MEB: Matching and Editing Databases

The MEB software was developed to aid researchers with the editing of transcriptome libraries, by using BLAST functions to pair the user's database (db), a Query, with a reference database, the Subject (Fig. 1). If the reference db is annotated, it is possible to mine genes of interest to the user, even if the query db is not annotated. These same processes can also be applied to short, non-codifying RNA (ncRNA) sequences. MEB also includes other editing tools, which permit the depuration of unwanted fragments and sequences, such as barcodes, low-quality fragments, and ribosomal RNA (rRNA), thus generating a new database free of these fragments. In addition, MEB can count reads or contigs, and convert FasQ files to Fasta files. The software was written in Pascal, and developed to run in Windows, in graphic mode, and it is user-friendly, being easy to install and use. MEB can be run on a conventional PC that has at least 2GB of RAM and a 1GHz processor. This freeware satisfies the demand from small projects and is capable of matching, depuring, and mining databases, generating new, more optimized files.

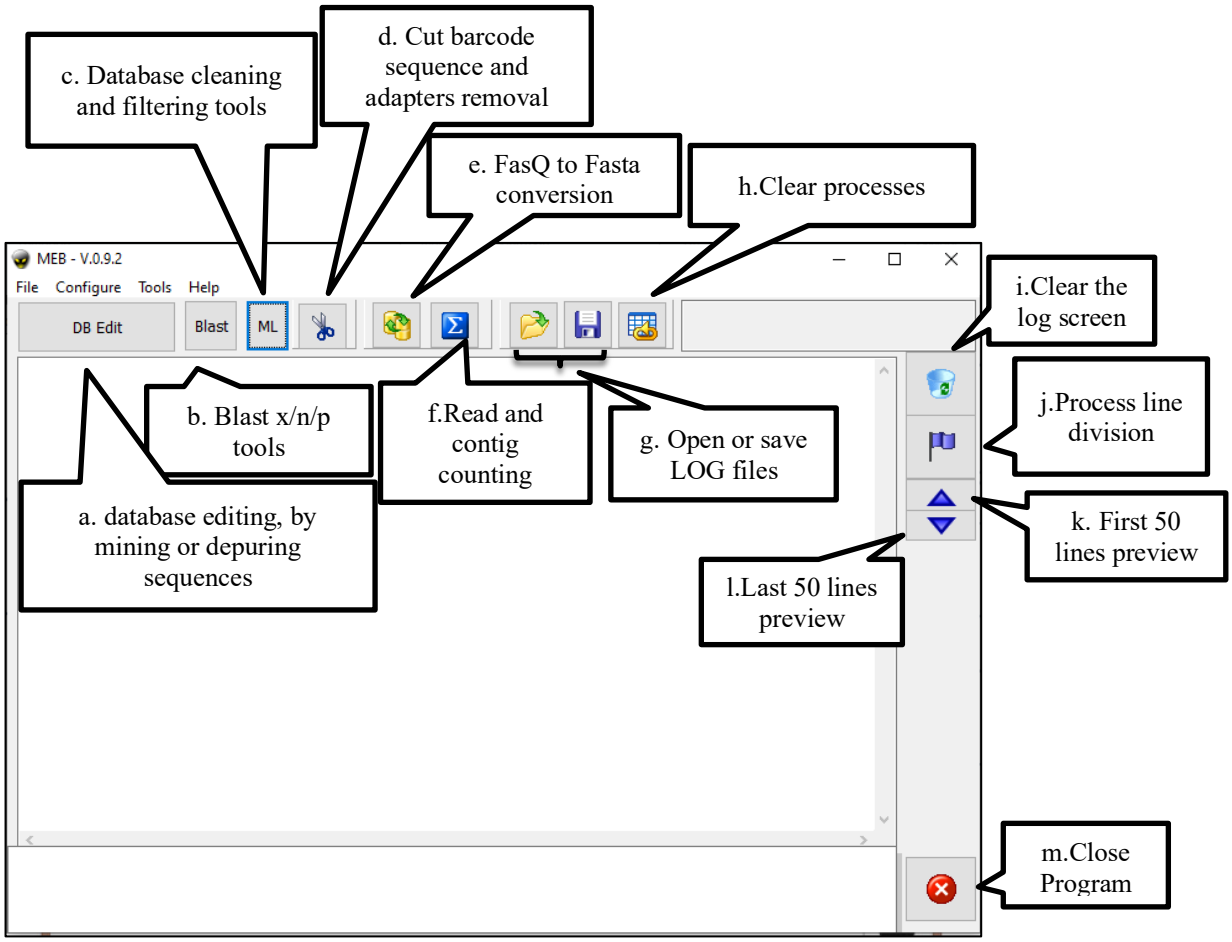

Fig. 1 – Description of the main form and its main functions available.

The principal tools available in MEB

- Depuring or mining of sequences
- Filtering of mined sequences
- Trimming of barcode or adapter sequences
- Conversion of FastQ to Fasta files
- Reads or contigs counting

Requirements:

- BLAST tools (https://www.ncbi.nlm.nih.gov/books/NBK279684/)
- Windows 7 or higher

How to start?

To use the software, the user must first launch it as an administrator. At first access, the user must identify a BLASTn executable folder in the *configure* menu. The MEB will then be ready to search sequences for depuring or mining:

Configure > *Preferences* > *Unit C* > *Program files* > *NCBI* > *blast-2.11.0+* > *Bin* > *blastn.exe* (Fig. 2).

If new versions of MEB are made available by the developers, it will be necessary to reinstall the program or update the *Ribohanterproj.exe* file in the principal MEB folder (C:\MEB).

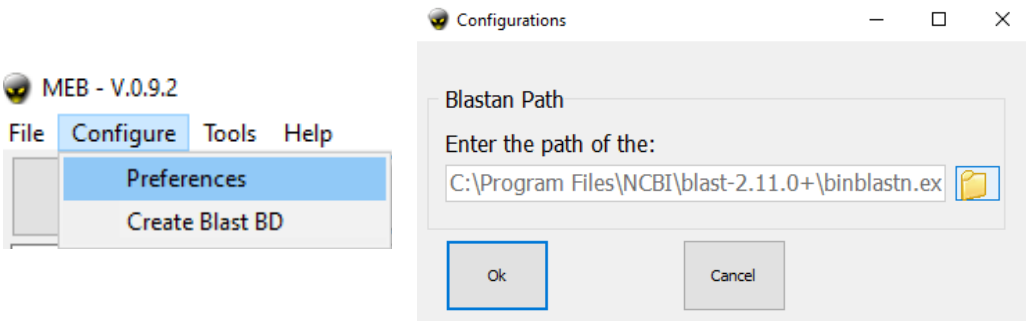

Fig. 2 – Blastn software loading.

File menu (Fig. 3):

- **Open Log/BLAST List File***:* This operation displays the file text in the principal MEB window.

- **Save Log Files***:* This operation saves Log files. These files contain the commands executed by the program.

- **Clear***:* This operation deletes the information displayed on the primary screen, which refers to the progress of the workflow requested by the user.

- **New Process***:* This operation clears all the variables to begin a new process and executes the Clear function.

- **Exit**: Closes the program.

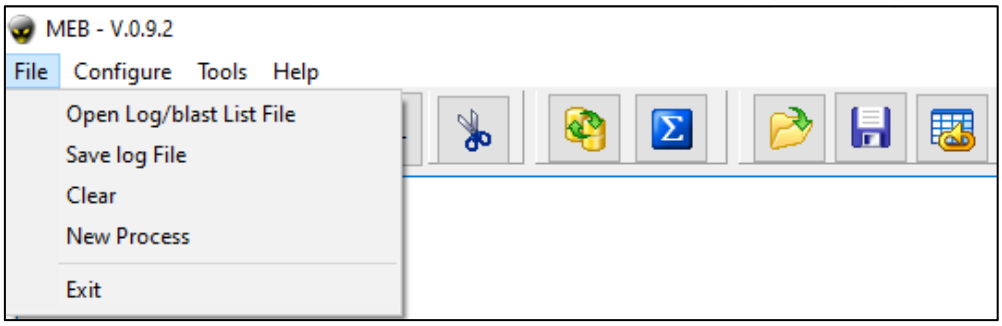

Fig. 3 – *File* Menu from MEB

Depuration or mining of sequences

1. Standard editing mode in MEB

The user can edit databases by clicking on *DB Edit* (Fig.1a) and opening the *Database editing* window (Fig.4).

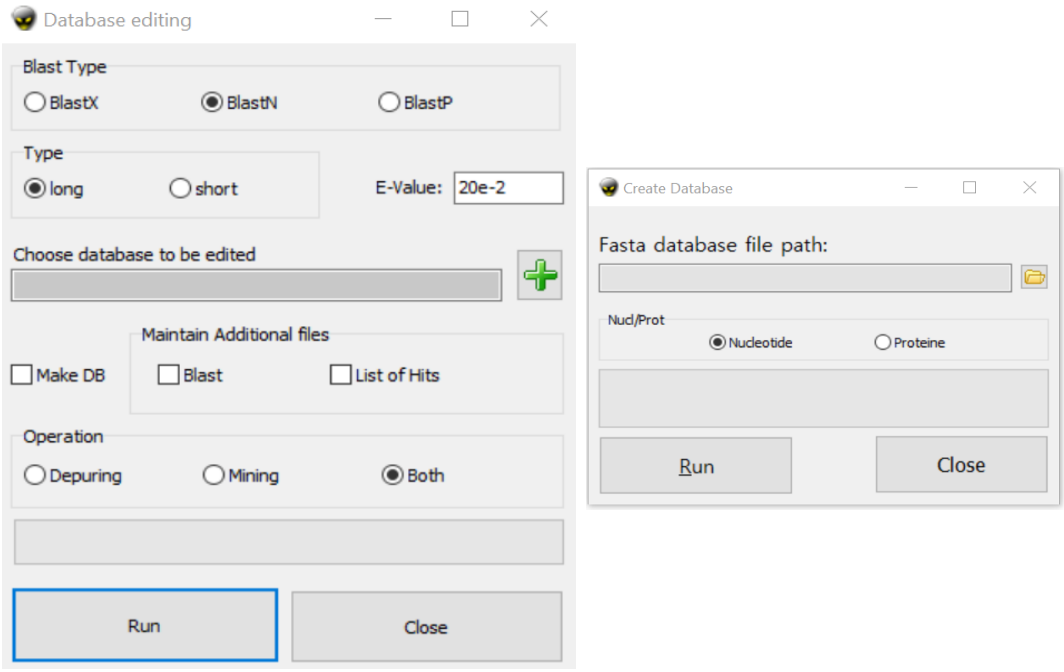

Fig. 4 – MEB's standard editing mode.

- BLAST Type: The software can use *Blast X*, *N* or *P* to execute the analysis of nucleotide or protein sequences.

- Type: Databases with cDNA (*long*) or short sequences (*short*), such as ncRNA, can be used, depending on the objectives of the study.

- *E-Value*: This option permits the selection of the precision of the MEB search, in order to process sequences of greater or lesser similarity.

- **Choose database to be edited:** Here, the user selects the target (Query) database to be edited.

- **Create DB:** The reference database for comparison can be loaded by selecting *Create DB* or by *Configure>Create Blast DB*. If *Create DB* is selected, followed by the *Run* command, the *Create Database* window will be opened for the user to define the reference database (Subject). *Create DB* does not need to be selected if the reference database has already been loaded.

- **Save Additional files:** If *BLAST* is selected, the BLAST search output file will be saved, and if *List of hits* is selected, the list of hits, with the names of the sequences common in both databases will be saved. If the user opts not to save these files, they will be used for the process and subsequently deleted.

- Operation: Here, the user can opt to conduct *Depuring* or *Mining* individually, or *Both*, depending on the objective of the study. *Depuring* creates a new database without sequences found in the BLAST hit list. *Mining* creates a new database made only from sequences defined in the BLAST hit list.

2. Personalized edition mode

The database can also be edited using the *Configure* menu, followed by the *BLAST* and *ML* commands. In this case, three steps have to be undertaken sequentially:

First step – Loading the reference database

The first step in the sequence mining is the loading of the reference database, which is the database used to search for specific sequences in the target (Query) database. Here, the user has to select *Configure>Create Blast DB* (Fig. 5).

- Fasta database file path: The user then selects the reference database file to use as reference for editing the target database.

- Nucl/Prot: Depending on the database used, the user must select the type of sequence, either *Nucleotide* or *Protein*, to be used in the BLAST X/N or P, respectively.

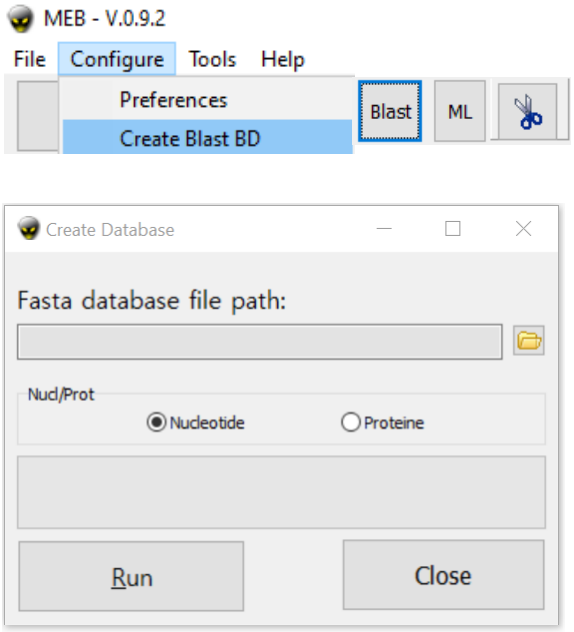

Fig. 5 – Loading of reference database (Subjct).

Second step – BLAST-ing

After selecting the database, the reference and targeted databases are paired, using the *BLAST* command (Fig. 1b), in the *BLAST–ing* window (Fig. 6). This will create the result.out file which will be used in the third step.

- Outfmt: Selects the output file format in the BLAST tool:

0 = pairwise,

1 = query-anchored showing identities,

2 = query-anchored no identities,

3 = flat query-anchored, show identities,

4 = flat query-anchored, no identities,

5 = XML BLAST output,

6 = tabular,

7 = tabular with comment lines,

8 = ASN.1 text,

9 = Binary ASN.1,

10 = Comma-separated values,

11 = BLAST archive format (ASN.1)

The pairwise file (0 = pairwise) format is recommended here, given that it is compatible with the next step in the MEB processing.

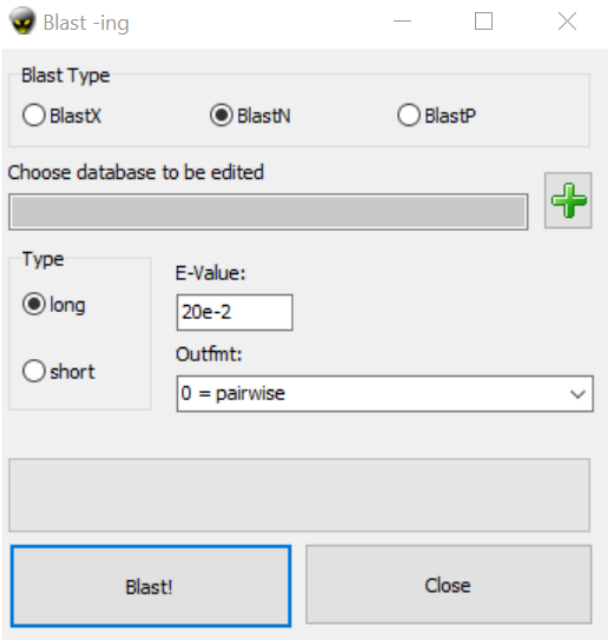

Fig. 6 – Blast function configuration.

Third step – Data mining/depuration

The *ML* button (MicroLimp, Fig. 1c; Fig.7) executes either the sequence mining or depuration process, using the file created in the previous step.

- **HitList**: This process requires the result.out file, created in the previous step. This command creates a Fasta format list of sequences with positive hits.

- **Mine genes**: This process requires two input files, the original target database and the list created using the *HitList* command, which must be in this order. A new database file is then created in Fasta format with only the positive BLAST hit sequences.

- **Depur fragments**: This process also requires the two input files, the target database and the list created by the *HitList* command, in this order. A new database file is then created in Fasta format with only the non-hit sequences.

- Clear BLAST Results: This process creates a new result.out file, without the sequences that had no hits, that is, a file with the BLAST information only on the hits obtained in the analysis.

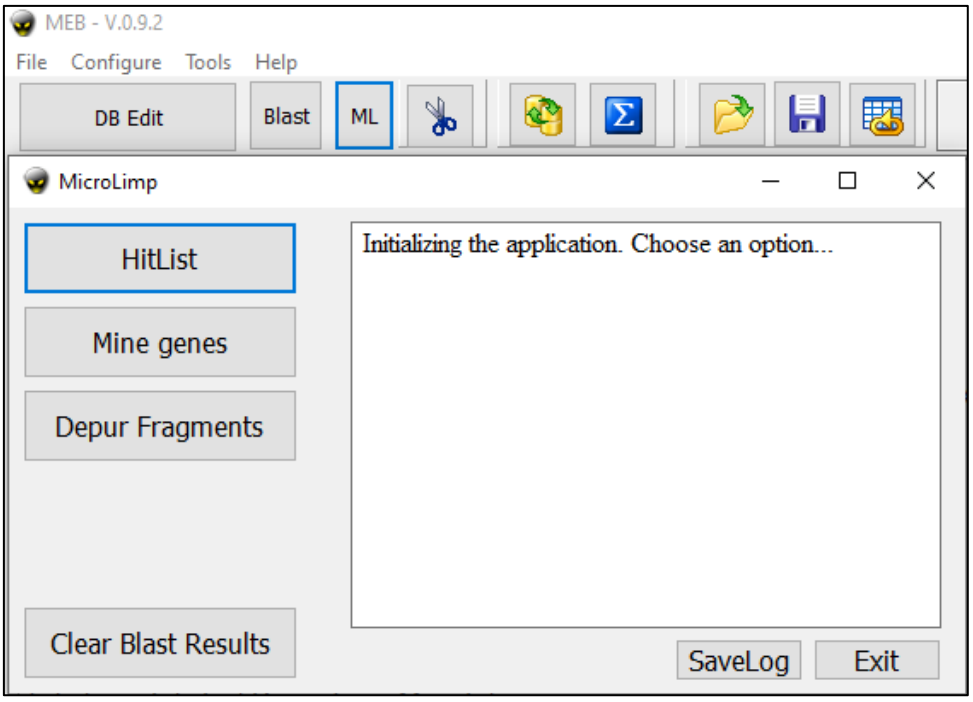

Fig. 7 – Personilized editing process of databases.

Barcode/Adapters and Low-quality region sequence trimming (Fig. 1d)

- **Cut Sequences**: This command provides access to the tools used to create a new depurated file of the barcodes or low-quality sequences. Clicking on *Barcode* will open the *Cutting Barcode* window, where it is possible to input the barcode sequence to be deleted from the beginning of each sequence (Fig. 8a). In *Leading and Trailing*, it is possible to delete small fragments from the 3' and 5' extremities, which correspond to the beginning and end reads (Fig. 8b). The user can set the tool to trim the *Leading*, *Trailing* or *Both* regions. The user must define the number of nucleotides to trimmed in *Beginning* and *Ending* boxes.

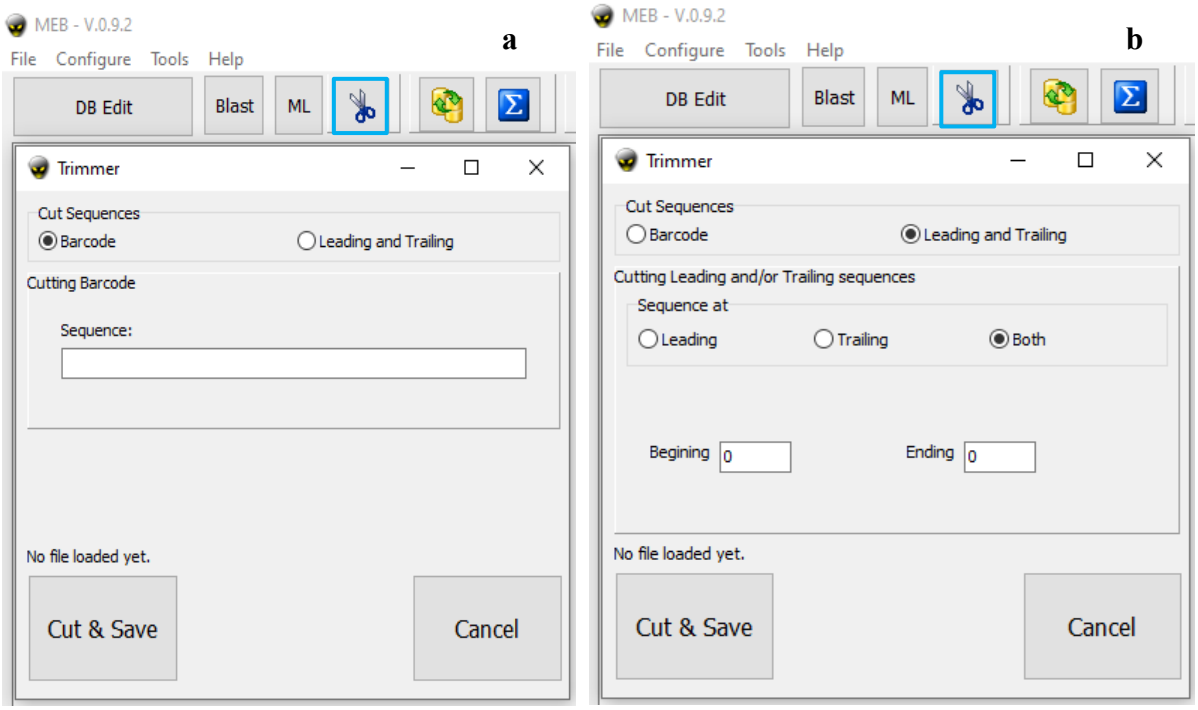

Fig. 8 – Depuring processes of barcode (a) and sequences trimming (b).

- FastQ > Fasta file conversion

The FastQ file created in the preliminary library mounting process contains information on the quality of each predicted base. Here, it is possible to convert FasQ files to the Fasta format (Fig.1e), which is important, given that the latter is the most compatible with other programs (Fig. 9).

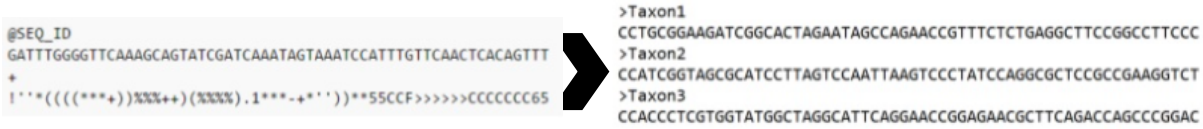

Fig. 9 – FastQ to Fasta file conversion.

Auxiliary MEB functions

- Count Reads: MEB can count the number of sequences present in database files, in both Fasta and FastQ formats. (Fig. 1f).

- Open or save LOG files: Opens a text file in the MEB LOG window, enabling the user to read processes that have already been run and saved (Fig. 1g).

- Clear Processes: LOG-related variables are cleared, so the user can initiate a fresh process (Fig. 1h).

- Clear the LOG screen: Erases the text in the MEB LOG window (Fig. 1j).

- Process line division: Here the user can add a dotted line to the LOG text to separate processes visually (Fig.1j).

- Lines preview: To follow the processes implemented in the files, the user can visualize the first (Fig. 1k) or the last (Fig. 1l) 50 lines of the database being analyzed in the MEB LOG window.

- Close Program: Closes MEB (Fig. 1m).

Help menu

- ***Software Manual***: This option opens the software manual, which describes the MEB functions, and supports the use of the software.

- ***About***: This command opens a window showing the basic information on MEB, including the current version and the contact data of the developers.

Tools Menu

Here, the user can access the functions of FastQ to Fasta conversion, sequence extremity trim, or count reads.