Bioinformatic analysis of biomarker genes using metagenomic shotgun sequence datasets

Version 2

Forked from a private protocol

Fernando Espinola[1]

[1]CESIMAR-CENPAT-CONICET

dx.doi.org/10.17504/protocols.io.u32eyqe

Fernando Espinola

ABSTRACT

This protocol describes the steps used to analyze shotgun sequences deposited in in Integrated Microbial Genomes and Microbiomes (IMG/M) by using functional anotation evidences. KEGG orthology terms and pathways are used as evidence.The abundance of each of the corresponding biomarker genes is estimated by calculating the proportion of amino acid sequences assigned to the KO of interest (estimated gene copies, assembled and unassembled metagenomes, as retrieved from IMG/M) and normalized with respect to the total number of sequences assigned to KOs in each metagenome.

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

Find metagenomes in the database

1

DATASET

Integrated Microbial Genomes and Microbiomes

NOTE

To select metagenome(s) of interest, click on the following:
Find Genomes -> Genome Search -> Quick Search with "Search Parameters: Search by ID ", and paste ID(s) in the corresponding box, separated by commas.
In our case, we searched by IDs: 3300000122, 3300000242, 3300000118, 3300000121, 3300000131, 3300000125

NOTE

Select the boxes on the retrieved genomes, and click "Add selected to Genome Cart"
With the genomes added to the genome cart it is possible to export them listed in an excel file.
Once the genomes have been selected and added to the genome cart all the tools present in the IMG website are available for different analyses.
In all cases be sure that the analyses were performed on your dataset. For that, check in the superior bar below the JGI/IMG icon where your data is depicted as "My Analysis Carts". It is important that the genomes number is in agreement with the genomes that have been selected.

Get the KO table from the database

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

2 Build a normalized table: divide each KO abundance in each metagenome by the number of sequences assigned to KOs
As the exported table from the IMG displays the KO therms in rows and samples columns in order to normalize the sequence abundance the total sequence count per sample should be obtained. For that, a sum function for total sum per column would be useful to get the sum of every KO therm count in every sample.
For normalization, each KO therm count wil be devided by the total sequence count per sample. This procedure has to be applied for every sample in all dataset.

Normalize the KO table

3

NOTE

To obtain the KO table from the database:
Go to the menu and select the Compare Genomes item and then Abundance Profiles. Finally click on Overview (all functions) item where you will determinate several features of the KO table settings different options.
Toggle the option Matrix and then for normalization method chose the Genome count (estimated gene copies).
As functional profile of the dataset could be represents using different database for functional evidence, to get KO table based based on KEGG database, set the KO option for the Function item.
The source for searching will be the metagenomes previously selected. Prior to consider the total secuence annoted as metagenomes those sequences included in the assembled as the unassembled fraction must be selected.For that purpose in the Genome option set both (assembled and unassembled) and all finished, permanent draft and draft for MER-FS Metagenome and Sequencing status items respectively.
Finally press Go. A new screen will be appeared where the option for download the KO table will be shown. Finally select for Download tab-delimted file for Excel

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited