding to a weighted sum of hits
The fifth column is the score of the next best classification for this read
The sixth column shows the approximate number of base pairs of the read matching the genomic sequence
The seventh column shows and the length of the read (or combined length of mate pairs)
The eigth column shows the number of classifications found for this reads

- ***.tsv**
  This output shows a classification summary for each genome or taxonomic unit. This output has 7 columns.

```
name    taxID   taxRank genomeSize  numReads    numUniqueReads  abundance
Azorhizobium caulinodans    7   species 5369772 3   1   0.0
Buchnera aphidicola 9   species 619958  18  8   0.0
Cellulomonas gilvus 11  species 3526441 11  9   0.0
Dictyoglomus    13  genus   0   1   0   0.0
Dictyoglomus thermophilum   14  species 1959987 1   0   0.0
Myxococcaceae   31  family  0   1   0   0.0
Myxococcus fulvus   33  species 10026214    4   1   0.0
Myxococcus xanthus  34  species 9139763 3   1   0.0
Stigmatella aurantiaca  41  species 10260756    3   2   0.0
Archangium  47  genus   0   1   0   0.0
Archangium gephyra  48  species 12489432    2   0   0.0
Chondromyces    50  genus   0   1   0   0.0
Chondromyces crocatus   52  species 11388132    2   1   0.0
Sorangium cellulosum    56  species 13907952    7   6   0.0
Planctomycetales    112 order   0   1   0   0.0
Pirellula staleyi   125 species 6196199 1   1   0.0
Planctomycetaceae   126 family  0   1   0   0.0
Isosphaera pallida  128 species 5529304 1   0   0.0
Borrelia    138 genus   0   1   0   0.0
Spirochaeta thermophila 154 species 2516433 2   1   0.0
Brachyspira hyodysenteriae  159 species 3055339 1   1   0.0
Leptospirillum ferrooxidans 180 species 2559538 1   0   0.0
Azospirillum brasilense 192 species 13978806    4   3   0.0
Azospirillum lipoferum  193 species 7223069 6   3   0
Campylobacter lari  201 species 1576113 1   0   0.0
Helicobacter    209 genus   0   5   0   0.0
Helicobacter pylori 210 species 2044699 27995   27982   0.00307586
Helicobacter acinonychis    212 species 1557588 12  3   9.47796e-269
Helicobacter mustelae   217 species 1578097 24  0   1.64306e-306
```

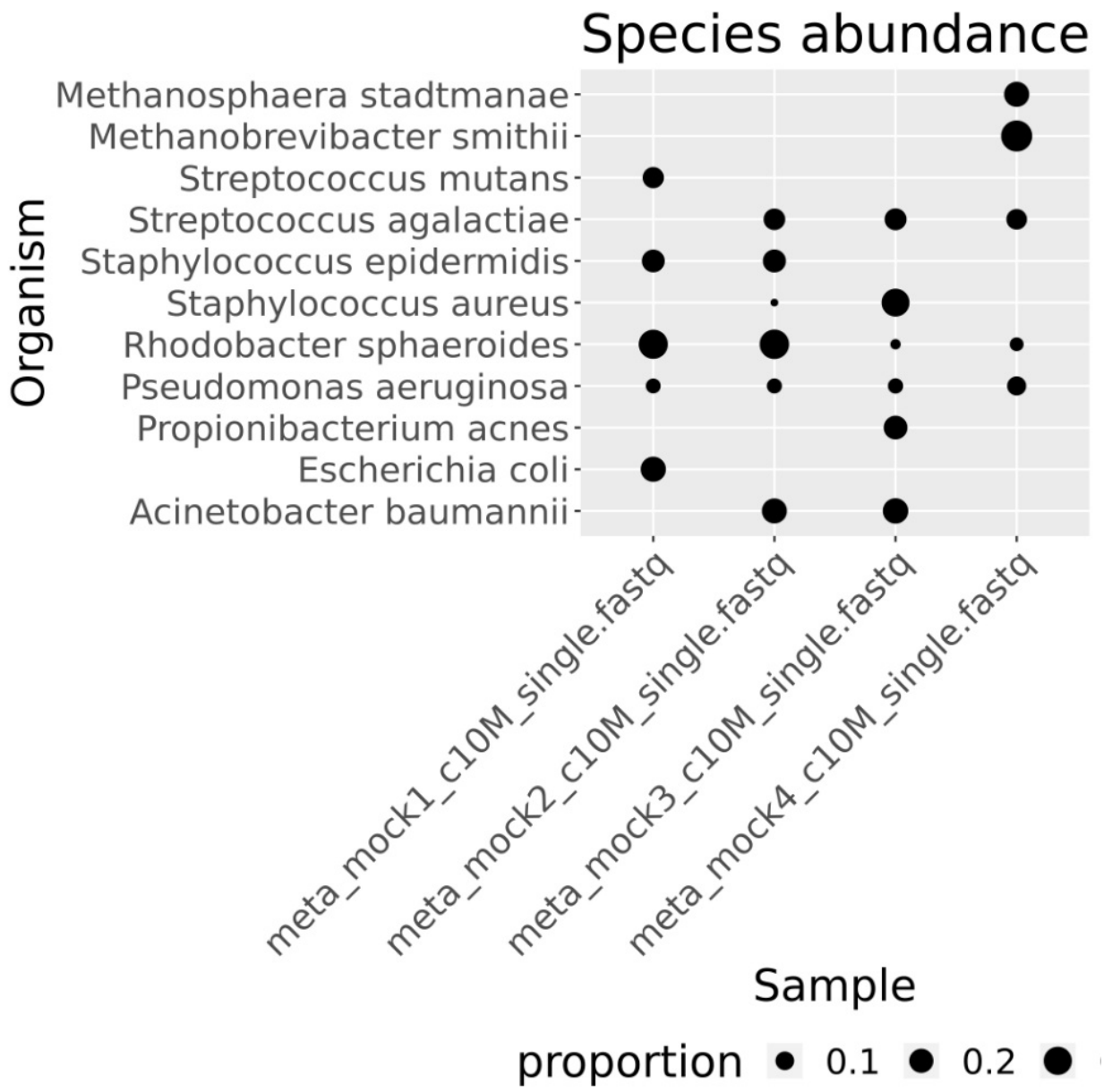
The first column is the name of the genome
The second column is the Taxonomic ID for this genome
The third column is the taxonomic rank of the genome
The fourth column is the length of the genome sequence in base pairs
The fifth column is the number of reads from the sample classified to this genome (including multi-classified reads)
The sixth column is the number of reads uniquely classified to this genome

The seventh column is the proportion of this genome normlized by its genomic length*reports*

DATASET

mock_communities1_to_4

EXPECTED RESULT



Running Centrifuge on 454 datasets excluding E. coli and P. aeruginosa

2 *This protocol section 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 centrifuge-1.0.4u1.

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

- **Index :** 'Bacteria, Archea, Viruses and human (compressed)'
- **File type** : 'Read Fastq'
- **Exclude :** 562, 287

This last parameter allows you to exclude from the analysis any reads matching the taxon 562 and 287 (E. coli and P. aeruginosa). The Taxon number correspond to the NCBI TaxID of the organism.

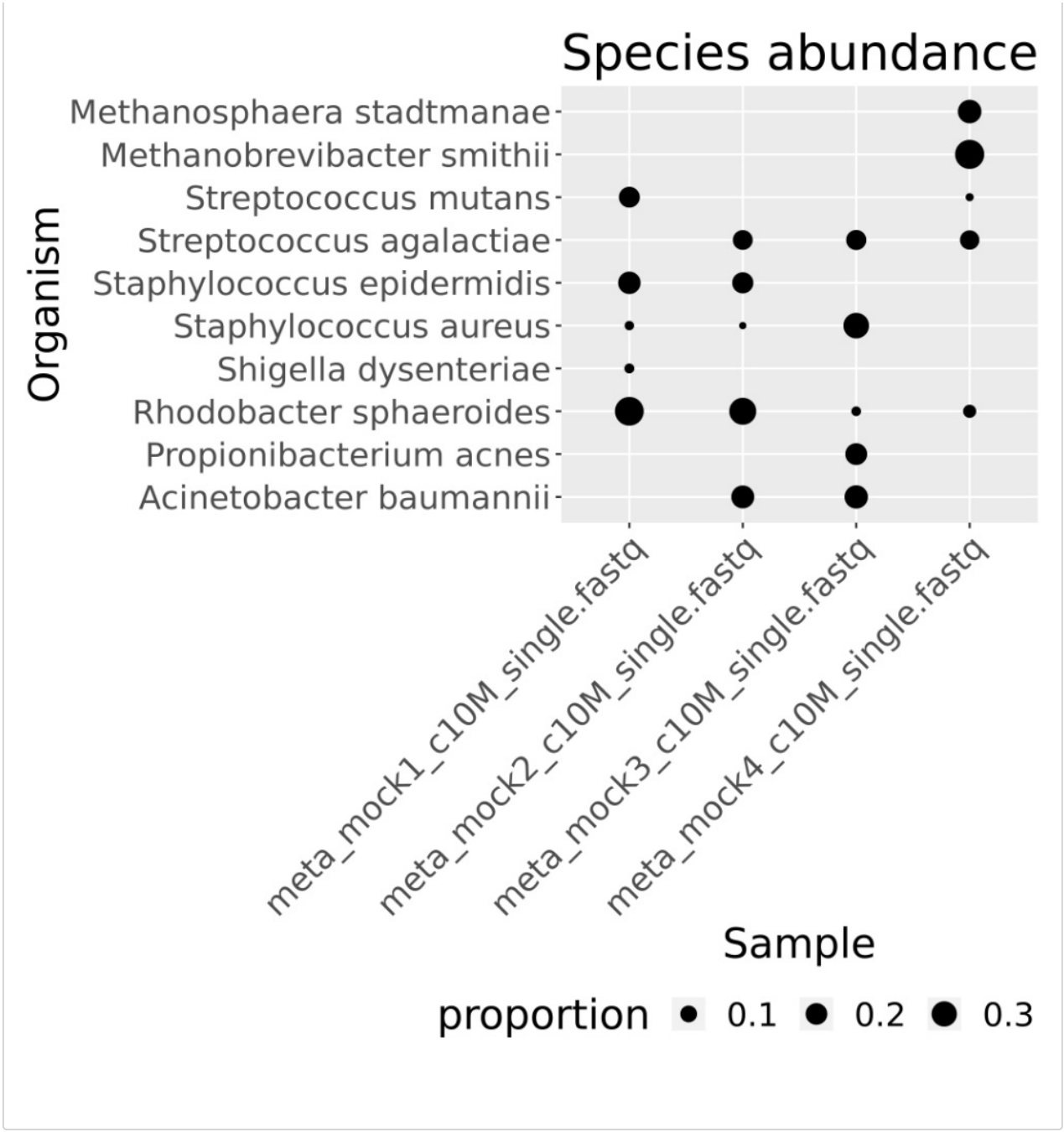
After the job is effectively ran, you can access your results using the drop-down menu 'Tools' and selecting 'Jobs'. Select the job corresponding to your centrifuge run, and go to the section 'Outputs'. The centrifuge output files are now in your cyverse datastore. Click on 'Browse and view output files in the CyVerse Datastore'

The different app outputs and vizualizations are detailed in step 1.

DATASET

mock_communities1_to_4

EXPECTED RESULT



Running Centrifuge on paired-end reads

3 *his protocol section uses illumina paired-end mock communities. These mock communities are artificially generated illumina reads (1 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 centrifuge-1.0.4u1.

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

- **Index :** 'Bacteria, Archea, Viruses and human (compressed)'
- **File type** : 'Read Fastq'

NOTE: *In order to match the two paired end files, the names of the submitted files should be in the following format : MYFILENAME_1.fast(a/q) and MYFILENAME_2.fast(a/q).*

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

The different app outputs and vizualizations are detailed in step 1.

DATASET

Mock communities - illumina

EXPECTED RESULT