 doi:10.20944/preprints201812.0184.v1

Method Article

***Title:** HoSeIn: a workflow for integrating various homology search results from a high-throughput sequence dataset
***Authors:** Christina B. McCarthy1,2*, Gastón Rozadilla1 and Jorgelina Moreiras Clemente1
***Affiliations:**
1 Centro Regional de Estudios Genómicos, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.
2 Departamento de Informática y Tecnología, Universidad Nacional del Noroeste de la Provincia de Buenos Aires, Pergamino, Buenos Aires, Argentina.
***Contact email:** mccarthychristina@gmail.com (corresponding author*)
***Keywords:** Metagenomics; Metatranscriptomics; Environmental sample; Homology searches; Taxonomic profile; Functional profile

ABSTRACT

Data generated by metagenomic and metatranscriptomic experiments is both enormous and inherently noisy [1]. When using taxonomy-dependent alignment-based methods to classify and label reads, such as MEGAN [2], the first step consists in performing homology searches against sequence databases. To obtain the most information from the samples, nucleotide sequences are usually compared to various databases (i.e., nucleotide and protein) using local sequence aligners such as BLASTN and BLASTX [3]. Nevertheless, the analysis and integration of these results can be problematic because the outputs from these searches usually show differences, which can be notorious when working with RNA-seq (Personal observation; Graphical abstract). These inconsistencies led us to develop the HoSeIn workflow to determine the unequivocal taxonomic and functional profile of environmental samples, based on the assumption that the sequences that correspond to a certain taxon are composed of (Graphical abstract):
1) sequences that were assigned to the same taxon by both homology searches, plus
2) sequences that were assigned to that taxon by one of the homology searches but returned no hits in the other one, and vice versa.

SPECIFICATIONS TABLE

Subject Area *Agricultural and Biological Sciences*	*Select one of the following subject areas:* • *Agricultural and Biological Sciences* • *Biochemistry, Genetics and Molecular Biology* • *Chemical Engineering* • *Chemistry* • *Computer Science* • *Earth and Planetary Sciences* • *Energy* • *Engineering* • *Environmental Science* • *Immunology and Microbiology* • *Materials Science* • *Mathematics* • *Medicine and Dentistry* • *Neuroscience* • *Pharmacology, Toxicology and Pharmaceutical Science* • *Physics and Astronomy* • *Psychology* • *Social Sciences* • *Veterinary Science and Veterinary Medicine*
More specific subject area: *Metagenomics, Metatranscriptomics*	*Describe narrower subject area*
Method name: HoSeIn: Homology Search Integration	*Please specify a name of the method that you customized. The method name should be a word or short phrase to describe the methods used in your paper*
Name and reference of original method	*If applicable, include full bibliographic details of the main reference(s) describing the original method from which the new method was derived.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

Resource availability	*If applicable, include links to resources necessary to reproduce the method (e.g. data, software, hardware, reagent)*
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ https://ab.inf.uni-tuebingen.de/software/megan6 https://www.sqlite.org/	

***Method details**

Note: This protocol describes the global procedure for analysing high-throughput metatranscriptomic and metagenomic sequences from an environmental sample, and focuses in detail on how to define its unequivocal taxonomic and functional profile.
It does not include a detailed description of the pre-processing of high-throughput sequences obtained from an environmental sample (for this, see [4][5]), nor on how to perform the homology searches (for this, see [5]), as this will vary according to the Next Generation Sequencing (NGS) method used; neither does it include a detailed description of how to use MEGAN (for this, see [2,6] and the MEGAN user manual).

Overview of the HoSeIn workflow:
Trimmed sequences from an environmental sample must first be compared to the chosen nucleotide (usually nt; non-redundant) and protein (usually nr; non-redundant) databases using BLASTN [3] and BLASTX [7], respectively. BLASTX [7] searches protein databases using a translated nucleotide sequence. MEGAN (MEtaGenome ANalyzer) [2,6] is a stand-alone interactive tool that uses homology search results as input, to analyse the taxonomic content of metagenomic and metatranscriptomic datasets, as well as their functional profile using SEED [8], COG [9] and KEGG [10].
The graphical abstract shows an overview of the rationale supporting the sequence analysis workflow that we developed, and summarises the different steps that were followed:
1) Sequences were submitted to homology searches against nucleotide and protein databases using BLASTN (BN) and BLASTX (BX), respectively. These homology search results were then analysed with MEGAN and, after performing a global comparison (using one of MEGAN´s features), it was clear that they showed differences and inconsistencies (Graphical abstract). For this reason, a workflow was developed to intersect the information from both analyses in order to define the unequivocal taxonomic and functional profile of the sample, based on the assumption that the sequences that correspond to a certain taxon are composed of (Graphical abstract):
a) sequences that were assigned to the same taxon by both homology searches, plus
b) sequences that were assigned to that taxon by one of the homology searches but returned no hits in the other one, and vice versa.

2) In order to be able to access, visualise, compare and combine the information from both homology searches easily, and so identify and extract the total sequences that corresponded to each taxon, the database (DB) browser for SQLite was used (https://sqlitebrowser.org/). This is an open source tool for users and developers wanting to create databases, search, and edit data, that uses a spreadsheet-like interface. For this, the taxonomic and functional information from both homology searches for all taxa was extracted individually from MEGAN (using one of its features), and then combined in the DB browser for SQLite.

3) After creating the database in the SQLite DB browser, the criss-crossing mentioned in 1) was performed in the browser, which enabled us to elucidate the inconsistencies observed in the homology search results, based on the aforementioned assumptions. In this way, the results from each homology search were visible for each sequence in a table, as well as their final taxonomic and functional assignment resulting from the data intersection. Thus was the unequivocal taxonomic and functional profile of the sample determined.

HoSeIn workflow tutorial:
What follows is a detailed tutorial in which we exemplify our HoSeIn workflow. To be able to follow the step-by-step tutorial, we provide sample data in the form of the "RMA" files generated by MEGAN6 after it processes the text files from the homology searches (*Supplementary material:* blastn_nt16sLep_total-contigs.rma6 and blastx_nr_total-contigs.rma6). The sequences that were used in the homology searches originated from total RNA extracted from an insect pest gut, which was pyrosequenced [11] and later assembled into contigs. A few remarks follow to explain certain features that are distinctive of this particular sample: The purpose of the mentioned study was to integrate gene expression data from *Spodoptera frugiperda* guts and their associated metatranscriptomes. For this, total RNA was extracted from fifth instar larval guts and submitted to a one-step reverse transcription and PCR sequence-independent amplification procedure, and then pyrosequenced [11]; the high-throughput reads were later assembled into contigs. As we were interested in identifying and differentiating the host (*S. frugiperda*) gut transcriptome as well as its associated metatranscriptome, the following NCBI databases were downloaded locally for the homology searches (ftp://ftp.ncbi.nlm.nih.gov/blast/db/):

i) Nucleotide:
- “Non-redundant” nucleotide sequence (nt),
- 16S microbial (16S),
- Lepidopteran whole genome shotgun (Lep) projects completed at the time of the analysis.

Sequences from nt, 16S, and Lep, were then combined in a single database (DB:nt16SLep) using the appropriate BLAST+ applications (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

ii) Protein:
- non-redundant protein sequence (nr).

Total contigs were then compared to the combined nucleotide database (db: nt16SLep) using BLASTN, and to the protein database (nr) using BLASTX. The Lep sequences of the combined nucleotide database simplified the identification of sequences that corresponded to the host (which represented the vast majority). The nt and 16S databases in the combined nucleotide database (db: nt16SLep) enabled the identification of the associated metatranscriptome.
In this tutorial we only show how we determined the taxonomical identity of these contigs (i.e., we do not include the identification of the functional profile as well). As mentioned previously, we provide sample data in the form of "RMA" files (*Supplementary material:* blastn_nt16sLep_total-contigs.rma6 and blastx_nr_total-contigs.rma6) to be able to follow the step-by-step tutorial from the point in which the taxonomical information is extracted from MEGAN.

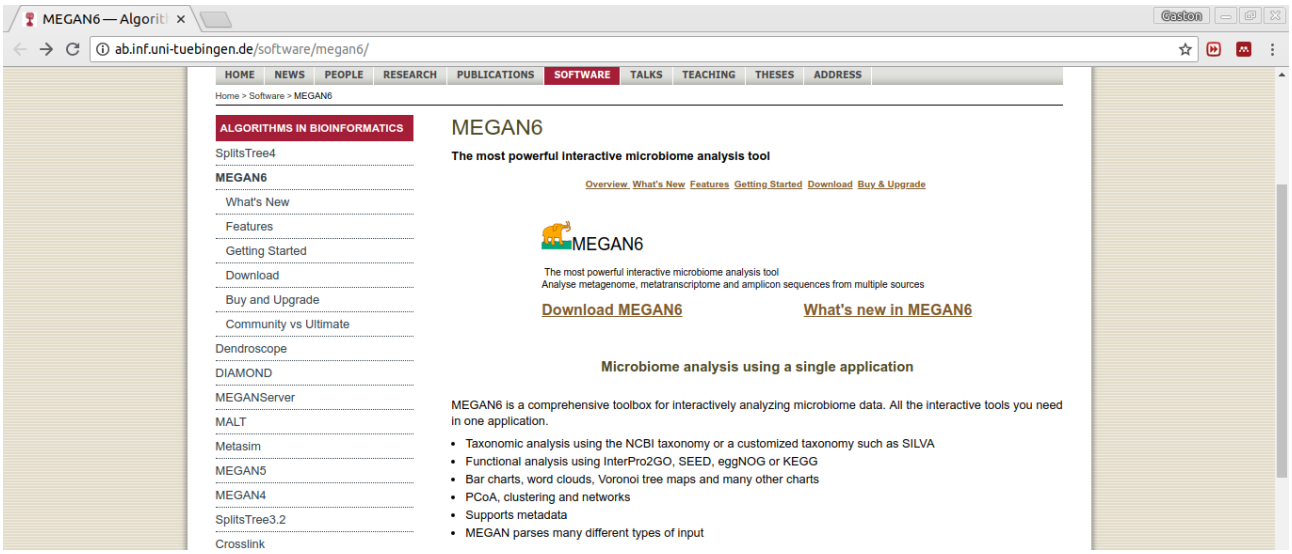
I) How to download MEGAN:

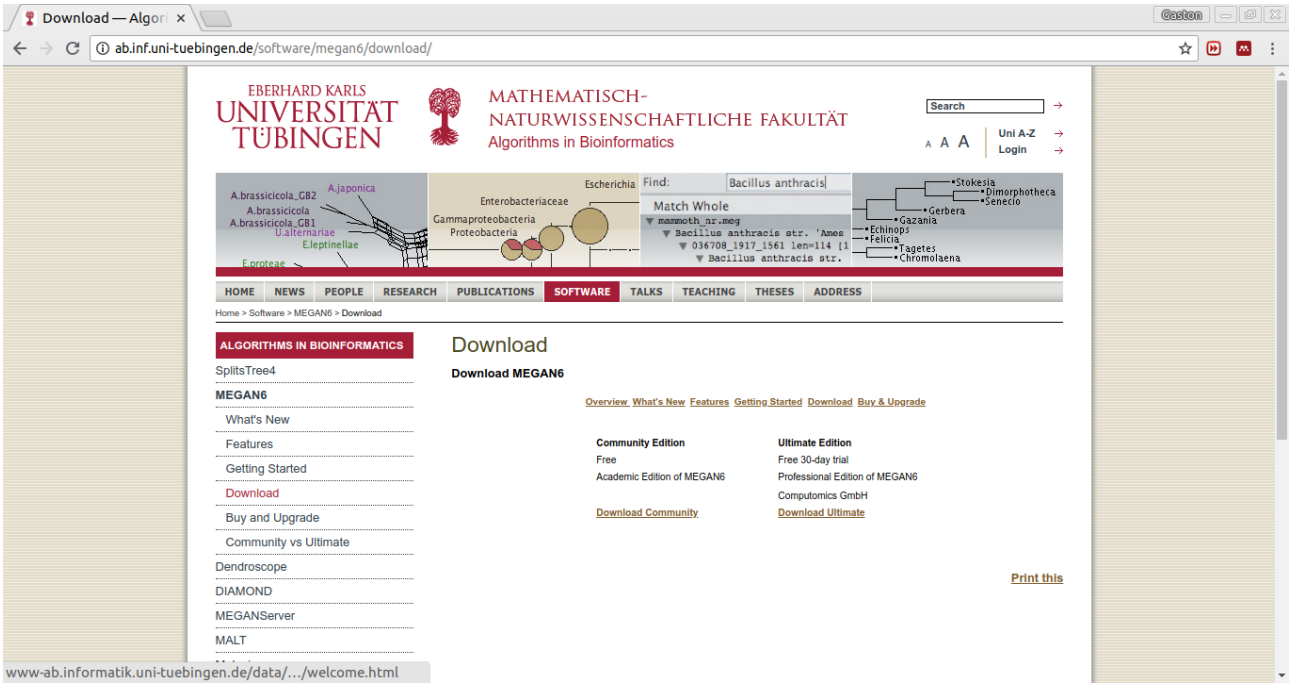
Summary:

MEGAN must be downloaded and installed locally to be able to process the text files from the homology searches and then extract the taxonomical and functional information.

Step by step tutorial:

MEGAN installation:





Download the MEGAN6 version that matches your operating system, and the mapping files (Protein accession to NCBI-taxonomy mapping file, and Nucleotide accession to NCBI-taxonomy mapping file):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

MEGAN6-download ×

ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html

MEGAN6 Download Page

This is MEGAN6, a new major update of MEGAN. This is the official download site for the most recent version. The Community Edition of the paper is described here: Huson et al., (2016), PLoS Computational Biology
Please post questions and bug reports to the community website.

MEGAN6 Community Edition installers:

MEGAN_Community_macos_6_12_0.dmg
MEGAN_Community_unix_6_12_0.sh
MEGAN_Community_windows-x64_6_12_0.exe
MEGAN_Community_windows_32bit_6_12_0.exe

MEGAN6 Ultimate Edition installers (require license key):

MEGAN_Ultimate_macos_6_12_0.dmg
MEGAN_Ultimate_unix_6_12_0.sh
MEGAN_Ultimate_windows-x64_6_12_0.exe
MEGAN_Ultimate_windows_32bit_6_12_0.exe

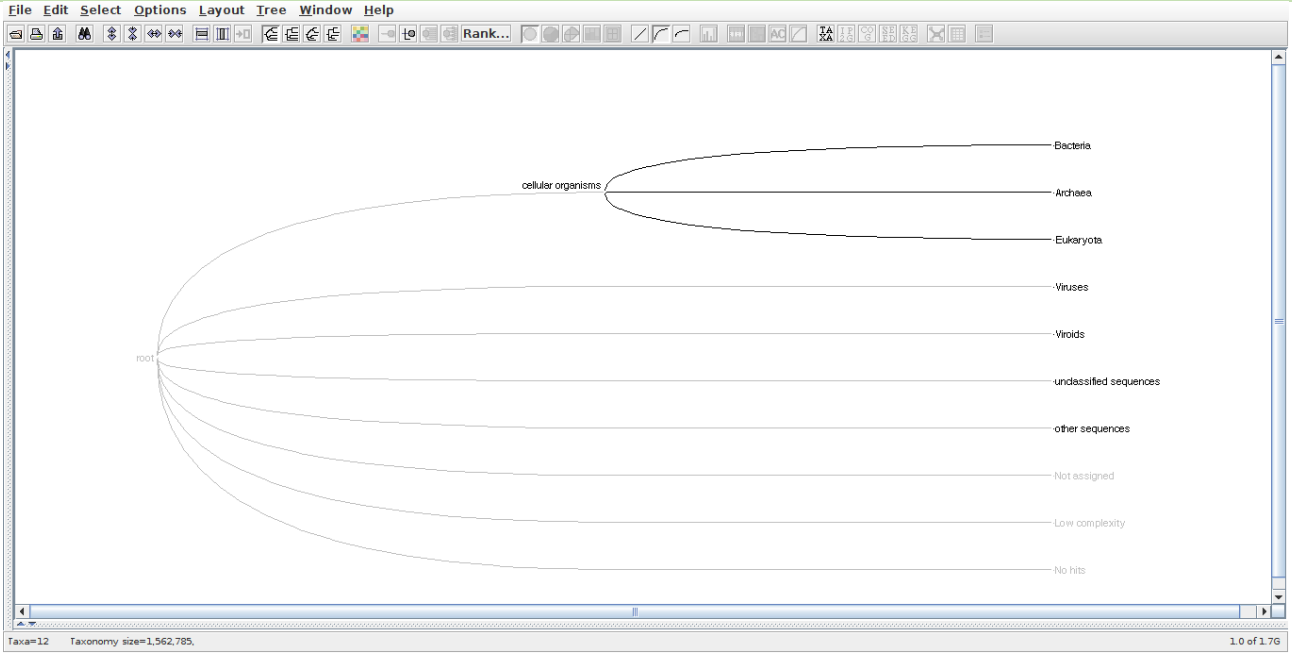
MEGAN6 Ultimate Edition is licensed by Computomics GmbH

Apply for a trial license here
New feature: The Ultimate Edition comes with the latest KEGG classification and pathway files. Computomics GmbH is a licensed reseller of KEGG.

Mapping files for current NCBI-nr protein database (not containing GI numbers):

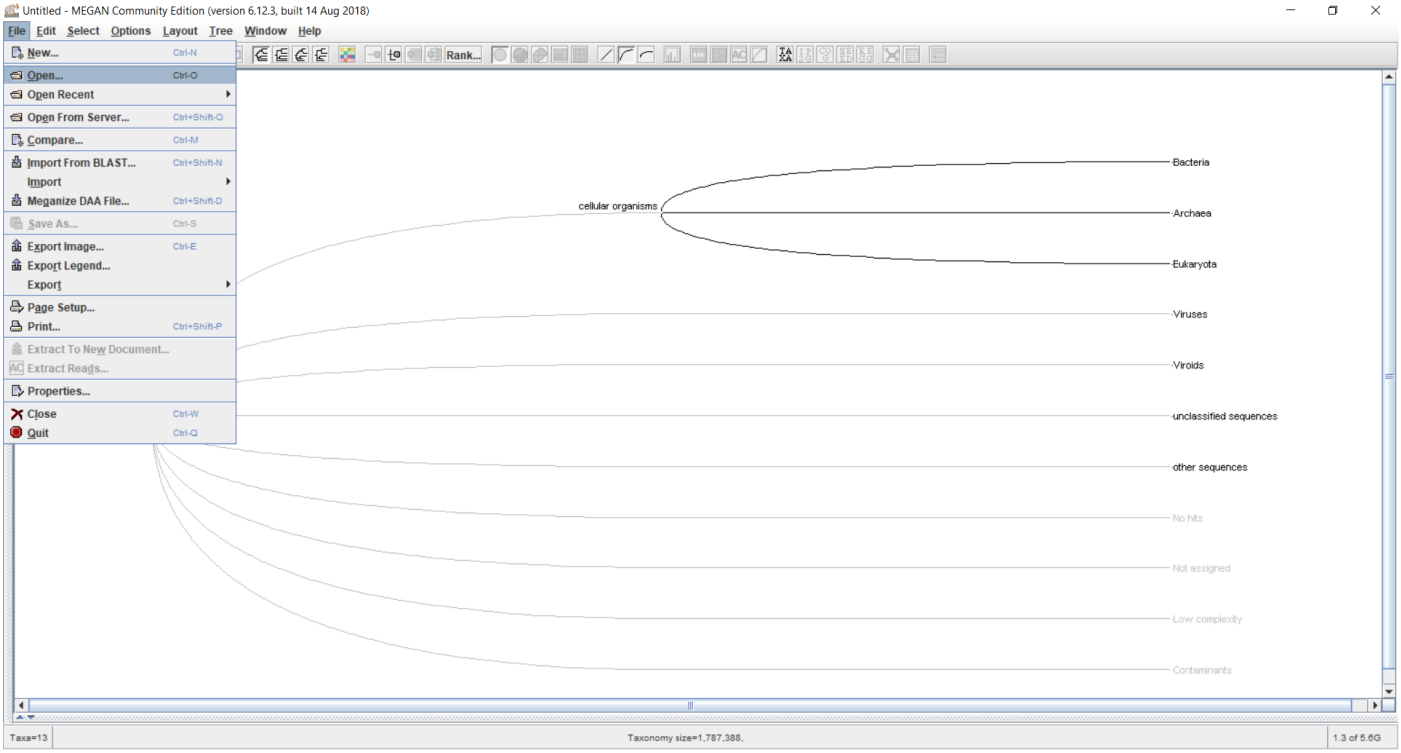
prot_acc2tax-June2018X1.abin.zip Protein accession to NCBI-taxonomy mapping file. Must be unzipped before use.
nucl_acc2tax-June2018.abin.zip Nucleotide accession to NCBI-taxonomy mapping file. Must be unzipped before use.
acc2interpro-June2018X.abin.zip Protein accession to InterPro mapping file. Must be unzipped before use.
acc2eggnog-Oct2016X.abin.zip Protein accession to eggNOG mapping file. Must be unzipped before use.
acc2seed-May2015XX.abin.zip Protein accession to SEED mapping file. Must be unzipped before use.

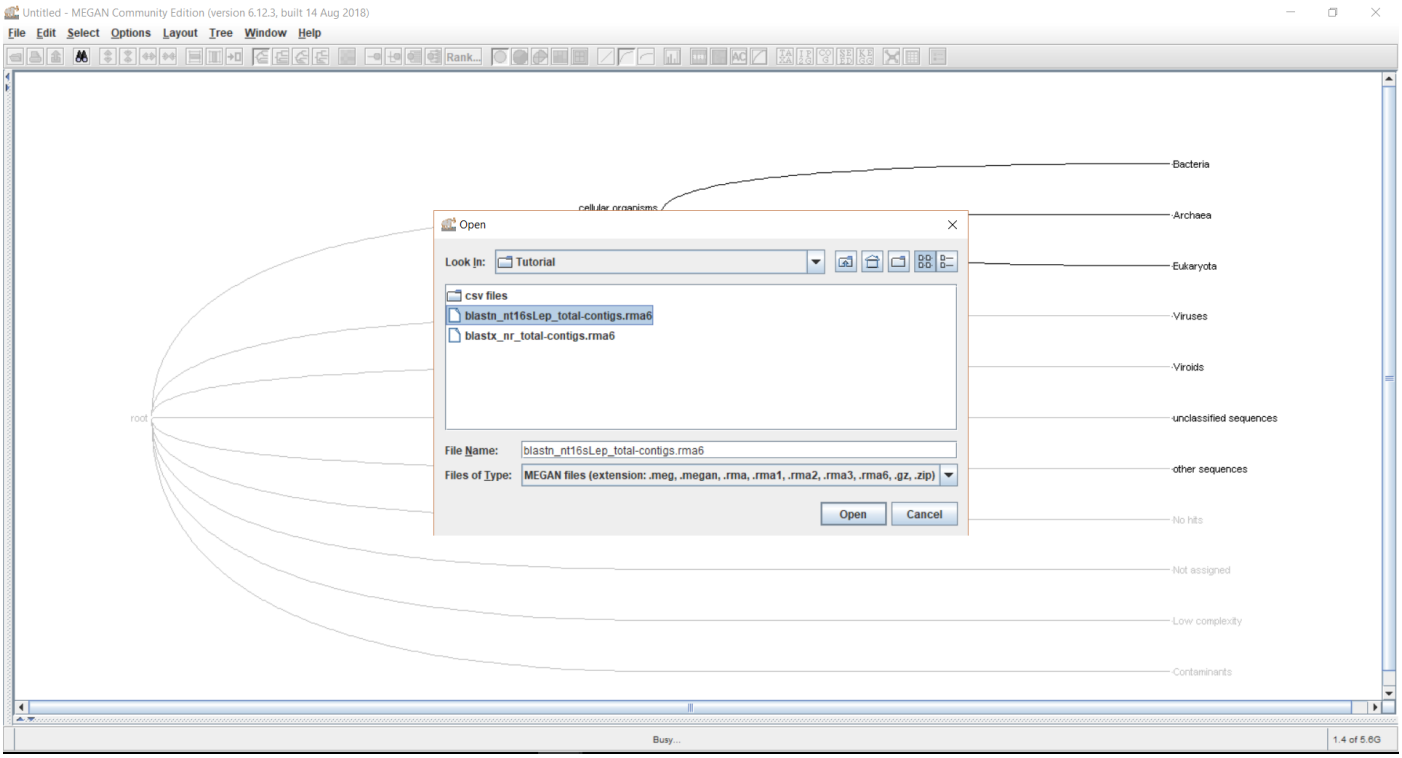
Run the installer.
This is the MEGAN6 main window:



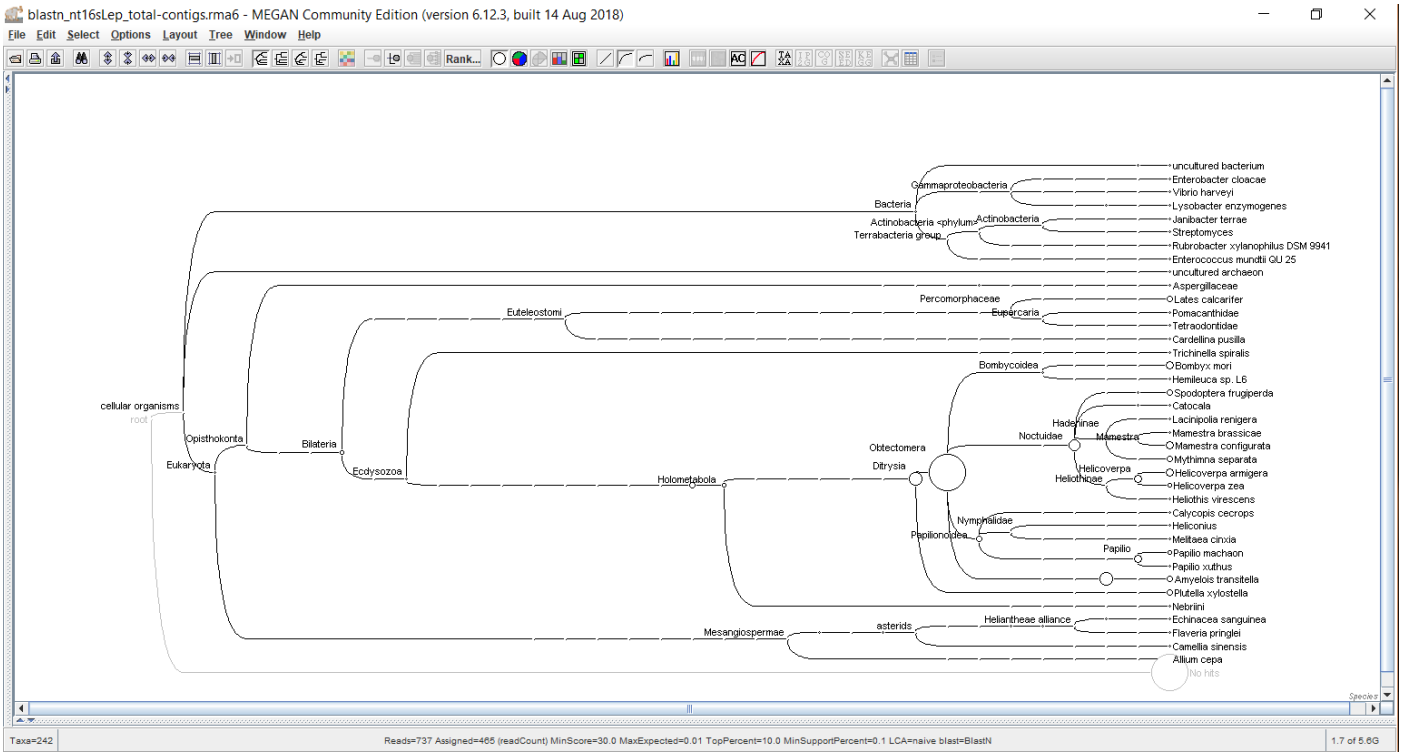
Open the provided RMA files (blastn_nt16sLep_total-contigs.rma6 and blastx_nr_total-contigs.rma6):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1





Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1



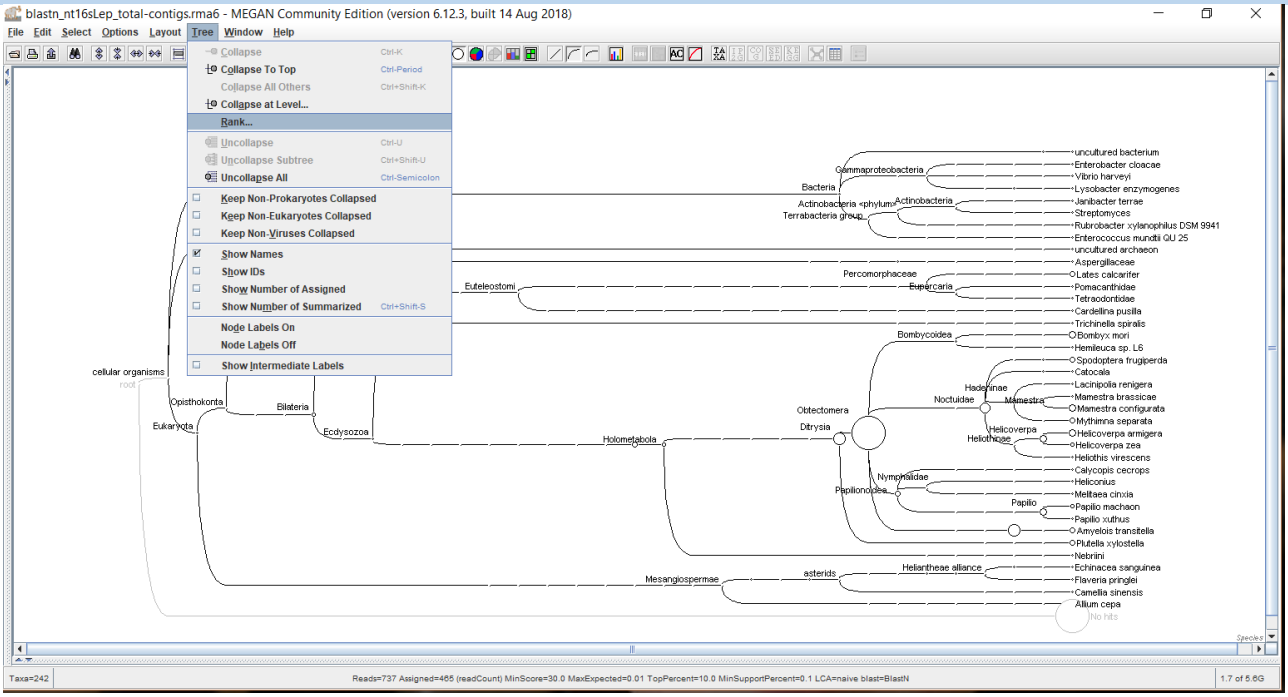
II) How to extract the information from MEGAN:

Summary:

To extract the taxonomic information from MEGAN, the taxonomic tree must be progressively expanded from Domain to species, selecting all the leaves, and extracting the text files in csv (comma-separated values) format. In each of these files there are two columns, one with the contig name (e.g., Contig479) and another with the corresponding assigned taxonomic level (Domain, Phylum, etc.). In this way, a series of files are obtained from both the BLASTN and BLASTX homology searches (blastn_domain, blastn_phylum, blastx_domain, blastx_phylum, etc.).

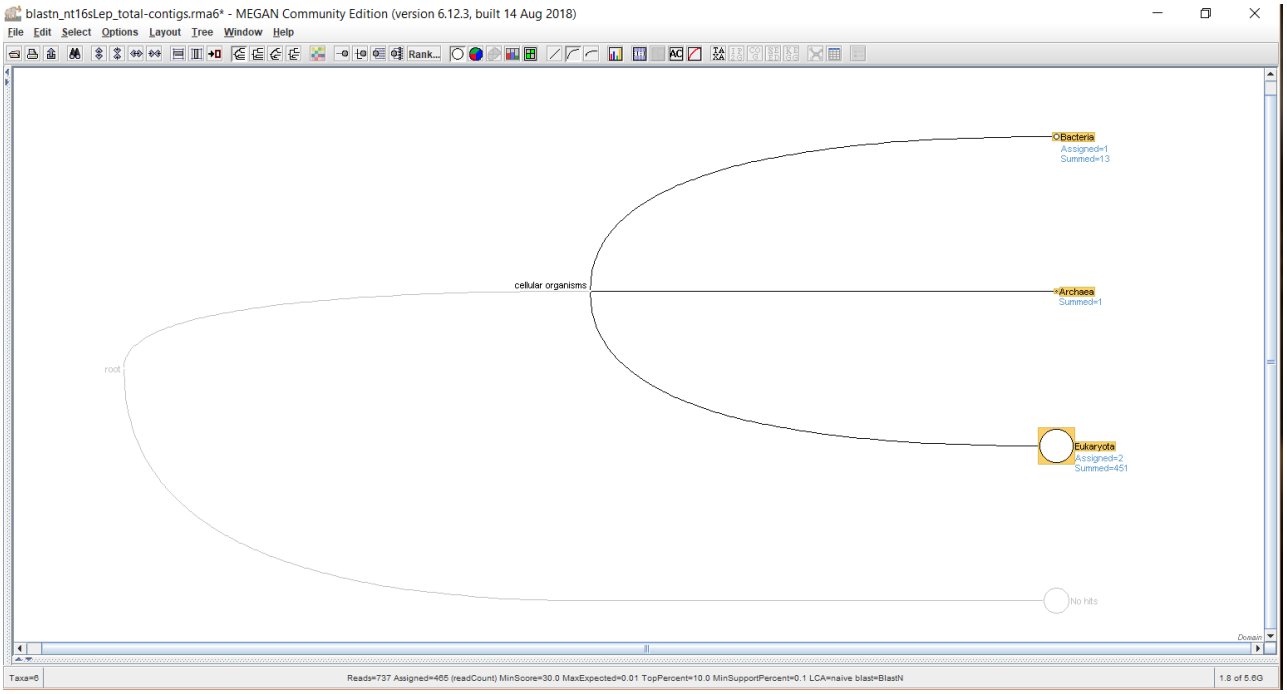
Step by step tutorial:

Choose the taxonomic level that you wish to extract (from Domain to Species):
"Tree" > "Rank"

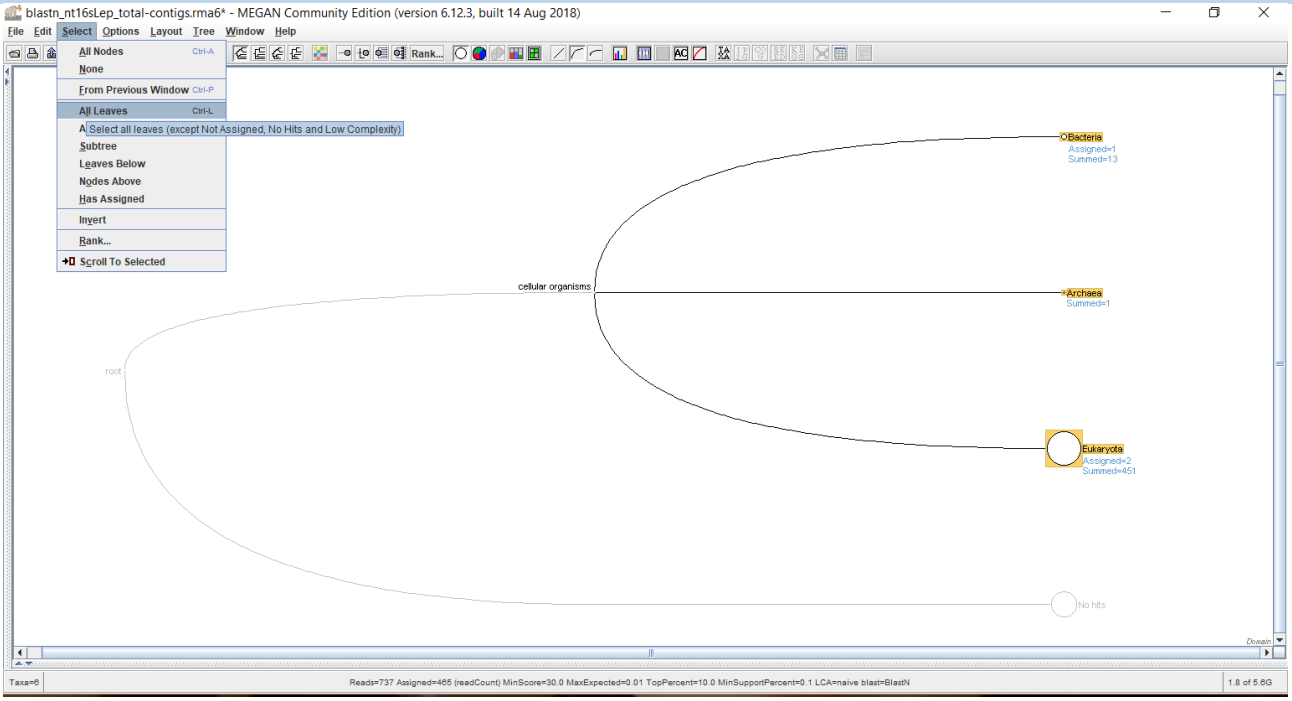


Here we show the tree collapsed to Domain level:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

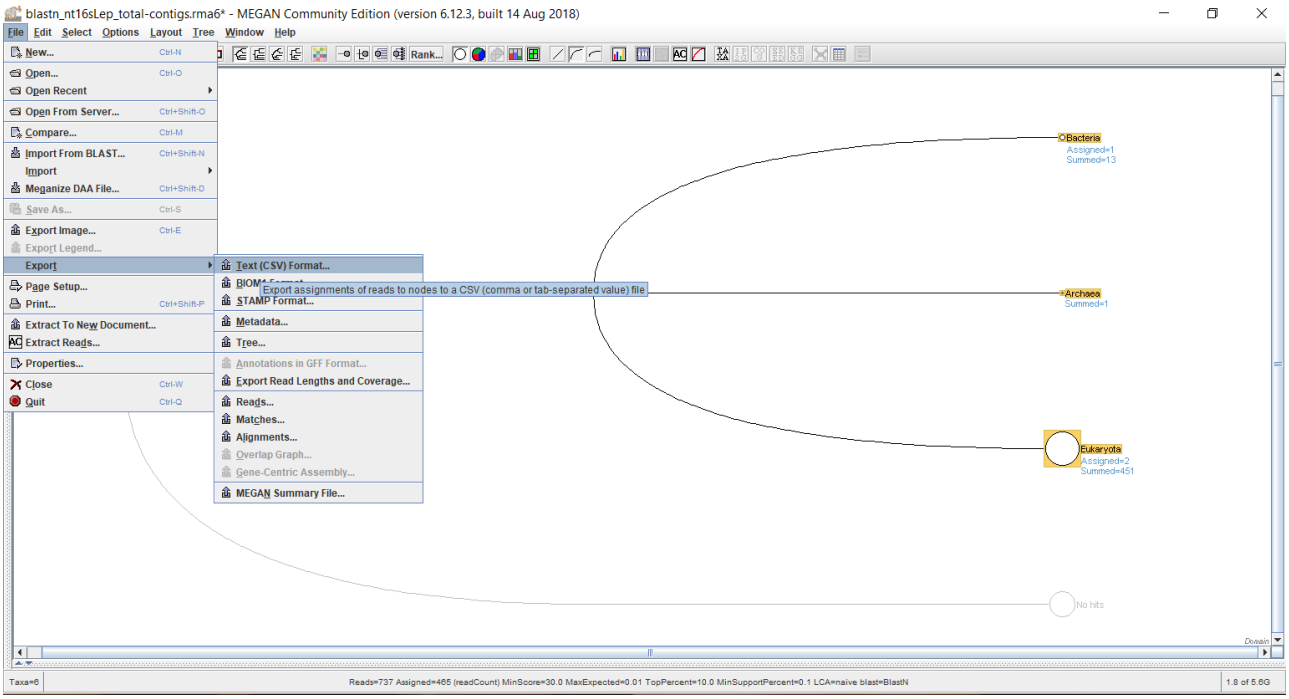


To select the leaves:
"Select" > "All Leaves"

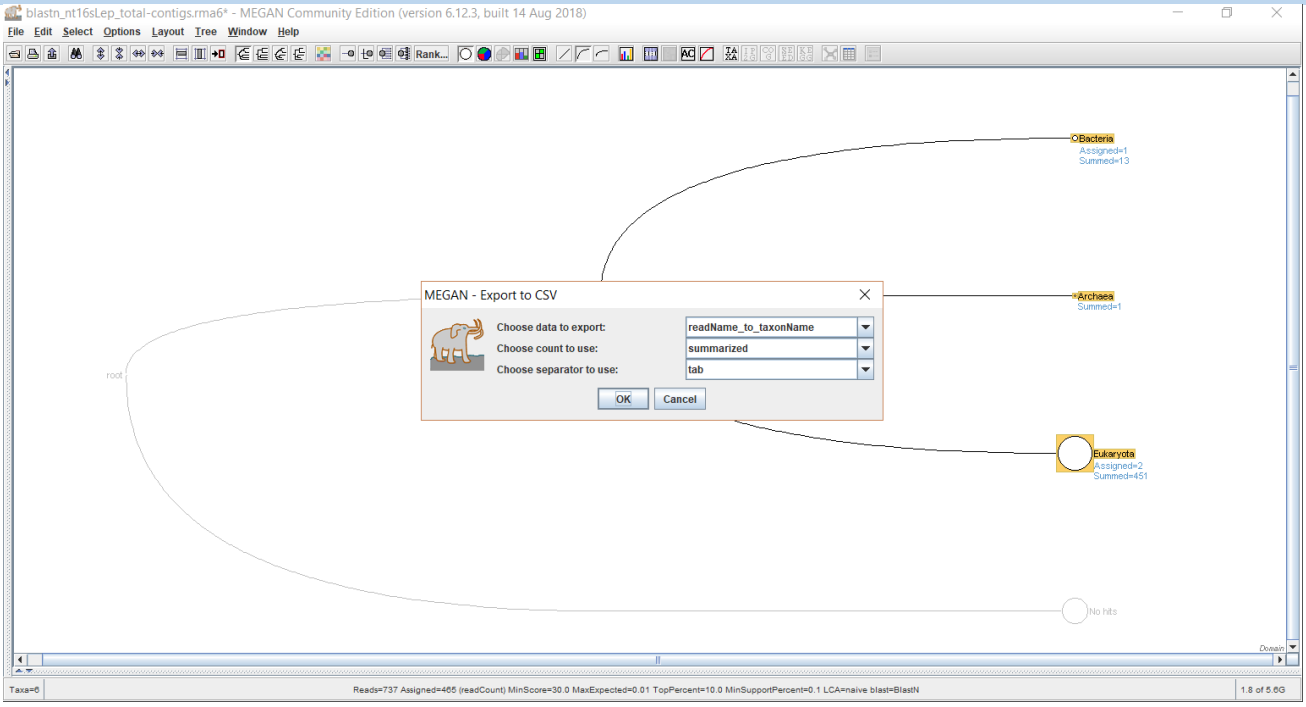


Without deselecting the leaves, export the file to csv format:
"File" > "Export" > "CSV Format"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

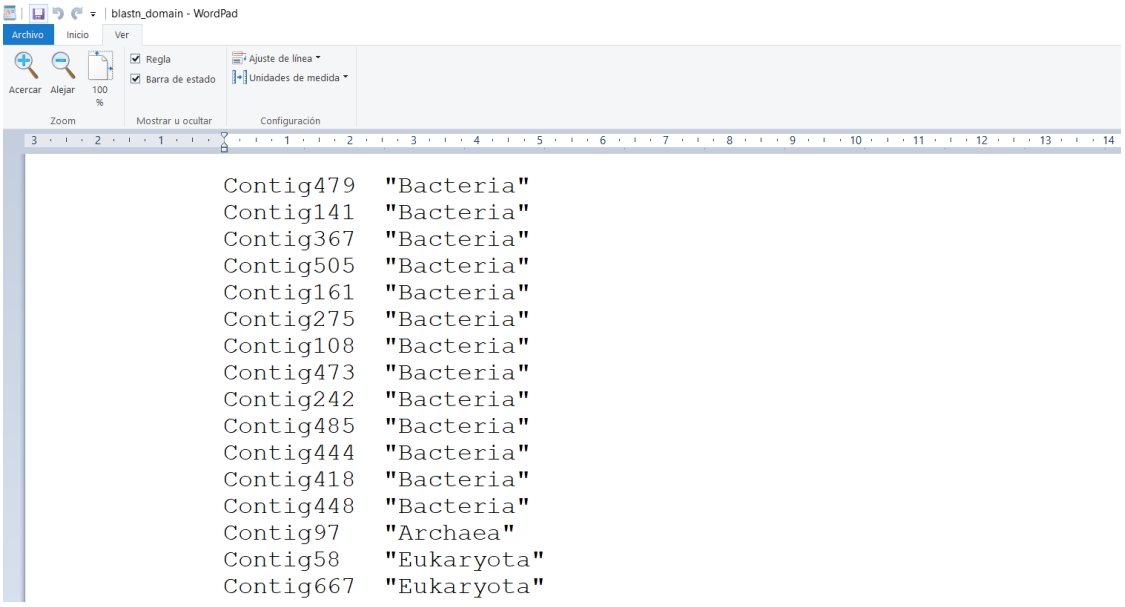


Choose what data you want to export and in what way it will be tabbed in the csv file (choose “Summarised” so it exports the sequences contained in the chosen taxonomic level as well as all the lower levels).

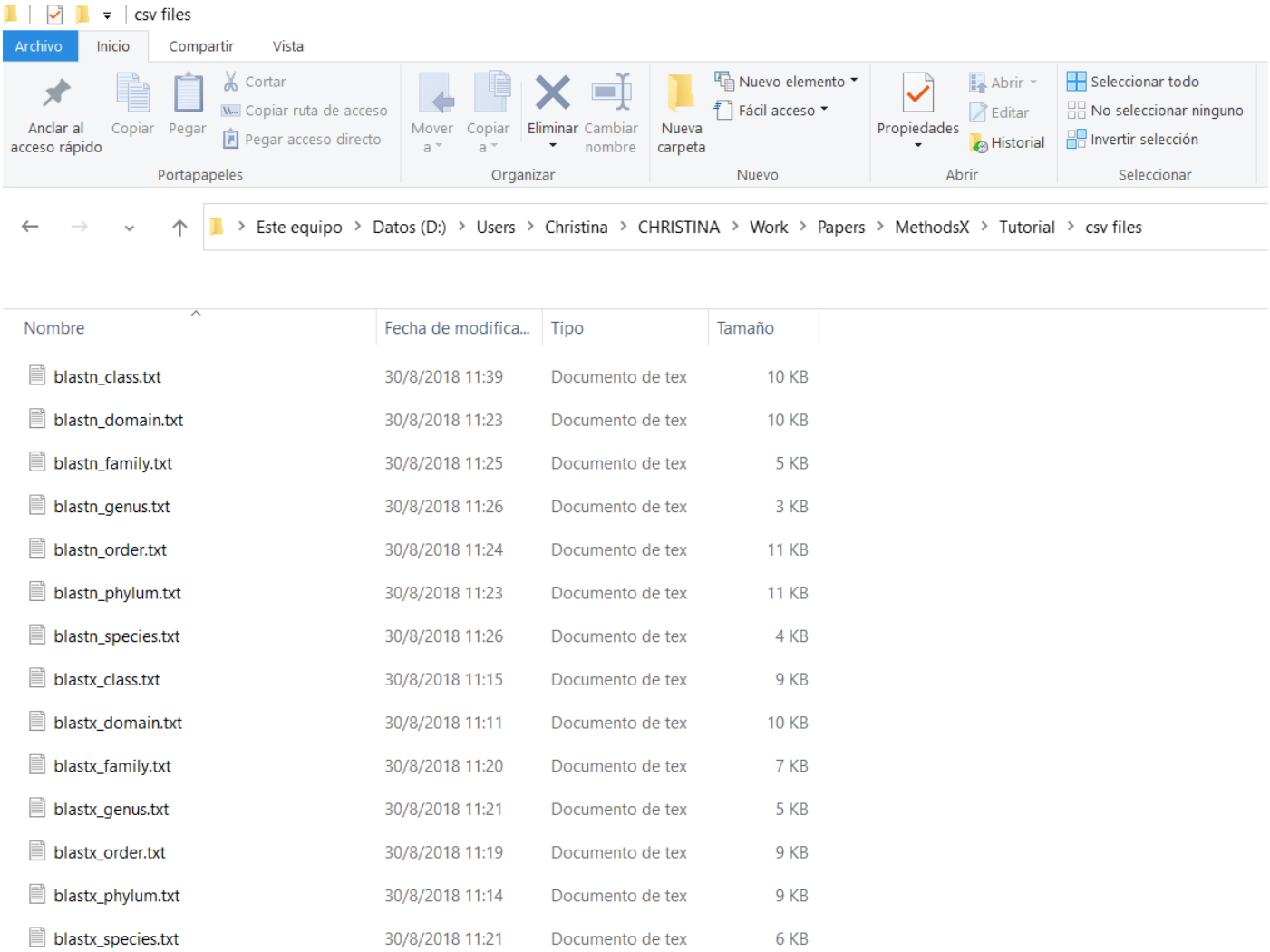


In this way a csv text file is obtained (which can be viewed in a basic word processor such as WordPad):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1



Repeat this procedure for each taxonomic level, and for the other homology search. In this way, 14 total files are obtained (7 files for each homology search), each one corresponding to a particular taxonomic level:



III) How to combine the information extracted from MEGAN to create a database in the DB browser for SQLite:

Below is a brief explanation of the DB browser for SQLite commands that were used to create the database:

- *Naming tables and columns*: the table name is separated from the column name by a period, e.g. "taxonomy.blastn-domain" indicates the "blastn-domain" column from the "taxonomy" table.
- *SELECT table1.column1, table2.column2, etc.*: indicates what columns we want to visualise, using the naming system explained in the previous paragraph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

- *FROM table1*: indicates what table to retrieve the columns from.
- *LEFT JOIN table2 ON table2.contig = table1.contig*: to join columns from various tables following a certain criterion (in this example, the contig number).
- *WHERE table1.column1 LIKE "Arthropoda"*: to filter the table following a certain criterion; in this example, to only visualise those contigs that were assigned to Arthropoda.
- *UPDATE column1 SET value1 WHERE condition1*: to update the information in column1 with value1 when a certain condition is fulfilled.

Summary:
The next step consisted in combining all the files extracted from MEGAN to create a single table which included both taxonomic assignments for each contig (the one from BLASTN and the one from BLASTX). To do this, each csv file was imported individually to the DB browser for SQLite, using the following commands:

SELECT
blastn_domain.contig,blastx_domain.blastx_domain,blastx_phylum.blastx_phylum,blastx_class.blastx_class,blastx_order.blastx_order,blastx_family.blastx_family,blastx_genus.blastx_genus,blastx_species.blastx_species,blastn_domain.blastn_domain,blastn_phylum.blastn_phylum,blastn_class.blastn_class,blastn_order.blastn_order,blastn_family.blastn_family,blastn_genus.blastn_genus,blastn_species.blastn_species FROM blastn_domain

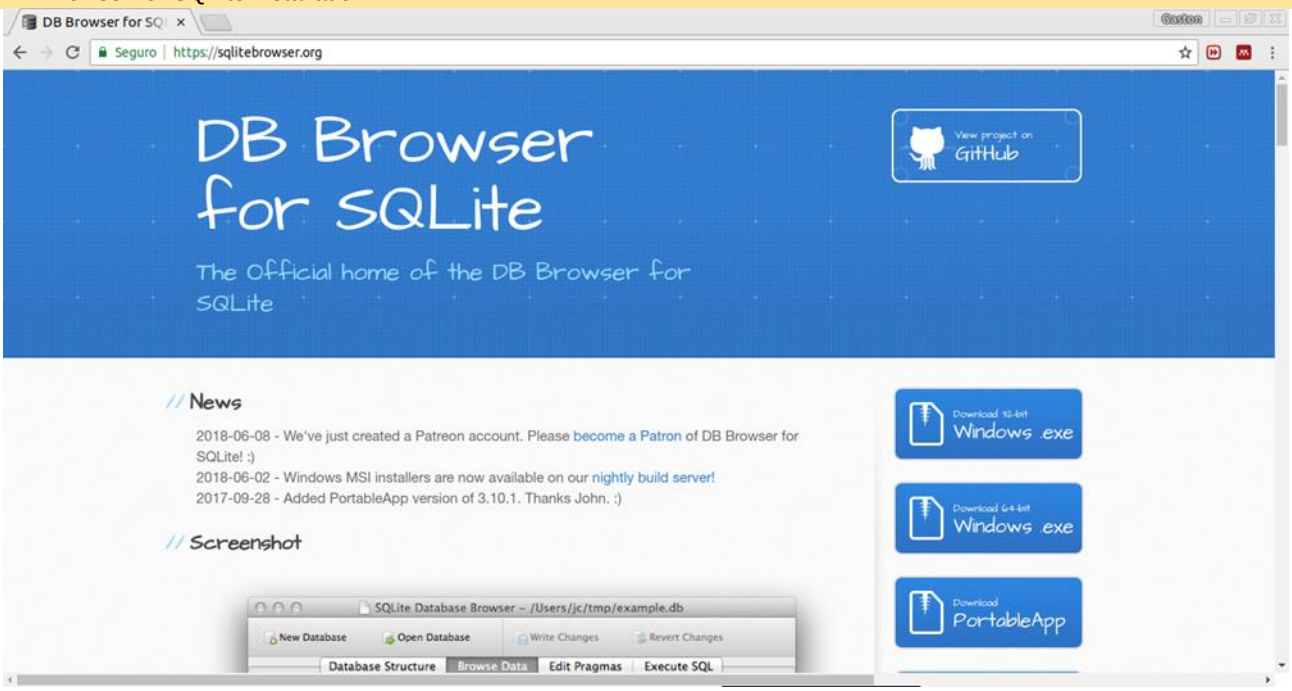
LEFT JOIN blastx_domain ON blastx_domain.contig = blastn_domain.contig
LEFT JOIN blastx_phylum ON blastx_phylum.contig = blastn_domain.contig
LEFT JOIN blastx_class ON blastx_class.contig = blastn_domain.contig
LEFT JOIN blastx_order ON blastx_order.contig = blastn_domain.contig
LEFT JOIN blastx_family ON blastx_family.contig = blastn_domain.contig
LEFT JOIN blastx_genus ON blastx_genus.contig = blastn_domain.contig
LEFT JOIN blastx_species ON blastx_species.contig = blastn_domain.contig

LEFT JOIN blastn_phylum ON blastn_phylum.contig = blastn_domain.contig
LEFT JOIN blastn_class ON blastn_class.contig = blastn_domain.contig
LEFT JOIN blastn_order ON blastn_order.contig = blastn_domain.contig
LEFT JOIN blastn_family ON blastn_family.contig = blastn_domain.contig
LEFT JOIN blastn_genus ON blastn_genus.contig = blastn_domain.contig
LEFT JOIN blastn_species ON blastn_species.contig = blastn_domain.contig

To simplify the naming of tables and columns, this result was exported in csv format and imported again to the DB browser for SQLite as the "Taxonomy" table. Various empty columns were added to this table (named final_domain, final_phylum, etc.) to be completed with the final result of the criss-crossing of the assignments, and another column "state" to indicate if the contig was assigned or not (to avoid multiple assignments).

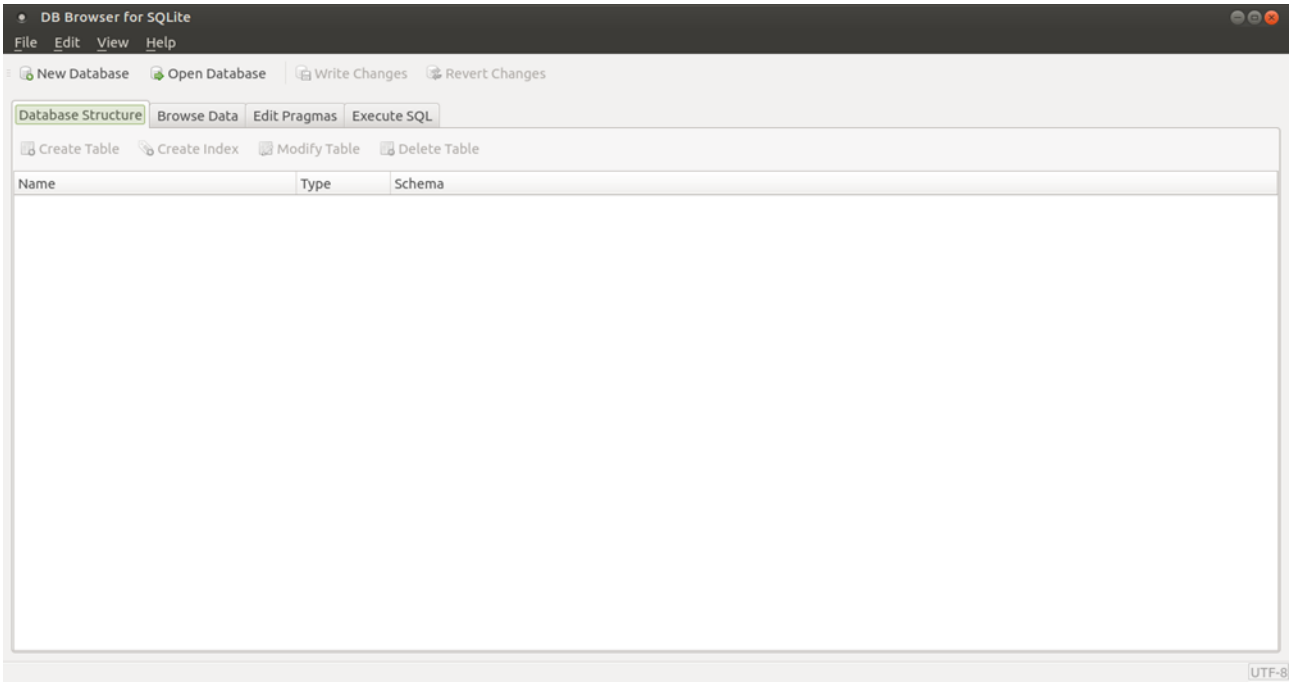
Step by step tutorial:

DB Browser for SQLite installation:

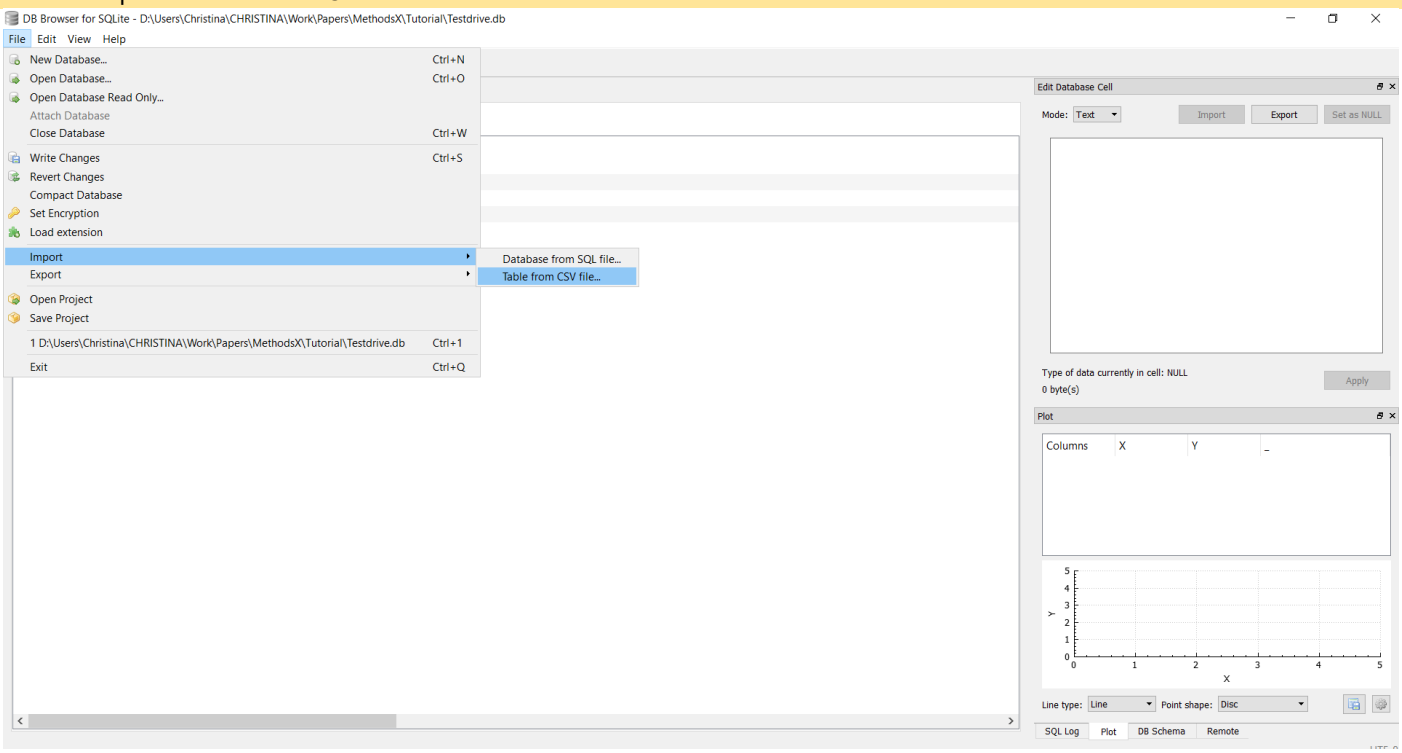


DB Browser for SQLite main window:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

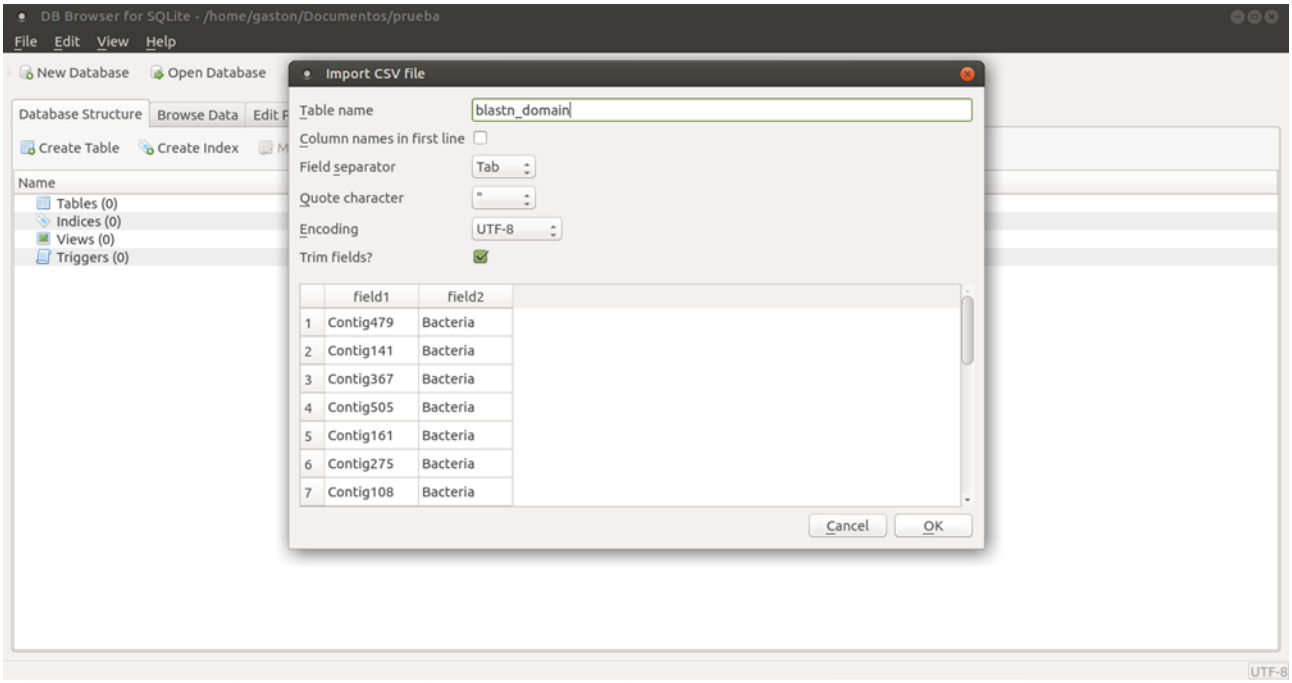


Create a new database clicking on “New Database” and choosing where to save it. Import csv files that were exported from MEGAN6:
“File” > “Import” > “Table from CSV file”

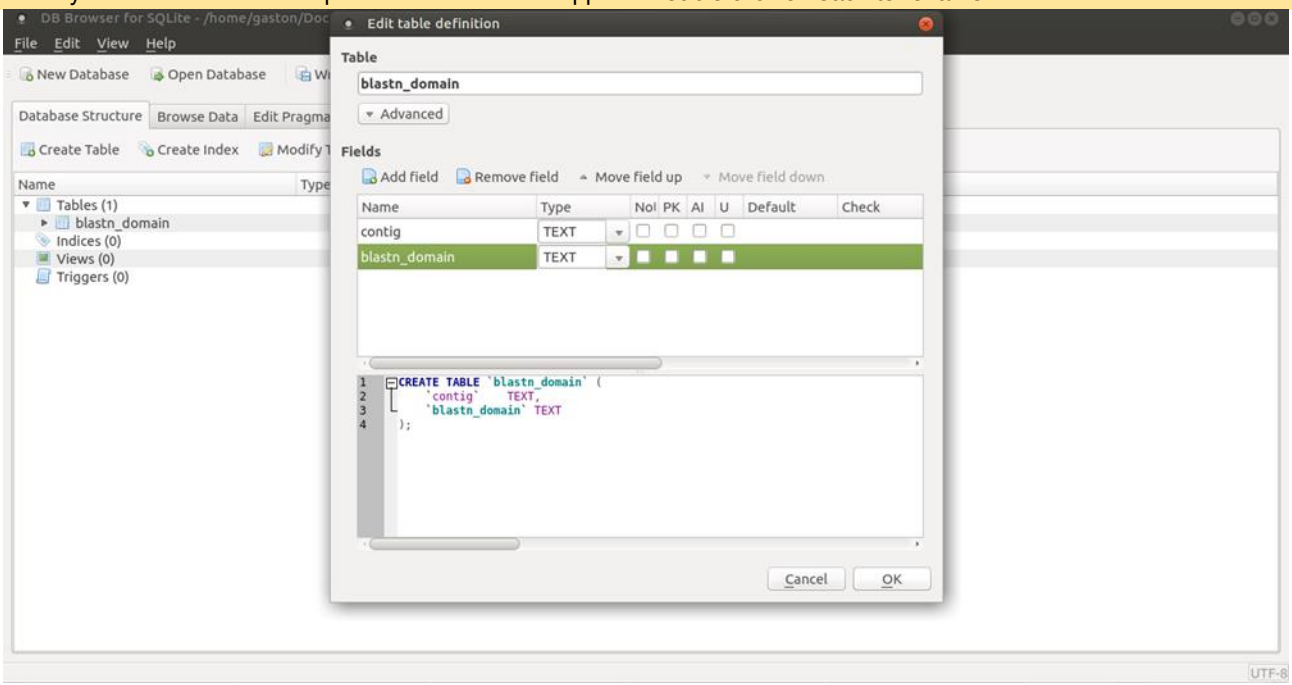


Choose a representative name for the table (in this example, blastn_domain), and indicate field separator (in this case, Tab):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

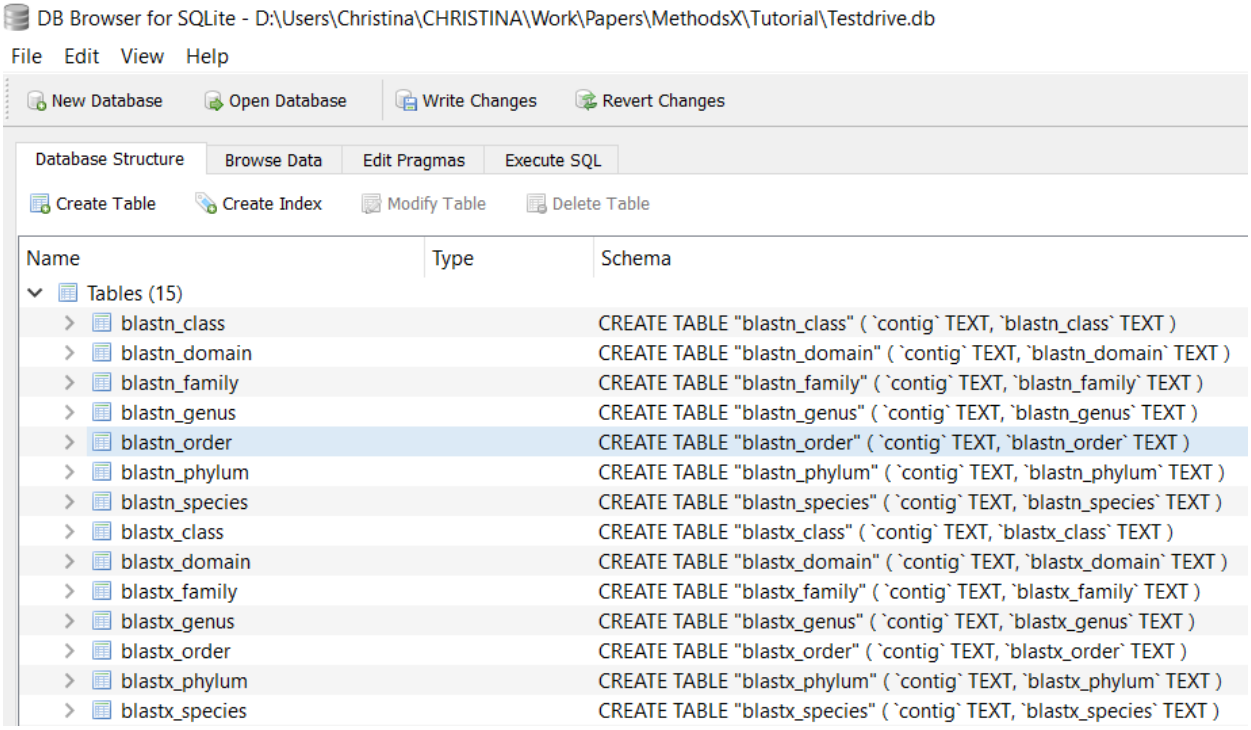


To simplify the interpretation of the commands and avoid mistakes, we recommend renaming the table columns with representative names (in this example, contig and blastn_domain). For this, in the main window select the table that will be modified and click “Modify table”. A window will open where both fields appear. Double click on each to rename:



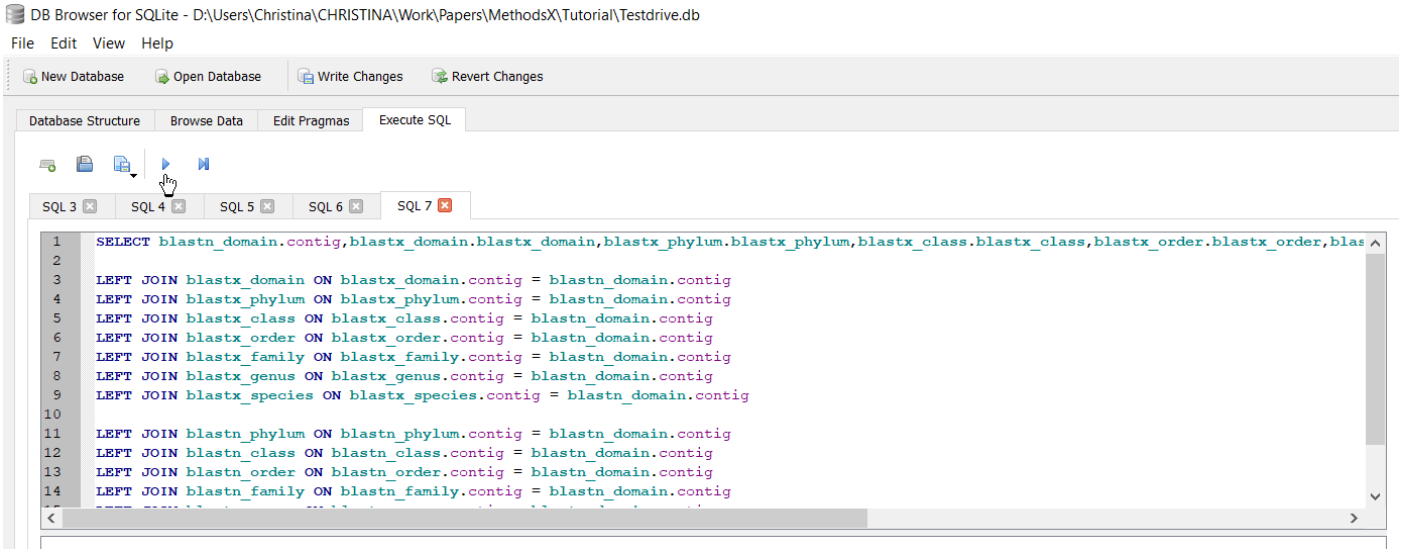
Repeat this procedure for each of the 14 csv files:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1



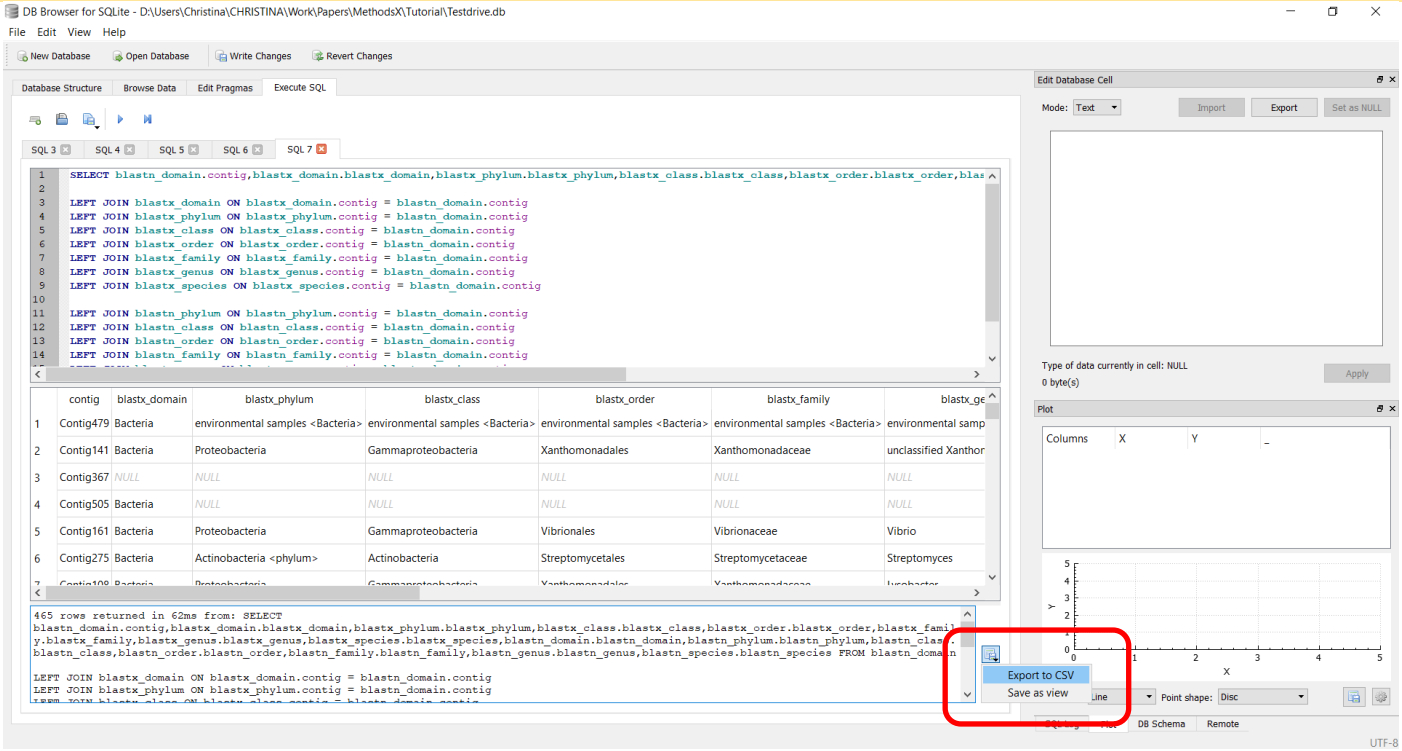
In the “Execute SQL” leaf, paste the following commands and execute them with “Play”:

```
SELECT
blastn_domain.contig,blastx_domain.blastx_domain,blastx_phylum.blastx_phylum,blastx_class.blastx_class,blastx_order.blastx_order,blastx_family.blastx_family,blastx_genus.blastx_genus,blastx_species.blastx_species,blastn_domain.blastn_domain,blastn_phylum.blastn_phylum,blastn_class.blastn_class,blastn_order.blastn_order,blastn_family.blastn_family,blastn_genus.blastn_genus,blastn_species.blastn_species FROM blastn_domain

LEFT JOIN blastx_domain ON blastx_domain.contig = blastn_domain.contig
LEFT JOIN blastx_phylum ON blastx_phylum.contig = blastn_domain.contig
LEFT JOIN blastx_class ON blastx_class.contig = blastn_domain.contig
LEFT JOIN blastx_order ON blastx_order.contig = blastn_domain.contig
LEFT JOIN blastx_family ON blastx_family.contig = blastn_domain.contig
LEFT JOIN blastx_genus ON blastx_genus.contig = blastn_domain.contig
LEFT JOIN blastx_species ON blastx_species.contig = blastn_domain.contig

LEFT JOIN blastn_phylum ON blastn_phylum.contig = blastn_domain.contig
LEFT JOIN blastn_class ON blastn_class.contig = blastn_domain.contig
LEFT JOIN blastn_order ON blastn_order.contig = blastn_domain.contig
LEFT JOIN blastn_family ON blastn_family.contig = blastn_domain.contig
LEFT JOIN blastn_genus ON blastn_genus.contig = blastn_domain.contig
LEFT JOIN blastn_species ON blastn_species.contig = blastn_domain.contig
```

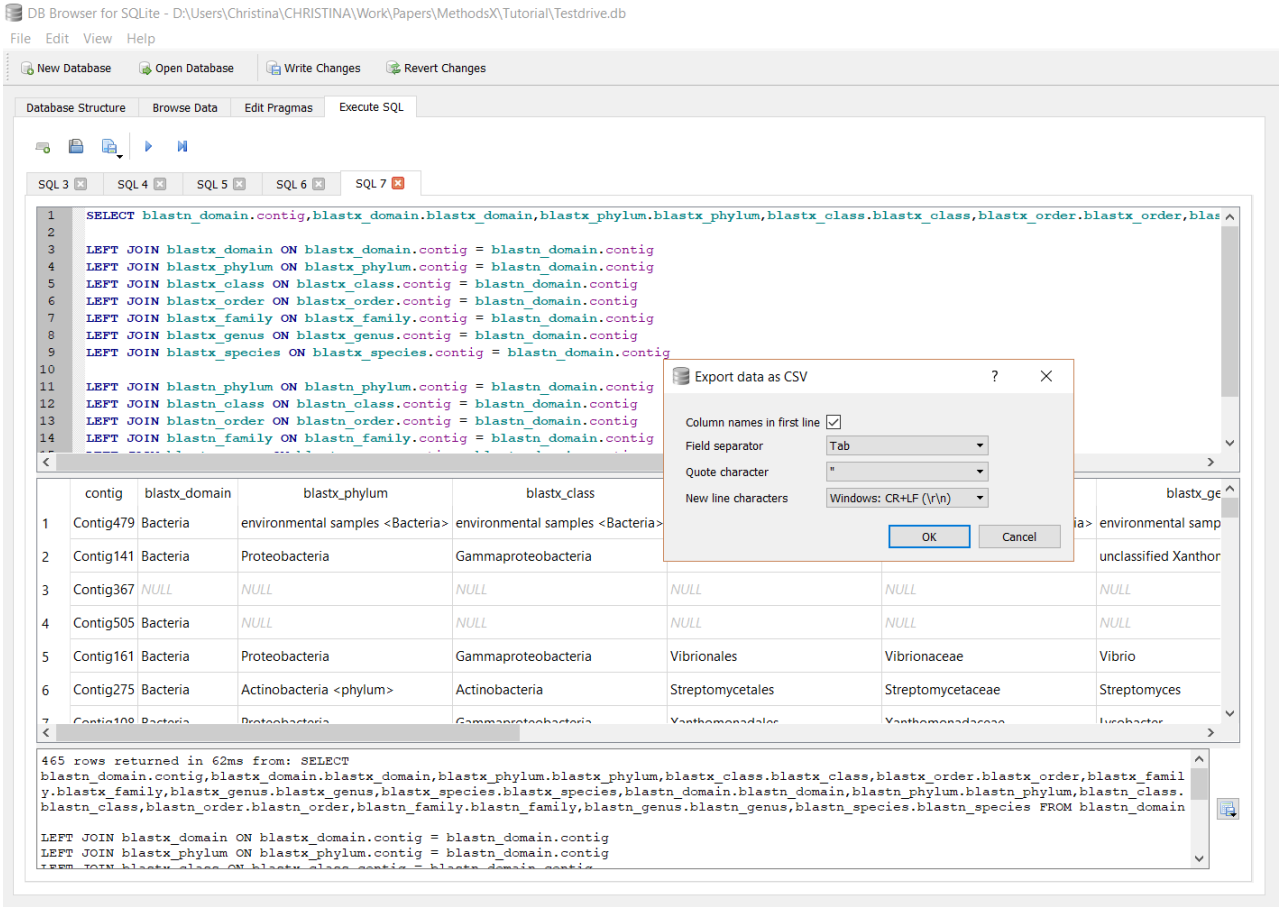


Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

Once it has been executed, the result will be shown in the bottom part of the window. Now this result can be saved in the database as “View”, or it can be exported in csv format to be analysed with other programs. To do this, we click on the save button and choose the desired option:

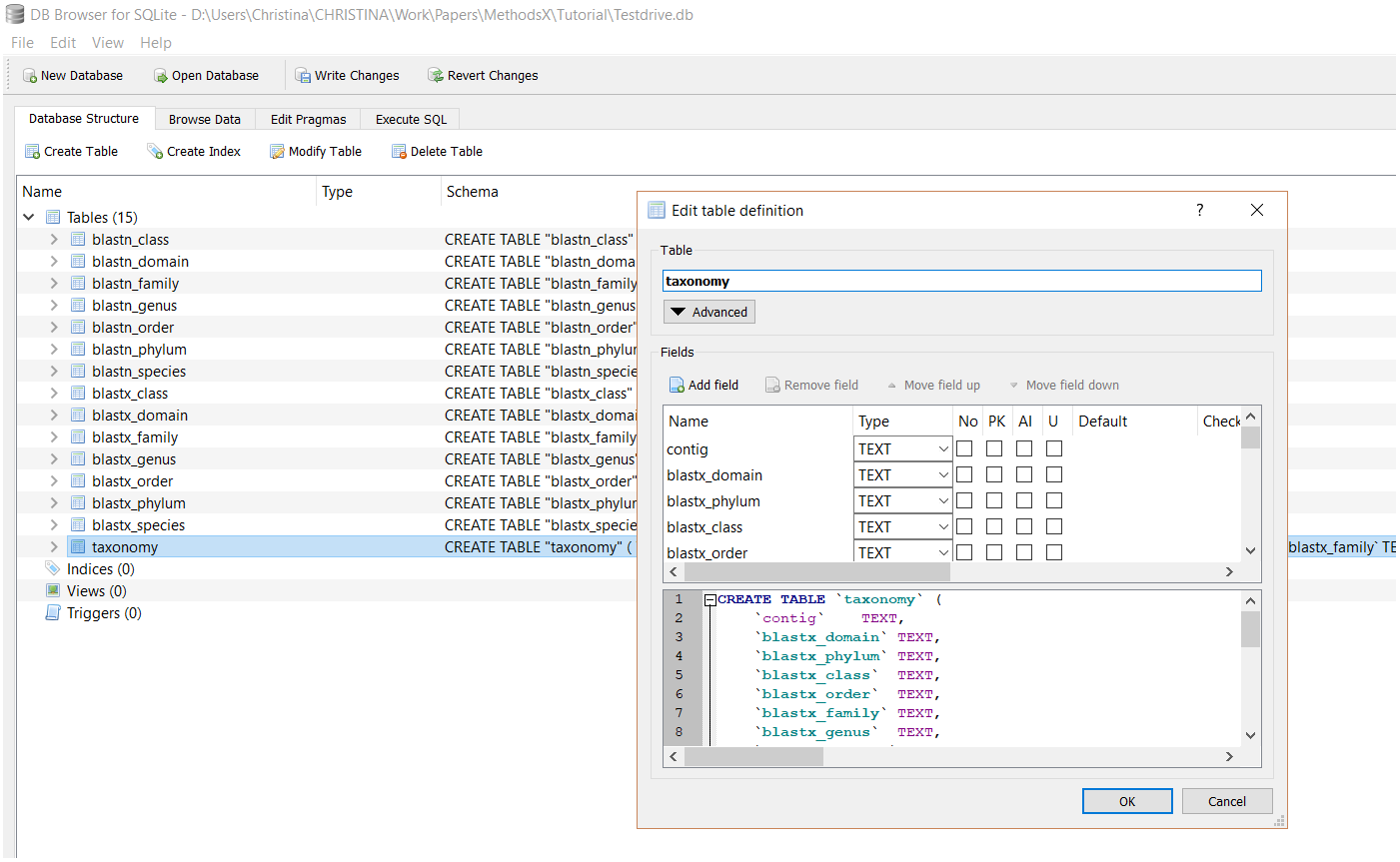


In this example, “Export to CSV” to obtain the complete table:

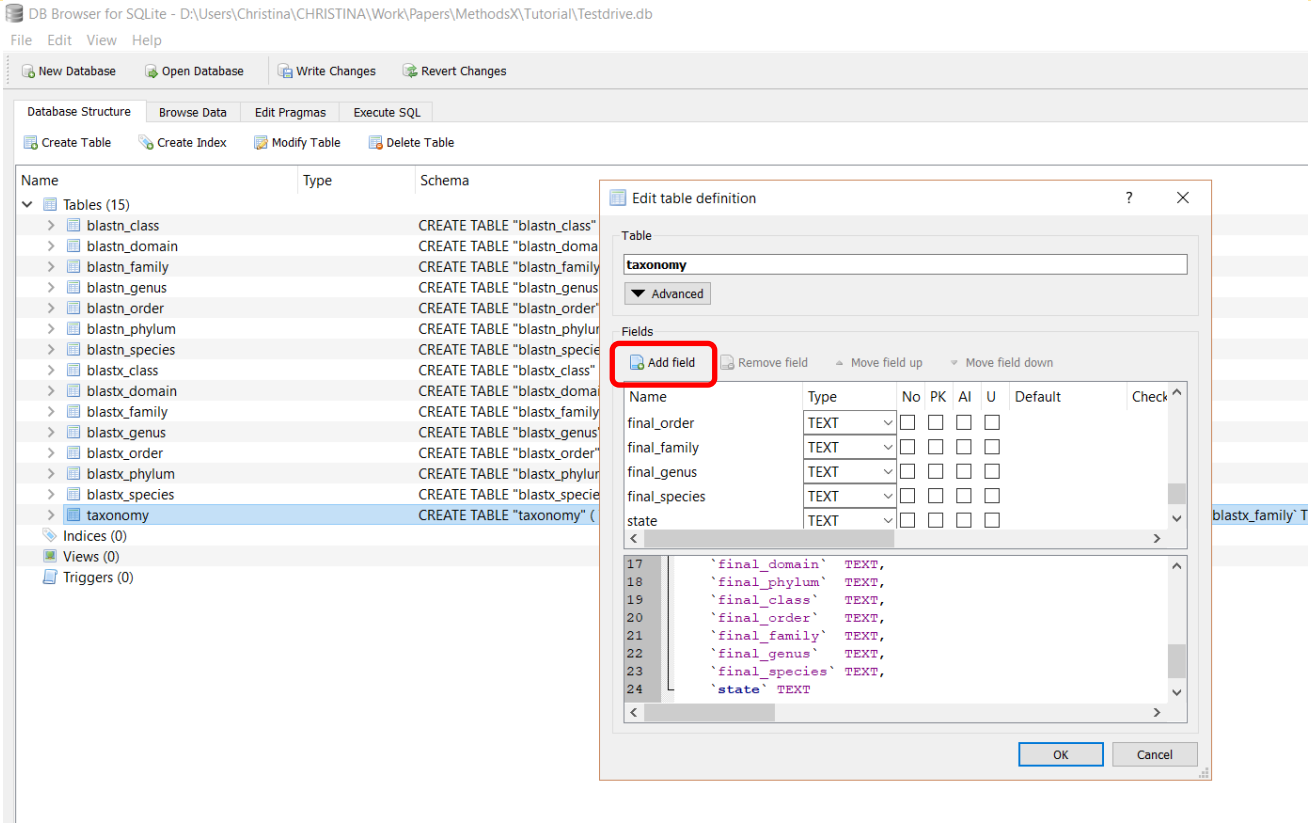


Having obtained the taxonomy.csv file, we now import it to the database to be able to perform the criss-crossing. Having selected the “taxonomy” table in the main window, we now choose “Modify Table” as indicated previously:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1



In this example, 7 columns were added to the “taxonomy” table (final_domain, final_phylum, etc.), to be completed with the criss-crossing result, and another column “state” was added to indicate if the contig was assigned or not (to avoid multiple assignments). To add more columns to the table, click on “Add field”:



IV) Intersection of taxonomic data:

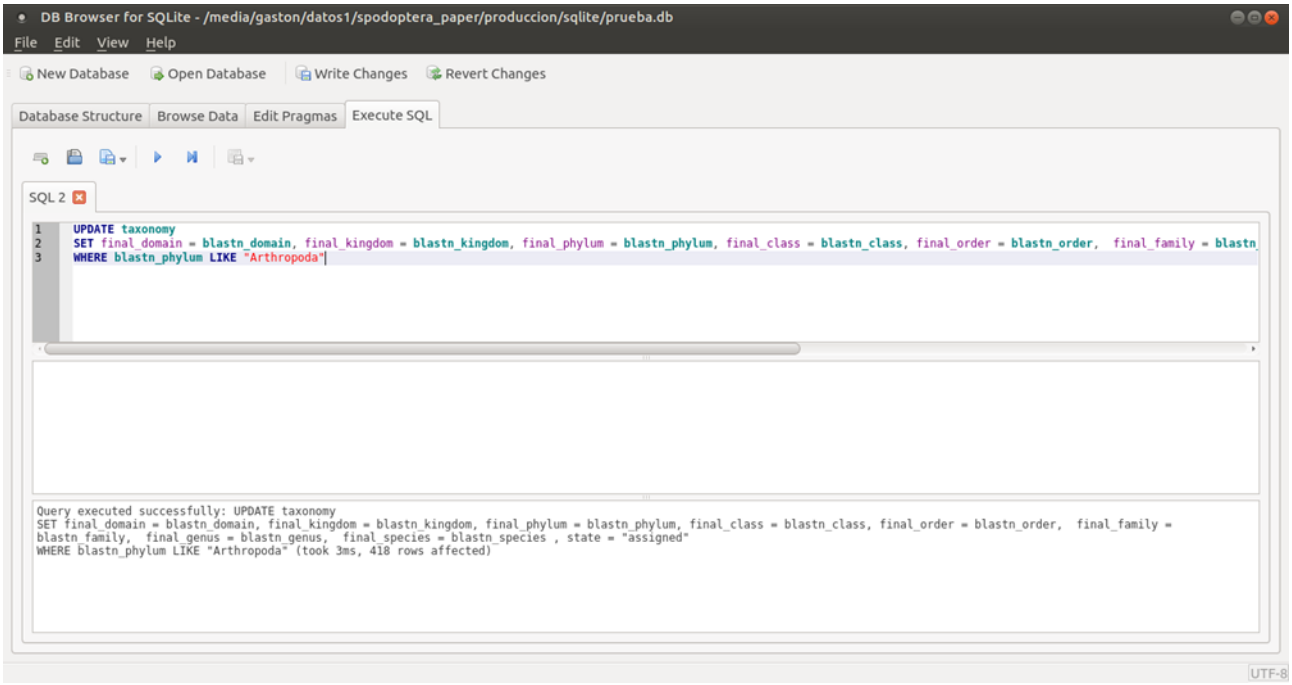
Summary:

The “taxonomy” database created in the DB browser for SQLite was then used to elucidate the inconsistencies observed in the homology search results, based on the aforementioned assumptions.

Step by step tutorial:

a) Contigs that showed hits against lepidopterans in the nt16SLep homology search, were directly annotated with the taxonomic assignment of that homology search using the following commands:

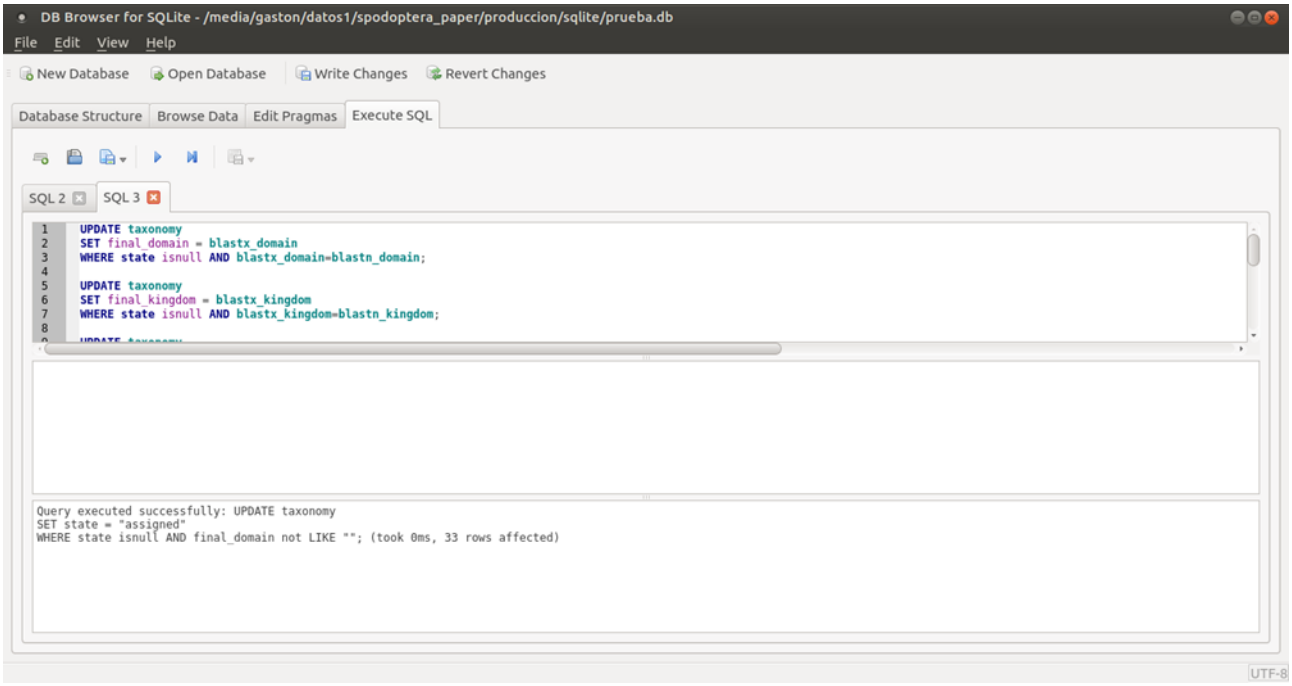
```
UPDATE taxonomy
SET final_domain = blastn_domain, final_phylum = blastn_phylum, final_class = blastn_class, final_order = blastn_order,  final_family = blastn_family, final_genus = blastn_genus, final_species = blastn_species , state = "assigned"
WHERE blastn_phylum LIKE "Arthropoda"
```



b) Contigs that were assigned to the same taxon by both homology searches, were annotated with the following commands:

```
UPDATE taxonomy
SET final_domain = blastx_domain
WHERE state isnull AND blastx_domain=blastn_domain;

UPDATE taxonomy
SET final_phylum = blastx_phylum
WHERE state isnull AND blastx_phylum=blastn_phylum;

UPDATE taxonomy
SET final_class = blastx_class
WHERE state isnull AND blastx_class=blastn_class;

UPDATE taxonomy
SET final_order = blastx_order
WHERE state isnull AND blastx_order=blastn_order;

UPDATE taxonomy
SET final_family = blastx_family
WHERE state isnull AND blastx_family=blastn_family;

UPDATE taxonomy
SET final_genus = blastx_genus
WHERE state isnull AND blastx_genus=blastn_genus;

UPDATE taxonomy
SET final_species = blastx_species
WHERE state isnull AND blastx_species=blastn_species;

UPDATE taxonomy
SET state = "assigned"
WHERE state isnull AND final_domain not LIKE "";
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

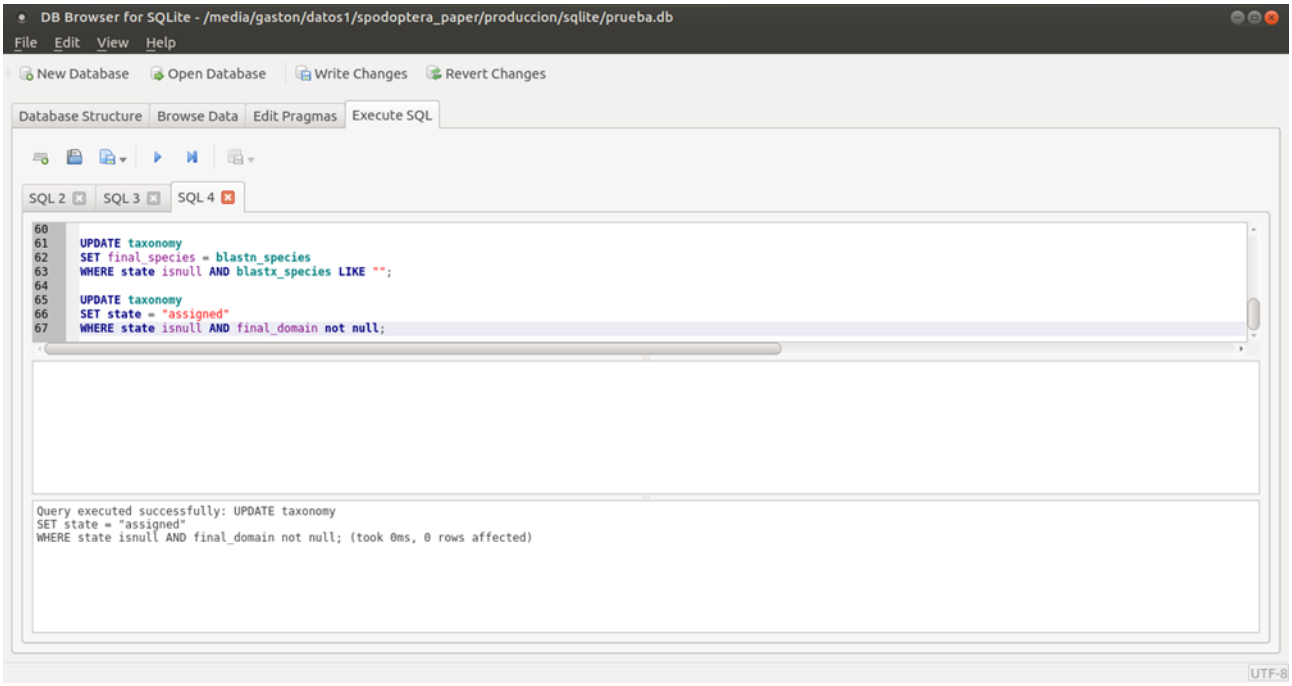


c) Contigs that showed hits in only one of the homology searches (and no hits in the other one), were annotated with the following commands:

```
UPDATE taxonomy
SET final_domain = blastx_domain
WHERE state isnull AND blastn_domain LIKE "";

UPDATE taxonomy
SET final_phylum = blastx_phylum
WHERE state isnull AND blastn_phylum LIKE "";

UPDATE taxonomy
SET final_class = blastx_class
WHERE state isnull AND blastn_class LIKE "";

UPDATE taxonomy
SET final_order = blastx_order
WHERE state isnull AND blastn_order LIKE "";

UPDATE taxonomy
SET final_family = blastx_family
WHERE state isnull AND blastn_family LIKE "";

UPDATE taxonomy
SET final_genus = blastx_genus
WHERE state isnull AND blastn_genus LIKE "";

UPDATE taxonomy
SET final_species = blastx_species
WHERE state isnull AND blastn_species LIKE "";

UPDATE taxonomy
SET final_domain = blastn_domain
WHERE state isnull AND blastx_domain LIKE "";

UPDATE taxonomy
SET final_phylum = blastn_phylum
WHERE state isnull AND blastx_phylum LIKE "";

UPDATE taxonomy
SET final_class = blastn_class
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

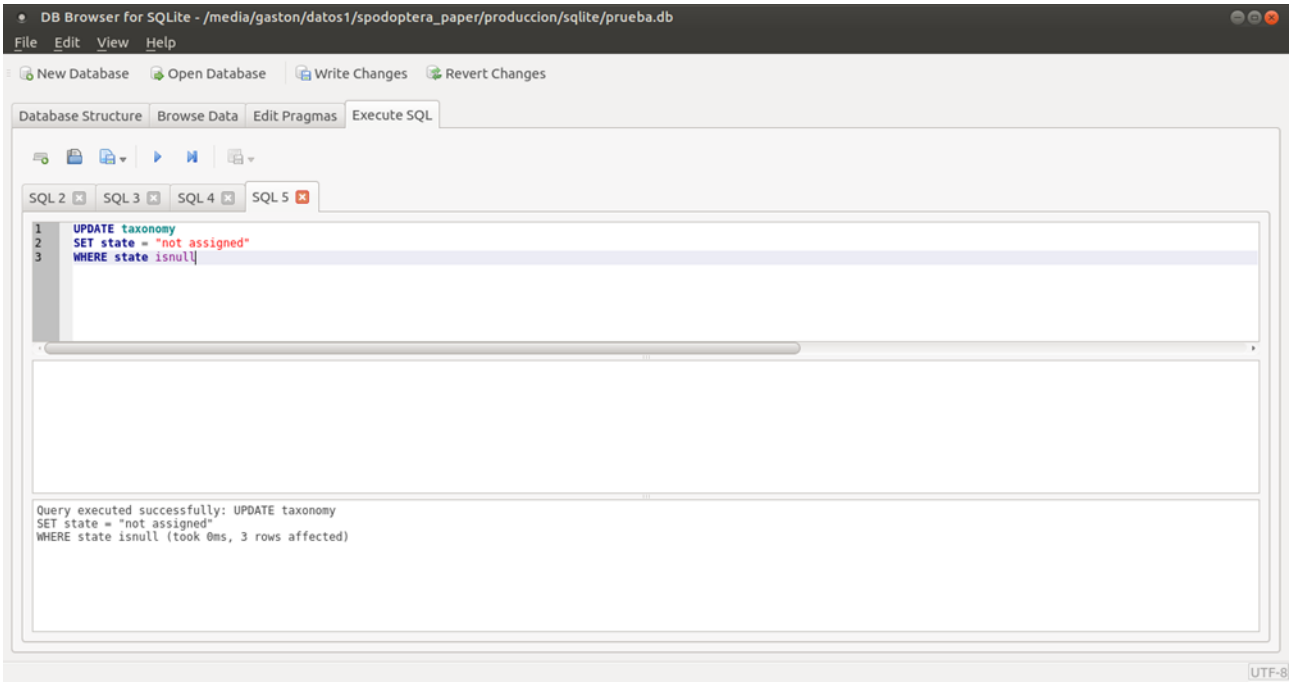
```
WHERE state isnull AND blastx_class LIKE "";

UPDATE taxonomy
SET final_order = blastn_order
WHERE state isnull AND blastx_order LIKE "";

UPDATE taxonomy
SET final_family = blastn_family
WHERE state isnull AND blastx_family LIKE "";

UPDATE taxonomy
SET final_genus = blastn_genus
WHERE state isnull AND blastx_genus LIKE "";

UPDATE taxonomy
SET final_species = blastn_species
WHERE state isnull AND blastx_species LIKE "";

UPDATE taxonomy
SET state = "assigned"
WHERE state isnull AND final_domain not null;
```



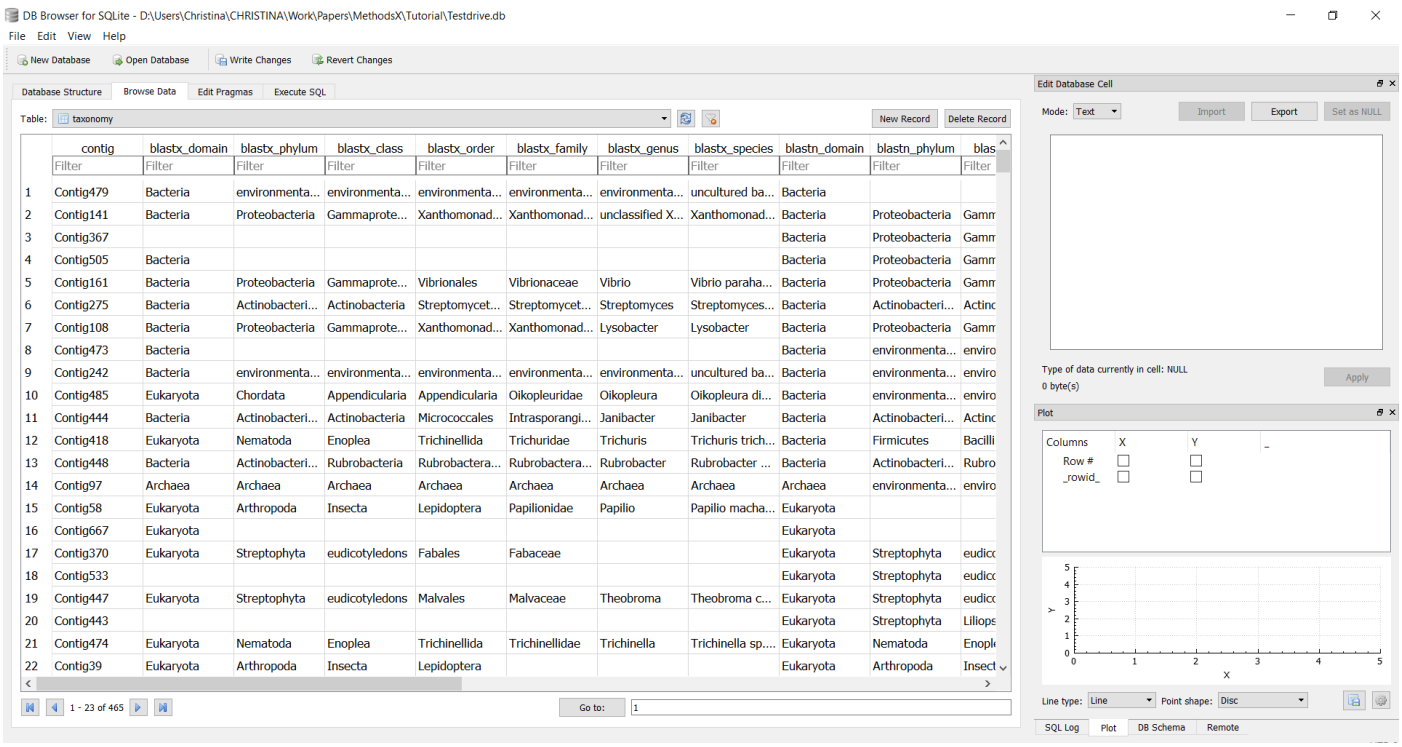
d) Contigs that showed hits in both homology searches but were assigned to a different taxon in each, were considered as “not assigned” with the following command:

```
UPDATE taxonomy
SET state = "not assigned"
WHERE state isnull
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1



e) We are finally ready to browse and analyse the updated “taxonomy” table. For this, go to the “Browse Data” leaf:



Acknowledgements:
This research was supported by Agencia Nacional de Promoción Científica y Tecnológica (PICT PRH 112 and PICT CABBIO 3632), and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) (PIP 0294) grants to CBM.
CBM is a member of the CONICET research career. GR and JMC are the recipients of CONICET fellowships.

Supplementary material:

Background:
This workflow was developed to analyse total RNA extracted from insects. Total RNA was submitted to a reverse transcription and polymerase chain reaction sequence-independent amplification procedure, which generated an unbiased random collection of transcripts in which the vast majority corresponded to the host [11–13]. The objective of these studies was to differentiate and determine the host transcriptomic and associated metatranscriptomic profiles of the sample. For this reason, we included all the

latest insect genome sequences in our local database to identify host sequences, as well as the non-redundant (nucleotide and protein) and 16S microbial databases, to identify the metatranscriptome. Due to the fact that we observed many differences in the BLASTN and BLASTX results, we developed this workflow to be able to resolve these inconsistencies. It must be noted that, even though this workflow was originally developed to analyse an unbiased collection of high-throughput metatranscriptomic sequences, it has since been used to analyse high-throughput metagenomic shotgun sequences, and works seamlessly.

Previous versions of the workflow:

Here we mention two previous versions of the workflow, which are described in chronological order to highlight the adjustments and upgrades that led to the final version presented here. In every case, the objective was to be able to observe, compare, and so determine, the correct result for each sequence (and thus for each taxon). Ideally, the results from each homology search would be readily available and visible for each sequence in a table, and the two previous versions of this workflow show how the procedure evolved towards this.

I) Workflow designed to analyse total RNA extracted from whole insects (*Lutzomyia longipalpis*) [12,13].
II) Workflow designed to analyse total RNA extracted from insect guts (*Spodoptera frugiperda*) [11].

I) Workflow designed to analyse total RNA extracted from whole insects (*Lutzomyia longipalpis*) [12,13]:

Main features:

- ⇒ Results were classified and sorted in excel spreadsheets using various custom scripts (written in Mathematica), and then manually combined and compared.
- ⇒ The taxonomic and functional profile of the sample was defined on the basis of the different homology search results (and BLAST2GO results, partially) after analysing and curing the spreadsheets manually.
- ⇒ MEGAN was not used;
- ⇒ No formal criss-crossing was performed.

Pros:

- ➢ It was possible to manually combine homology results in a spreadsheet (and so the results from each homology search for each sequence were combined and visible in a single table).

Cons:

- ➢ A lot of manual curation required;
- ➢ Very time-consuming;
- ➢ Custom scripts had to be modified each time the format of the homology search outputs differed (however minimal the differences)

Workflow:

A preliminary version of the *L. longipalpis* genome (Llon v1.0 contigs, here referred to as Llon_contigs) was downloaded from the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) website (http://www.hgsc.bcm.tmc.edu). Additionally, five databases were downloaded locally:
- non-redundant nucleotide sequences (nt) (NCBI),
- 16S ribosomal RNA sequences (Bacteria and Archaea) (16S) (NCBI),
- non-human, non-mouse ESTs (est-others) (NCBI),
- non-redundant protein sequences (nr) (NCBI), and
- UniProt Knowledgebase (uniprotKB), consisting of the Swiss-Prot Protein Knowledgebase (fully annotated curated entries) and the TrEMBL Protein Sequence Database (computer-generated entries enriched with automated classification and annotation) (uniprotKB = uniprot_sprot + uniprot_trembl).

In a first stage, sequences from Llon_contigs, nt and 16S were combined in a single database (DB:Llon_contigs+nt+16S) using the appropriate BLAST+ applications (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Trimmed reads were compared to this combined database using BLASTN (nucleotide homology) [3], with a 1e-50 cutoff E-value (Figure S1). Reads which mapped to Llon_contigs by sequence homology (here referred to as mapped-reads) were then separated from those that showed homology to nt16S (here referred to as read-nt16S), and an accurate record was kept of which Llon_contig each mapped-read showed homology to, using custom applications written in Mathematica (Wolfram Mathematica 7; available upon request) (Figure S1). Results from the first stage of analysis were organised in three datasets: mapped-reads (reads that mapped to Llon_contigs by sequence homology), read-nt16S (reads that showed homology to nt16S) and no hits (reads that returned no significant hits). Results from the read-nt16S dataset were then classified using custom applications written in Mathematica (Wolfram Mathematica 7; available upon request) and those read-nt16S that did not show homology to insect rDNA or taxa other than insects, were selected and separated from the rest to be further analysed (Figure S1).

In a second stage, database sequences from nt and 16S were combined in a single database (DB:nt16S) using the appropriate BLAST+ applications (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), and the mapped-reads dataset was compared to DB:nt16S using BLASTN [3] with a 1e-50 cutoff E-value. After analysing the results of this homology search, it was observed that mapped-reads which in the first stage showed homology to Llon_contigs 15379, 23694, 25834, 27903, 27904 and 9281, after BLASTN against DB:nt16S, overall showed homology to insect rDNA (these were the only Llon_contigs that were partially annotated after blasting their corresponding mapped-reads against DB:nt16S, i.e. the second stage of analysis), and mapped-

reads that in the first stage showed homology to Llon_contig 31202, after BLASTN against DB:nt16S, overall showed no significant hits. Therefore, these mapped-reads were excluded from the third stage of analysis (Figure S1).

In the third stage of analysis, all mapped-reads excluding the aforementioned ones (a total of 4910 mapped-reads considering all four samples) and the selected read-nt16S (i.e. those that did not show homology to insect rDNA or taxa other than insects, see above), were then compared separately to three databases: DB:est-others using BLASTN [3] with a 1e-50 cutoff E-value, and DB:nr and DB:uniprotKB using BLASTX [7] with a 1e-6 cutoff E-value; these mapped-reads and read-nt16S were also analysed and annotated using Blast2GO [14] (Figure S1). Hits from all four databases and Blast2GO results were compared and individually revised and confirmed, and only those which showed unequivocal results were included in the final analysis. Putative functional assignment and categorisation for each selected mapped-read and read-nt16S was individually and manually revised and assigned on the basis of BLASTX results (mainly DB:uniprotKB) and Blast2GO annotation.

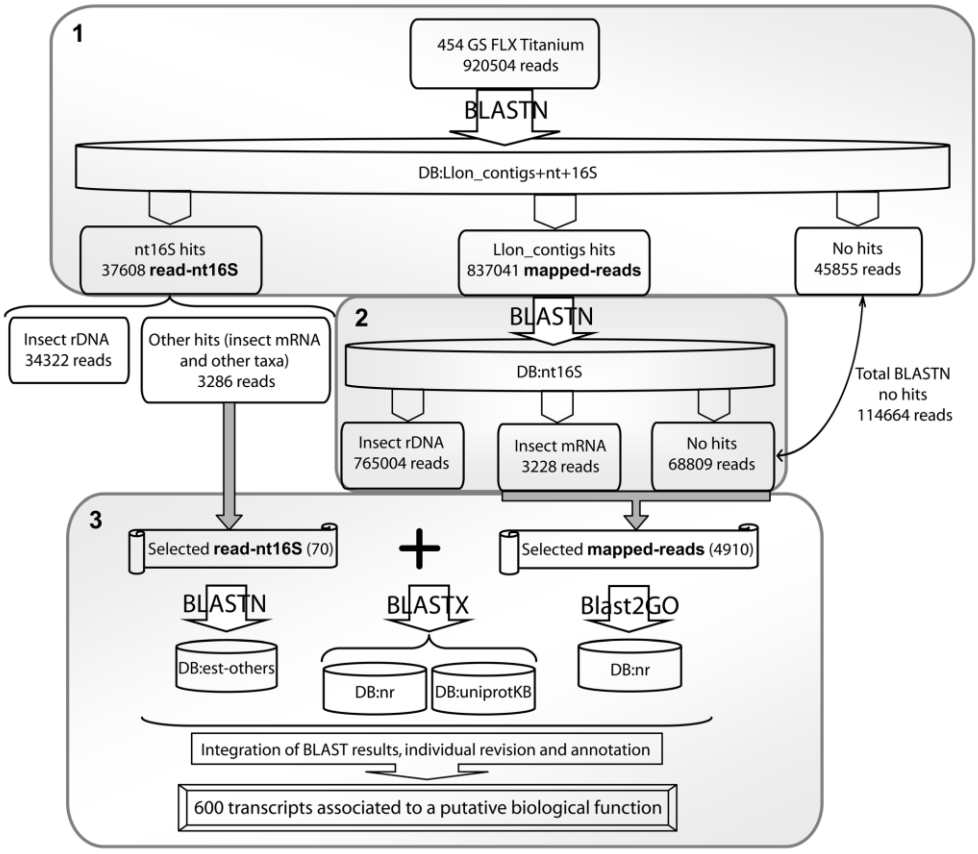


Figure S1. Sequence analysis workflow.
This figure shows an overview of the rationale supporting the analysis of the reads and summarises the different steps that were followed.
1) Indicates the first stage of analysis, in which all the reads were blasted against DB:Llon_contigs+nt+16S. Results were classified in three datasets: reads that mapped to Llon_contigs by sequence homology (mapped-reads); reads that showed homology to nt16S (read-nt16S); and reads that returned no significant hits (no hits) (see text for details).
2) Indicates the second stage of analysis, in which the mapped-reads dataset was blasted against DB:nt16S (see text for details).
3) Indicates the third stage of analysis, in which selected reads (mapped-reads and read-nt16S) were separately blasted against three databases (est-others, nr and uniprotKB) and annotated using Blast2GO (see text for details).
The "selected mapped-reads (4910)" excluded mapped-reads 15379, 23694, 25834, 27903, 27904 and 9281, which showed homology to insect rDNA after BLASTN against DB:nt16S, and contig 31202, with unknown function after BLASTN against DB:nt16S (see text for details).
The "selected read-nt16S (70)" included read-nt16S that did not show homology to either insect rDNA or taxa other than insects (see text for details).

II) Workflow designed to analyse total RNA extracted from insect guts (*S. frugiperda*) [11]:
Main features:

- ⇒ MEGAN was used to classify and sort the homology results, in place of the various custom scripts that were used for this same purpose in the previous workflow;
- ⇒ Consequently, results were not classified and sorted in excel spreadsheets;
- ⇒ The taxonomic and functional profile of the sample was determined based on the assumption that the sequences that correspond to a certain taxon are composed of (Figure S2):
  a) sequences that were assigned to the same taxon by both homology searches (1 in Figure S2), plus
  b) sequences that were assigned to that taxon by one of the homology searches but returned no hits in the other one, and vice versa (2 and 3 in Figure S2);

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

⇒ For this, the information from both homology searches for each taxon was extracted using one of MEGAN's features, and a custom script (written in Mathematica) was used to intersect this information to identify all the reads that unequivocally corresponded to a certain taxon, and then combine them in a single file.

⇒ Formal criss-crossing was performed using a custom script (written in Mathematica).

Pros:

- MEGAN greatly simplifies the classification, sorting and manipulation of homology search results (compared to the custom scripts that were used in the previous workflow);
- Custom scripts were only used to intersect the data.

Cons:

- Homology results were not combined in a spreadsheet and it was thus difficult to visualise the information for each sequence.
- A lot of manual curation was still required;
- Time-consuming.

Workflow:

The following NCBI databases were downloaded locally (ftp://ftp.ncbi.nlm.nih.gov/blast/db/):
- "Non-redundant" nucleotide sequence (nt) (NCBI),
- 16S ribosomal RNA sequences (Bacteria and Archaea) (16S) (NCBI),
- sequences from lepidopteran whole genome shotgun projects completed at the time of the analysis (Lep) (NCBI), and
- non-redundant protein sequence (nr) (NCBI).

Sequences from nt, 16S, and Lep, were then combined in a single database (DB:nt+16S+Lep) using the appropriate BLAST+ applications (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Trimmed reads were compared to this combined database (DB:nt+16S+Lep) using BLASTN (BN) (nucleotide homology) [3], with a 1e-50 cutoff E-value. Rapsearch2 (RS) [15], which achieves more than 90X acceleration when compared to BLASTX, was used for the protein similarity search against DB:nr with a 1e-7 cutoff E-value. MEGAN (MEtaGenome ANalyzer) [2] was downloaded locally (http://ab.inf.uni-tuebingen.de/data/software/megan5/download/welcome.html) and used to analyse both homology search results. A custom script (written in Mathematica) was used to compare the results from the different analyses (see below). This script not only enabled the comparison of the results, but also generated fasta files of the reads that were found in common, and of the reads that were found in one group but not the other, and vice versa.

Figure S2 shows an overview of the rationale supporting the sequence analysis workflow used in this study, summarising the different steps that were followed to intersect the information from both analyses and so be able to define the unequivocal taxonomic and functional profile of the sample, based on the assumption that the sequences that correspond to a certain taxon are composed of:

a) those sequences that were assigned to the same taxon by both homology searches (1 in Figure S2), plus

b) those sequences that were assigned to that taxon by one of the homology searches but returned no hits in the other one, and vice versa (2 and 3 in Figure S2).

General workflow:

Trimmed reads were submitted to homology searches against DB:nt+16S+Lep and DB:nr using BLASTN (BN) and RAPSearch2 (RS), respectively. The homology search results were analysed with MEGAN and a detailed comparison was then performed in the following way (Figure S2):

1) For each major group, reads were extracted using the appropriate MEGAN tool, and were named reads-nr or reads-nt depending on the homology search they came from (RS or BN, respectively).

2) Reads-nr and reads-nt from the same group (for e.g., Bacteria; Figure S3) were then compared using a custom script (written in Mathematica), and those reads that were found in common were directly assigned to that group (1 in Figures S2 and S3).

3) To retrieve the reads that were assigned to a certain taxon by one of the homology searches but returned no hits in the other one, and vice versa (2 and 3 in Figures S2 and S3), two different analyses were performed (which produced the same end result):

First analysis:

3) a. i. Using the custom script, reads-nr were compared with all the reads that retrieved no hits in the global BN analysis (Total no hits BN in Figures S2 and S3), and reads-nt were compared with all the reads that retrieved no hits in the global RS analysis (Total no hits RS in Figures S2 and S3).

3) a. ii. Reads that were found in common (i.e., reads-nr that had no hits in BN, 2 in Figures S2 and S3, and reads-nt that had no hits in RS, 3 in Figures S2 and S3), were also assigned to the taxonomical group that was being analysed (for e.g., Bacteria; Figure S3).

3) a. iii. Reads that showed hits to different taxa in the RS and BN analyses (for e.g., Bacteria in one analysis and Fungi in the other), were considered "not assigned". The only exception to this was Arthropoda, because various taxonomical groups that appeared in the RS analysis did not figure in the BN analysis, and when the reads for these groups were extracted and separately analysed with BN, they showed unequivocal homology to Arthropoda. For this reason, they were subsequently assigned to Arthropoda.

Second analysis:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

3) b. i. All the reads that retrieved no hits in the global RS (nohits-nr) and BN (nohits-nt) analyses, were extracted using the appropriate MEGAN tool, and compared using the custom script. Here again, reads that were found in common were unquestionably assigned as “no hits”.

3) b. ii. On the other hand, reads that only retrieved no hits in either the RS or BN analyses, were submitted to BN and RS homology searches, respectively, and the results were subsequently analysed with MEGAN.

3) b. iii. If the results in one of these searches included reads that were assigned to a certain taxon (for e.g., Bacteria), they were assigned to that taxon (2 and 3 in Figure S3).

4) After this criss-crossing, reads belonging to the taxonomical group being analysed were combined in a single multifasta (“total reads = 1 + 2 + 3” in Figures S2 and S3), submitted to RS and BN analysis, and these results were then processed with MEGAN. In this way, since all the reads that unequivocally corresponded to a certain taxon were combined in a single file, it was possible to discriminate and characterise the taxonomic and functional profile of the sample.

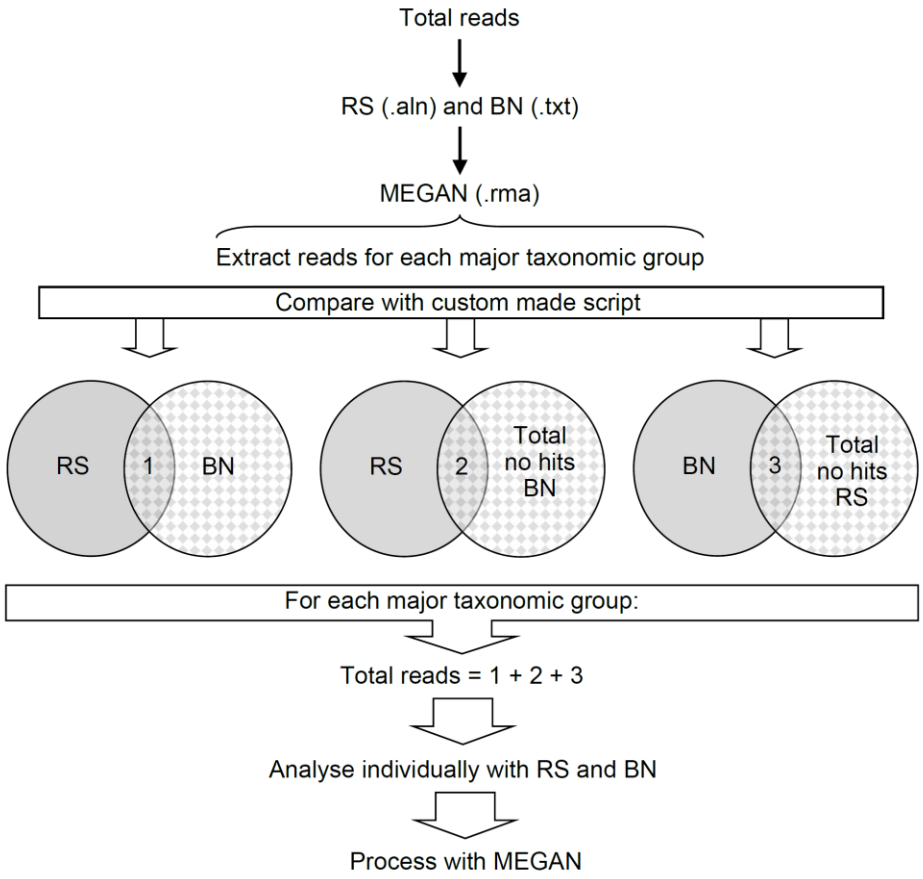


Figure S2. General workflow to define the unequivocal taxonomical and functional profile of a sample. RS = RapSearch2; BN = BLASTN; .aln = RS file extension; .txt = BN file extension; .rma = MEGAN file extension; 1 = sequences that were assigned to taxon X by both homology searches; 2 = sequences that were assigned to taxon X by RS but returned no hits in BN; 3 = sequences that were assigned to taxon X by BN but returned no hits in RS.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2018 doi:10.20944/preprints201812.0184.v1

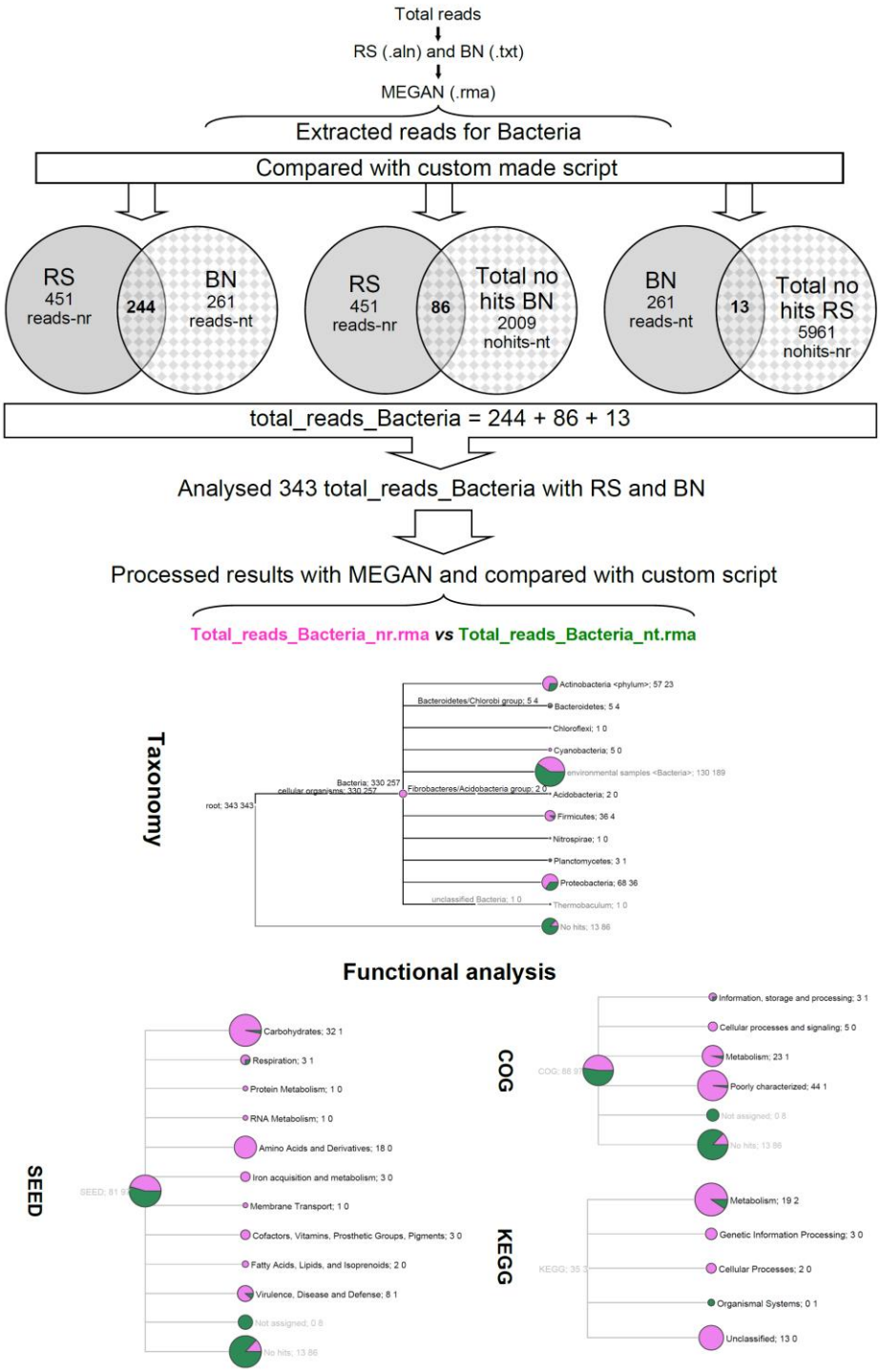


Figure S3. Example of how the workflow was used to define the unequivocal taxonomical and functional profile of Bacteria in the sample. RS = RapSearch2; BN = BLASTN; .aln = RS file extension; reads-nt = reads that showed homology to Bacteria in the BN analysis; reads-nr = reads that showed homology to Bacteria in the RS analysis; nohits-nt = reads that retrieved no results in the BN analysis; nohits-nr = reads that retrieved no results in the RS analysis; .txt = BN file extension; .rma = MEGAN file extension; 1 = sequences that were assigned to taxon X by both homology searches; 2 = sequences that were assigned to taxon X by RS but returned no hits in BN; 3 = sequences that were assigned to taxon X by BN but returned no hits in RS.

***References:**

[1] J.C. Wooley, A. Godzik, I. Friedberg, A primer on metagenomics, PLoS Comput. Biol. (2010). doi:10.1371/journal.pcbi.1000667.

[2] D.H. Huson, A.F. Auch, J. Qi, S.C. Schuster, MEGAN analysis of metagenomic data, Genome Res. (2007). doi:10.1101/gr.5969107.

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[4] M. Kim, K.-H. Lee, S.-W. Yoon, B.-S. Kim, J. Chun, H. Yi, Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era, Genomics Inform. (2013). doi:10.5808/GI.2013.11.3.102.

[5] V. Aguiar-Pulido, W. Huang, V. Suarez-Ulloa, T. Cickovski, K. Mathee, G. Narasimhan, Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis, Evol. Bioinforma. (2016). doi:10.4137/EBO.S36436.

[6] D.H. Huson, S. Mitra, H.J. Ruscheweyh, N. Weber, S.C. Schuster, Integrative analysis of environmental sequences using MEGAN4, Genome Res. 21 (2011) 1552–1560. doi:gr.120618.111 [pii]10.1101/gr.120618.111.

[7] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402. doi:10.1093/nar/25.17.3389.

[8] R. Overbeek, T. Begley, R.M. Butler, J. V Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A.C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, V. Vonstein, The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, Nucleic Acids Res. 33 (2005) 5691–5702. doi:33/17/5691 [pii]10.1093/nar/gki866.

[9] R.L. Tatusov, The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Res. (2000). doi:10.1093/nar/28.1.33.

[10] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. (1999). doi:10.1093/nar/27.1.29.

[11] C.B. McCarthy, N.A. Cabrera, E.G. Virla, Metatranscriptomic analysis of larval guts from field-collected and laboratory-reared Spodoptera frugiperda from the South American subtropical region, Genome Announc. 3 (2015) e00777--15.

[12] C.B. McCarthy, M.S. Santini, P.F.P. Pimenta, L. a. Diambra, First Comparative Transcriptomic Analysis of Wild Adult Male and Female Lutzomyia longipalpis, Vector of Visceral Leishmaniasis, PLoS One. 8 (2013) e58645. doi:10.1371/journal.pone.0058645.

[13] C.B. McCarthy, L. a Diambra, R. V Rivera Pomar, Metagenomic analysis of taxa associated with Lutzomyia longipalpis, vector of visceral leishmaniasis, using an unbiased high-throughput approach., PLoS Negl. Trop. Dis. 5 (2011) e1304. doi:10.1371/journal.pntd.0001304.

[14] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics. 21 (2005) 3674–3676. doi:bti610 [pii]10.1093/bioinformatics/bti610.

[15] Y. Zhao, H. Tang, Y. Ye, RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data, Bioinformatics. 28 (2012) 125–126. doi:10.1093/bioinformatics/btr595.