User Manual for MEGAN V4.70.4

Daniel H. Huson

September 11, 2012

Contents

1 Introduction

Disclaimer: This software is provided "AS IS" without warranty of any kind. This is developmental code, and we make no pretension as to it being bug-free and totally reliable. Use at your own risk. We will accept no liability for any damages incurred through the use of this software. Use of the MEGAN is free, however the program is not open source.

Type-setting conventions: In this manual we use e.g. `Edit`→Find... to indicate the `Find...` menu item in the `Edit` menu.

How to cite: If you publish results obtained in part by using MEGAN , then we require that you acknowledge this by citing the program as follows:

- D.H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber and S.C. Schuster, Integrative analysis of environmental sequences using MEGAN 4, Genome Res. 2011. 21:1552-1560, software freely available for academic purposes from www-ab.informatik.uni-tuebingen.de/software/megan.

The first version of the program, as described in [6], was designed by Daniel H. Huson and Stephan C. Schuster. The program was written by Daniel H. Huson. Suparna Mitra, Daniel C. Richter, Paul Rupek, Hans Ruscheweyh and Nico Weber contributed many ideas and some supporting code.

The term *metagenomics* has been defined as "The study of DNA from uncultured organisms" (Jo Handelsman), and an approximately 99% of all microbes are believed to be unculturable. A *genome* is the entire genetic information of one organism, whereas a*metagenome* is the entire genetic information of an *ensemble* of organisms. Metagenome projects can be as complex as large-scale vertebrate projects in terms of sequencing, assembly and analysis.

The aim of MEGAN is to provide a tool for studying the taxonomical content of a set of DNA reads, typically collected in a metagenomics project. In a preprocessing step, a sequence comparison of all reads with a suitable database of reference DNA or protein sequences must be performed to produce an input file for the program. MEGAN is suitable for DNA reads (metagenome data), RNA reads (metatranscriptome data), peptide sequences (metaproteomics data) and, using a suitable synonyms file that maps SILVA ids to taxon ids, on 16S rRNA data (amplicon sequencing).

At start-up, MEGAN first reads in the current NCBI taxonomy (consisting of over 670,000 taxa). A first application of the program is that it facilitates interactive exploration of the NCBI taxonomy.

However, the main application of the program is to parse and analyze a the result of a BLAST comparison of a set of reads against one or more reference databases, typically using BLASTN, BLASTX or BLASTP to compare against NCBI-NT, NCBI-NR or genome specific databases. The result of a such an analysis is an estimation of the taxonomical content ("species profile") of the sample from which the reads were collected. The program uses a number of different algorithms to "place" reads into the taxonomy by assigning each read to a taxon at some level in the NCBI hierarchy, based on their hits to known sequences, as recorded in the BLAST file.

Alternatively, MEGAN can also parse files generated by the RDP website [4] or the Silva website [14]. Moreover, MEGAN can parse files in SAM format [9].

As of version 4, MEGAN provides functional analysis using both the SEED classification [12] and also using KEGG pathways [8].

For an example of its application, see [13], where an early version of this software (called GenomeTaxonomyBrowser) was used to analyze the taxonomical content of a collection of DNA reads sampled from a mammoth.

This document provides both an introduction and a reference manual for MEGAN . Follow MEGAN on facebook at http://www.facebook.com/meganMetagenomeAnalyzer.

2 Getting Started

This section describes how to get started.

First, download an installer for the program from
www-ab.informatik.uni-tuebingen.de/software/megan, see Section 3 for details.

Upon startup, the program will automatically load its own version of the NCBI-taxonomy and will then display the first three levels of the taxonomy. To explore the NCBI taxonomy further, leaves of this overview tree can be uncollapsed. To do so, first click on a node to select it. Then, use the `Tree→Uncollapse` item to show all nodes on the next level of the taxonomy, and use the `Tree→Uncollapse Subtree` item to show all nodes in the complete subtree below the selected node (or nodes).

To analyze a data set of reads, first BLAST the reads against a database of reference sequences, such as NCBI-NR [2] using BLASTX [1] or BLASTP, NCBI-NT [2] using BLASTN [1]. In addition, the output of a number of other programs can also be parsed, for example, RapSearch2 [16].

Then import the BLAST file into MEGAN using the `File→Import From BLAST...` menu item. The Import wizard will ask you to enter the name of the BLAST file, a reads file containing all the read sequences in multi-FastA format (if available), and the name of the new output RMA file. As of version 4, you can also specify more than one BLAST file and one more than one reads file.

Alternatively, instead of supplying a BLAST file, one can also specify a file obtained from the RDP website or from the Silva website. In additional, MEGAN can also parse files in SAM format.

Some implementations or output formats of BLAST suppress those reads for which no alignments were found. In this case, use the `Options→Set Number Of Reads...` menu item to set the total number of reads in the analysis.

Clicking on a node will cause the program to display the exact number of hits of any given node, and the number of hits in the subtree rooted at the node. Right-clicking on a node will show a popup-menu and selecting the first item there, `Inspect` , will open the `Inspector` window which is used to explore the hits associated with any given taxon. A node is selected by clicking on it. Double-clciking on a node will select the node and the whole subtree below it. Double-clicking on the label of a node will open the node in the `Inspector` window.

Example files are provided with the program. They are contained in the `examples` subdirectory of the installation directory. The precise location of the installation directory depends upon your operating system.

3 Obtaining and Installing the Program

MEGAN is written in Java and requires a Java runtime environment version 1.5 or newer, freely available from www.java.org.

MEGAN is installed using an installer program that is freely available from
www-ab.informatik.uni-tuebingen.de/software/megan. There are four different installers, targeting different operating systems:

- `MEGAN_windows_4.70.4.exe` provides an installer for a 32-bit version of MEGAN for Windows-XP.
- `MEGAN_windows-64x_4.70.4.exe` provides an installer for a 64-bit version of MEGAN for Windows 7.
- `MEGAN_macos_4.70.4.dmg` provides an installer for MacOS X.
- `MEGAN_unix_4.70.4.sh` provides a shell installer for Linux and Unix.

The 32-bit Windows version of MEGAN is configured to use 1.1 GB of memory. For all other versions of the software, the installer will allow you to set the maximal amount of memory during the installation process. By default, the program will suggest to use 2 GB. If your computer has more memory available, then it is a good idea to set this limit higher. For example, if you have 4 GB of main memory, then set the limit for MEGAN to 3 GB. This is because the program runs faster, the more memory it is given. To change the maximum amount of memory used after installation of the program, see Section 36.

To install MEGAN using a command-line dialog, launch the installer from the command line and pass the command-line option `-c`. For example, under MacOS X, type the following:

```
/Volumes/MEGAN/MEGAN\ Installer.app/Contents/MacOS/JavaApplicationStub -c
```

4 Program Overview

In this section, we give an overview over the main design goals and features of this program. Basic knowledge of the underlying design of the program should make it easier to use the program.

MEGAN is written in the programming language Java. The advantages of this is that we can provide versions that run under the Linux, MacOS, Windows and Unix operating systems.

Typically, after generating a RMA file (read-match archive) from a BLAST file, the user will then interact with the program, using the Find toolbar to determine the presence of key species, collapsing or un-collapsing nodes to produce summary statistics and using the `Inspector` window to look at the details of the matches that are the basis of the assignment of reads to taxa. The assignment of reads to taxa is computed using the LCA-assignment algorithm, see [6] for details.

The program is designed to operate in two different modes: in a GUI mode, the program provides a GUI for the user to interact with the program. In command-line mode, the program reads commands from a file or from standard input and writes output to files or to standard output.

5 Importing, Reading and Writing Files

To open an existing RMA file or MEGAN text file, select the `File→Open...` menu item and then browse to the desired file. Alternatively, if the file was recently opened by the program, then it may be contained in the `File→Open Recent` submenu.

By default, when parsing an input file, for each read, taxon and RefSeq id, only one best-scoring match is kept. For example, if read R has two equally high-scoring matches M_1 and M_2 to two

sequences from *E. coli*, say, then MEGAN will discard one of the two matches, unless they have different RefSeq accession numbers, or unless exactly one of the two matches does not have a RefSeq accession number. To turn this filter off, use the `Window→Input Command...` menu item to enter the following command `setprop one_match_per_taxon=false`.

5.1 Blast Files

New input to the program is usually provided as a BLAST file obtained from a BLAST comparison of the given set of reads to a database such as NCBI-NR or NCBI-NT, see Section 32 for details of the file formats used. MEGAN supports BLASTN, BLASTX and BLASTP standard text-format, and BLAST XML format. MEGAN can read gzipped BLAST files directly, so there is no need to un-gzip them (although at present MEGAN processes uncompressed files much faster than compressed ones).

MEGAN can also parse tabular BLAST output (generated using BLAST option `-m 8`, however as this form of output does not contain the subject line for sequences matched, it is unsuitable for MEGAN because MEGAN cannot determine the taxon or gene associated with the database sequence. However, if you add an additional column to this format containing the associated taxon name or numerical NCBI taxon-id for each line then MEGAN will parse these and use them as input. For unknown taxa, write either `unknown` or `-1` in the column.

Note that, in all cases, the reads file should be given to use the full potential of the program.

The BLAST file and reads file are supplied to MEGAN when setting up a new *MEGAN project* . Both files are parsed and all information is stored in the project file. The input data is then analyzed and can be interactively explored. All reads and BLAST matches are contained in the project file and MEGAN provides different mechanisms for extracting them again. A MEGAN project file contains all reads and all significant BLAST matches (by default, up to 100 matches per read) in a binary and incrementally compressed format. The size of such a project file is around 20% of the size of the original input files and is thus usually smaller than the file that one obtains by simply compressing the BLAST file. As of version 4, MEGAN provides more control over whether and how BLAST matches and reads are stored, see the discussion of the `Import` window.

As of version 4.41, MEGAN uses a new algorithm for determining the taxon associated with a given reference sequence. In all previous versions, the program looked in the header line of a reference sequence for the longest substring that matches some valid taxon name (or synonym) in the NCBI taxonomy. This determined which taxon to assign to the match. However, because many entries in the NR database mention multiple different species for a given match, the program now determines only maximal matching names in the header line and assigns the match to the LCA of the taxa mentioned. (So, in particular, the LCA algorithm is used twice in MEGAN, namely once to figure out which taxon to assign to a match and then, based on this, again to determine which taxon to assign to a given read.)

5.2 SAM Files

MEGAN can now parse files in *SAM* format [9]. Note, however, that SAM files usually do not contain the names of the taxa associated with the reference sequences and so one must supply a

synonyms file that maps identifiers used for the reference sequences to NCBI taxon names or ids.

5.3 RDP Files

In addition, MEGAN can import rRNA analysis files downloaded from the *RDP* website at `http://rdp.cme.msu.edu/` [4]. Go to the website and upload your rRNA sequences and then let the website process them for you. Please note that the RDP website allows one to download two types of files, namely a *hierarchy as text* file from its `Classifier :: Hierarchy View` window and a *text* file obtained from its `Classifier :: Assignment Detail` window. Input to MEGAN must be of the latter type. The RDP website recommends using a `Min Score` setting of 80. MEGAN calls this the *RDP-Assignment-Detail* format.

If you use the standalone RDP classifier, then the output has a different format. MEGAN calls this the `RDP-standalone` format. In this case, MEGAN expects the format to be a tab-separated file in which each line corresponds to one read:

```
read-name [-] [taxon-name rank-name score] [taxon-name rank-name score] ...
```

In more detail, the first token is a string that identifies the read. The next token is either empty, or a minus, in the latter case indicating that the read is reverse complemented. Then all further tokens come in groups of three, indicating the name of a taxon, the name of the rank of the taxon and a score between 0 and 1 (which MEGAN will multiple by 100). MEGAN reports each such taxon as a separate hit for the read.

5.4 Silva Files

Similarly, MEGAN can import rRNA analysis files downloaded from the *Silva* website at `http://www.arb-silva.de` [14]. To create a file using the Silva website that can be imported into MEGAN, go to the Aligner tab of the Silva website and upload your sequences and then press the *align sequences* button. Once the Silva website has computed an alignment, you will be able to download two files, an *arb file* and a *log file*. MEGAN requires the log file as input, *not* the arb file. When importing such a file into MEGAN, one must specify that MEGAN uses the synonyms file called `silva2ncbi.map` to map Silva accession numbers to NCBI taxa. This file is available from the MEGAN download page.

5.5 CSV Files

MEGAN supports import of data from other programs in a comma-separated format from a CSV file.

5.6 BIOME Format Files

BIOME is a new file format for exchanging data between different metagenome analysis tools. MEGAN can import and export data in BIOME format, see `http://biom-format.org/`. For example, can import OTU classification data generated by the *QIIME* package [3], or taxonomic and functional classifications generated by *MG-RAST* [5], using the `File→Import BIOME Format...`

menu item. To export data in BIOME format, open the viewer for the type of data that you would like to export. For example, if you want to export a SEED classification, then open the SEED Analyzer window and select those nodes that you want to export. Then use the Export→BIOME Format... to save the data to a file. The suffix of a BIOME file is .biom.

6 The NCBI Taxonomy

The *NCBI taxonomy* provides unique names and IDs for over 660,000 taxa, including approximately 25,000 prokaryotes, 84,000 animals, 65,000 plants, and 17,000 viruses. The individual species are hierarchically grouped into clades at the levels of: Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species (and some unofficial clades in between).

At startup, MEGAN automatically loads a copy of the complete NCBI and then displays the taxonomy as a rooted tree. The taxonomy is stored in an NCBI tree file and an NCBI mapping file, which are supplied with the program.

7 The NCBI-NR and NCBI-NT Databases

The *NCBI-NR* ("non-redundant") protein sequence database is available from the NCBI website. It contains entries from GenPept, Swissprot, PIR, PDF, PDB and RefSeq. It is non-redundant in the sense that identical sequences are merged into a single entry.

The *NCBI-NT* nucleotide sequence database is available from the NCBI website. It contains entries from GenBank and is not non-redundant. It contains untranslated gene coding sequences and also mRNA sequences.

8 Assigning Reads to Taxa

The main problem addressed by MEGAN is to compute a "species profile" by assigning the reads from a metagenomics sequencing experiment to appropriate taxa in the NCBI taxonomy. At present, this program implements the following naive approach to this problem:

1. Compare a given set of DNA reads to a database of known sequences, such as NCBI-NR or NCBI-NT [2], using a sequence comparison tool such as BLAST [1].
2. Process this data to determine all hits of taxa by reads.
3. For each read r, let H be the set of all taxa that r hits.
4. Find the lowest node v in the NCBI taxonomy that encompasses the set of hit taxa H and assign the read r to the taxon represented by v.

We call this the LCA-assignment algorithm (LCA = "lowest common ancestor"). In this approach, every read is assigned to some taxon. If the read aligns very specifically only to a single taxon, then it is assigned to that taxon. The less specifically a read hits taxa, the higher up in the taxonomy

it is placed. Reads that hit ubiquitously may even be assigned to the root node of the NCBI taxonomy.

If a read has significant matches to two different taxa a and b, where a is an ancestor of b in the NCBI taxonomy, then the match to the ancestor a is discarded and only the more specific match to b is used.

The program provides a threshold for the bit score of hits. Any hit that falls below the threshold is discarded. Secondly, a threshold can be set to discard any hit whose score falls below a given percentage of the best hit. Finally, a third threshold can be used to report only taxa that are hit by a minimal number of reads. By default, the program requires at least five reads to hit a taxon, before the taxon is deemed present. All reads that are initially assigned to a taxon that is not deemed present are pushed up the taxonomy until a node is reached that has enough reads.

Taxa in the NCBI taxonomy can be excluded from the analysis. For example, taxa listed under `root - unclassified sequences - metagenomes` may give rise to matches that force the algorithm to place reads on the `root` node of the taxonomy. This feature is controlled by `Options→Taxon Disabling` menu. At present, the set of disabled taxa is saved as a program property and not as part of the Megan document.

Note that the *LCA-assignment algorithm* is already used on a smaller scale when parsing individual blast matches. This is because an entry in a reference database may have more than one taxon associated with it. For example, in the NCBI-NR database, an entry may be associated with up to 1000 different taxa. This implies, in particular, that a read that may be assigned to a high level node (even the root node), even though it only has one significant hit, if the corresponding reference sequence is associated with a number of very different species.

Note that as of version 4.60.1, the list of disabled taxa is also taken into consideration when parsing a BLAST file. Any taxa that are disabled are ignored when attempting to determine the taxon associated with a match, unless all recognized names are disabled, in which case the disabled names are used.

9 Identification of SEED Functional Classes

The *SEED* classification of gene function consists of a collection of biologically defined *subsystems* [12]. The SEED classification can be displayed as a tree containing about 10,000 nodes and edges. Genes are mapped onto *functional roles* and these are present in one or more subsystems. The program will attempt to map each read onto a gene that has an known functional role and then into one or more subsystems.

To perform this analysis, MEGAN uses a mapping of *RefSeq* ids to SEED functional roles. Hence, if a SEED-based analysis is desired, then the database that is used in the BLAST comparison must contain RefSeq-ids. This is the case for the NCBI-NR database.

10 Mapping of Reads to KEGG Pathways

The *KEGG* database provides a collection of *metabolic pathways* and other pathways [8]. The KEGG classification can be displayed as a trees. Genes are mapped onto so-called *KO* identifiers and these are present in one or more pathways. The program will attempt to map each read onto a gene that has a valid KO identifier and thus to one or more pathways.

To perform this analysis, MEGAN uses a mapping of RefSeq-ids to KO identifiers. Hence, if a KEGG-based analysis is desired, then the database that is used in the BLAST comparison must contain RefSeq-ids. This is the case for the NCBI-NR database.

11 Comparison of datasets

Multiple datasets can be opened simultaneously and then displayed together in a comparison view.

12 Main Window

The`Main` window is used to display the taxonomy and to control the program via the main menus. Initially, at startup, before reopening or creating a new RMA file, the `Main` window displays the NCBI taxonomy. By default, the taxonomy is only drawn to its second level. Parts of the taxonomy, or the full taxonomy, can be explored using the menu items of the window.

Once a data set has been read in, the full NCBI taxonomy is replaced by the taxonomy that is induced by the data set. The size of nodes indicates the number of reads that have been assigned to the nodes using the algorithm described in Section 8.

Double-clicking on a node will produce a textual report stating how many reads have been assigned to the corresponding taxon and how many reads have been assigned in total to the taxon and to any of the taxa below the given node in summary.

Subtrees can be collapsed and expanded, as described below. Most windows in MEGAN provide access to their functionality through menus, a tool bar that contains a selection of the menu items, and popup menus that also provide contextual access to menu items.

We now discuss all menus of the `Main` window.

12.1 The File Menu

The `File` menu contains the following items:

- The `File`→`New...` item: Open a new empty document.
- The `File`→`Open...` item: Open a MEGAN file (ending on .rma, .meg or .megan). Under Linux or Windows, multiple files can be selected and opened. To open multiple files simutaneously on a Mac, press the shift-key when selecting this menu item so as to obtain a different file dialog that allows selection of multiple files.

- The `File→Open Recent` submenu.
- The `File→Import From BLAST...` item: Import BLAST file (RDP, or Silva) and reads files to create a new MEGAN file.
- The `File→Save As...` item: Save current data set.
- The `File→Export` submenu.
- The `File→Export Image...` item: Export the tree to an image file.
- The `File→Page Setup...` item: Setup the page for printing.
- The `File→Print...` item: Print the main panel. item The `File→Extract Reads...` item: Extract reads for the selected nodes.
- The `File→Import CSV Format...` item: Load data in comma-separated-values (CSV) format: taxon,sum or readid,taxon,score.
- The `File→Import BIOME Format...` item is used to import data from QIIME, or from other systems that support the BIOME format, such as MG-RAST.
- The `File→Properties...` item: Show document properties.
- The `File→Close` item: Close the window.
- The `File→Quit` item: Exit the program (Windows and Linux only).

12.2 The Open Recent Menu

The `Open Recent` menu contains a list of recently opened documents.

12.3 The Export Menu

The `Export` menu contains the following items:

- The `Export→CSV Format...` item: Export assignments of reads to taxa to a CSV (comma-separated values) file.
- The `Export→BIOME Format...` item: Export data in BIOME format.
- The `Export→Taxonomic Paths...` item: Export taxonomic classification in an RDP-like path syntax.
- The `Export→Tree...` item: Export the phylogenetic tree induced by the data.
- The `Export→Reads...` item: Export reads to a FastA file.
- The `Export→Matches...` item: Export matches to a file.

- The `Export→Alignments...` item: Export all multiple sequence alignments associated with the selected nodes.
- The `Export→Summary...` item: Export as summary file.

12.4 The Edit Menu

The `Edit` menu contains the following items:

- The `Edit→Cut` item: Cut.
- The `Edit→Copy` item: Copy.
- The `Edit→Copy Image` item: Copy image to clipboard.
- The `Edit→Paste` item: Paste.
- The `Edit→Edit Node Label` item: Edit the node label.
- The `Edit→Edit Edge Label` item: Edit the edge label.
- The `Edit→Format...` item: Format nodes and edges.
- The `Edit→Find...` item: Open the Find toolbar.
- The `Edit→Find Again` item: Find the next occurrence.
- The `Edit→Preferences` submenu.

12.5 The Preferences Menu

The `Preferences` menu contains the following items:

- The `Preferences→Show Legend` item: Show legend identifying different datasets.
- The `Preferences→Edit Comparison Colors...` item: Edit the color palette used in comparison views.
- The `Preferences→Use Alternative Taxonomy...` item: Allows one specify an alternative taxonomy. For example, this allows one to use a Silva-based taxonomy.
- The `Preferences→Use Default NCBI Taxonomy` item: Switches back to the NCBI taxonomy shipped with MEGAN.

12.6 The Select Menu

The `Select` menu contains the following items:

- The `Select`→`All Nodes` item: Select all nodes.
- The `Select`→`None` item: Deselect all nodes.
- The `Select`→`From Previous Window` item: Select from previous window.
- The `Select`→`All Leaves` item: Select all leaves.
- The `Select`→`All Internal Nodes` item: Select all internal nodes.
- The `Select`→`All Intermediate Nodes` item: Select all intermediate nodes.
- The `Select`→`Subtree` item: Select subtree.
- The `Select`→`Leaves Below` item: Select allow leaves below.
- The `Select`→`Invert` item: Invert selection.
- The `Select`→`Level` submenu.

12.7 The Level Menu

The `Level` menu contains the following items:

- The `Level`→`Kingdom` item: Select Kingdom.
- The `Level`→`Phylum` item: Select Phylum.
- The `Level`→`Class` item: Select Class.
- The `Level`→`Order` item: Select Order.
- The `Level`→`Family` item: Select Family.
- The `Level`→`Genus` item: Select Genus.
- The `Level`→`Species` item: Select Species.

12.8 The Options Menu

The `Options` menu contains the following items:

- The `Options`→`Set Number Of Reads...` item: Set the total number of reads in the analysis.

- The `Options→Change LCA Parameters...` item: Rerun the LCA analysis with different parameters.
- The `Options→Taxon Disabling` submenu.
- The `Options→List Summary...` item: List summary of hits for selected nodes of tree.
- The `Options→List Path...` item: List path from root to selected nodes.
- The `Shannon-Weaver Index...→c`omputes the Shannon-Weaver diversity index on the set of selected nodes. (This is $-\sum p_i \log p_i$, where p_i is the proportion of reads assigned to node i.)
- The `Simpson-Reciprocal Index...→c`omputes the reciprocal Simpson diversity index on the set of selected nodes. (This is $1/\sum p_i^2$), where p_i is the proportion of reads assigned to node i.)
- The `Options→List Microbial Attributes...` item: List NCBI microbial attributes for selected microbes.
- The `Options→Compare...` item: Open compare dialog to produce a comparison of multiple datasets.
- The `Options→Reorder or Rename...` item: Change the order of names of the datasets in a comparison.
- The `Options→Open NCBI Web Page...` item: Open NCBI Taxonomy web site in browser.
- The `Options→Inspect...` item: Inspect the read to taxon assignments.

12.9 The Taxon Disabling Menu

The `Taxon Disabling` menu contains the following items:

- The `Taxon Disabling→Enable All` item: Enable all taxa.
- The `Taxon Disabling→Disable...` item: Disable all selected taxa. If none are selected, asks for taxa to disable (comma separated).
- The `Taxon Disabling→Enable...` item: Enable all selected taxa. If none are selected, asks for taxa to enable (comma separated).
- The `Taxon Disabling→List Disabled...` item: List all disabled taxa.

12.10 The Layout Menu

The `Layout` menu contains the following items:

- The `Layout→Expand/Contract` submenu.

- The `Layout→Layout Labels` item: Layout labels.
- The `Layout→Scale Nodes By Assigned` item: Scale nodes by number of reads assigned to taxon.
- The `Layout→Scale Nodes By Summarized` item: Scale nodes by number of reads assigned to and below a taxon.
- The `Layout→Set Max Node Radius...` item: Set the maximum node radius in pixels.
- The `Layout→Zoom To Selection` item: Zoom to the selection.
- The `Layout→Fully Contract` item: Contract tree vertically.
- The `Layout→Fully Expand` item: Expand tree vertically.
- The `Layout→Draw Circles` item: Draw data as circles.
- The `Layout→Draw Pies` item: Draw data as pie charts.
- The `Layout→Draw Heatmaps` item: Draw data as heat maps.
- The `Layout→Draw Bars` item: Draw nodes as bars.
- The `Layout→Cladogram` item: Draw the tree as a cladogram with all leaves aligned right.
- The `Layout→Phylogram` item: Draw the tree as a phylogram in which all edges have length one.
- The `Layout→Use Magnifier` item: Turn magnifier on and off.
- The `Layout→Draw Leaves Only` item: Only draw leaves.
- The `Layout→Highlight Differences` submenu.

12.11 The Expand/Contract Menu

The `Expand/Contract` menu contains the following items:

- The `Expand/Contract→Expand Horizontal` item: Expand view horizontally.
- The `Expand/Contract→Contract Horizontal` item: Contract view horizontally.
- The `Expand/Contract→Expand Vertical` item: Expand view vertically.
- The `Expand/Contract→Contract Vertical` item: Contract view vertically.

12.12 The Highlight Differences Menu

The `Highlight Differences` menu contains the following items:

- The `Highlight Differences→Uncorrected` item: In a comparison of exactly two datasets, highlight statistically significant differences, using no correction.
- The `Highlight Differences→Holm-Bonferroni Corrected` item: In a comparison of exactly two datasets, highlight statistically significant differences, using Holm-Bonferroni correction.
- The `Highlight Differences→Bonferroni Corrected` item: In a comparison of exactly two datasets, highlight statistically significant differences, using Bonferroni correction.

12.13 The Tree Menu

The `Tree` menu contains the following items:

- The `Tree→Collapse` item: Collapse selected nodes.
- The `Tree→Collapse at Level...` item: Collapse all nodes at given depth in tree.
- The `Tree→Collapse At Taxonomic Level` submenu.
- The `Tree→Uncollapse` item: Uncollapse selected nodes.
- The `Tree→Uncollapse Subtree` item: Uncollapse whole subtree beneath selected nodes.
- The `Tree→Show Taxon Names` item: Display the full names of taxa.
- The `Tree→Show Taxon Ids` item: Display the NCBI ids of taxa.
- The `Tree→Show Number of Reads Assigned` item: Display the number of reads assigned to a taxon.
- The `Tree→Show Number of Reads Summarized` item: Display the total number of hits to a taxon and its descendants.
- The `Tree→Node Labels On` item: Show labels for selected nodes.
- The `Tree→Node Labels Off` item: Hide labels for selected nodes.
- The `Tree→Show Intermediate Labels` item: Show intermediate labels at nodes of degree 2.

12.14 The Collapse At Taxonomic Level Menu

The `Collapse At Taxonomic Level` menu contains the following items:

- The `Collapse At Taxonomic Level→Kingdom` item: Collapse Kingdom.
- The `Collapse At Taxonomic Level→Phylum` item: Collapse Phylum.
- The `Collapse At Taxonomic Level→Class` item: Collapse Class.
- The `Collapse At Taxonomic Level→Order` item: Collapse Order.
- The `Collapse At Taxonomic Level→Family` item: Collapse Family.
- The `Collapse At Taxonomic Level→Genus` item: Collapse Genus.
- The `Collapse At Taxonomic Level→Species` item: Collapse Species.

12.15 The Window Menu

The `Window` menu contains the following items:

- The `Window→About...` item: Information about the program (Windows and Linux only).
- The `Window→How to Cite...` item: Show how to cite the program.
- The `Window→Website...` item: Go to the program website.
- The `Window→Register...` item: Register program for free.
- The `Window→Message Window...` item: Open the message window.
- The `Window→Set Window Size...` item: Set the window size.
- The `Window→Inspector Window...` item: Open inspector window.
- The `Window→Alignment Viewer...` item: Open alignment viewer for the selected nodes.
- The `Window→Main Viewer...` item: Brings the main viewer to the front.
- The `Window→SEED Analyzer...` item: Opens the SEED Analyzer window.
- The `Window→KEGG Analyzer...` item: Opens the KEGG Analyzer window.
- The `Window→Microbial Attributes Window...` item: Open Microbial Attributes window.
- The `Window→Chart...` item: Draw different types of charts.
- The `Window→Chart Microbial Attributes...` item: Chart attributes of all found microbes in datasets.

- The `Window→Network Comparison...` item: Open a network comparison window.
- The `Window→Rarefaction Analysis...` item: Compute and chart the species rarefaction curve.
- The `Window→Command-Line Syntax...` item: Shows the command-line syntax for commands relevant for current window.
- The `Window→Input Command...` item: Enter and execute a command.

12.16 The MEGAN Menu

Under MacOS, there is an additional, standard menu associated with the program, called the `MEGAN` menu. As usual, this contains the `Window→About...` and `File→Quit` menu items.

12.17 Wheel Mouse and Special Keys

Use of a wheel mouse is recommended for zooming of graphics displayed in different windows. The default is *vertical zoom*. For *horizontal zoom*, additionally press the shift key.

To scroll the graph, either press and drag the mouse (using the right mouse button), or use the arrow keys. To zoom the graph in verticial or horizontal direct, press the shift-key while using the arrow keys. To increase the zoom factor, additionally press the alt key or the control key.

To select a region of nodes using the mouse, click, hold for a second until the cursor changes to an arrow and then drag the mouse to capture the nodes to be selected.

13 SEED Window

The `SEED` window is used to display a SEED analysis of gene function, based on [12]. The SEED classification is displayed as a tree. Genes are mapped onto functional roles and these are present in one or more subsystems. Modes of interaction and available menu items are similar to those of the main window.

The window is split into two panes. The right pane contains a tree-based display of the result of the SEED classification. The left pane contains two tabs, one containing a textual tree-based view and the other using a heat-map style listing of the current leaf nodes of the tree displayed in the right pane.

14 KEGG Window

The `KEGG` window is used to display a KEGG analysis of gene function, based on [8]. The KEGG classification is displayed as a tree. Genes are mapped onto *enzymes* and these are present in one or more pathways. Modes of interaction and available menu items are similar to those of the main window.

The window is split into two panes. The right pane contains a tree-based display of the result of the KEGG classification. The left pane contains two tabs, one containing a textual tree-based view and the other using a heat-map style listing of the current leaf nodes of the tree displayed in the right pane.

Additionally, in the KEGG viewer the right pane of the window is tabbed. Initially, only the tree-based display of the KEGG classification is visible. However, by double-clicking on any item in the left pane for which a KEGG-pathway diagram exists, a new *pathway tab* is opened containing the corresponding pathway. Different shades of green are used to indicate how many reads were assigned to any given enzyme of gene-product in the pathway.

Another way to open a pathway tab is to use the following menu item, which is available in the `Options` menu and from context menus associated with nodes:

- The `Options`→`Show KEGG Pathway...` item: Show the specified KEGG pathway.

15 Network Window

The `Network` window provides methods for comparing multiple datasets. It is available after generating a comparison of multiple datasets, containing at least four datasets. The network window allows one to compute a distance matrix of the compared datasets using a number of different ecological indices. The calculation can be based on the results of a taxonomic, SEED or KEGG analysis. If no nodes are selected, then the distances will be based on the number of reads assigned to the current leaves of the analysis. If some nodes are selected, then only those nodes are used in the calculation.

The distance matrix can be visualized either using a split network calculated using the neighbor-net algorithm, or using a multi-dimensional scaling plot. See [10] for details.

16 Import Dialog

The `Import` dialog is used to import new data from BLAST (or a similar tool) and to create a new RMA file. The dialog has five tabbed panes.

The first tabbed pane titled the *Wizard pane* provides an *Import wizard* for creating a new RMA file. The user is first asked to specify a BLAST file (alternatively, the output of a number of other tools can also be specified), then a reads file and finally, the name of the new RMA file to be created. The program allows one to open more than one BLAST file or reads files, for the case that reads and matches are spread across multiple files. If the reads are from a paired-read project, then selecting the `Paired reads` check box will request MEGAN to perform a paired-read analysis (see [11]). Once this information has been collected, the user can press the *Apply* button to import the data.

The other four panes are for advanced settings.

The second tabbed pane titled the *Content pane* can be used to specify whether the SEED or KEGG content shall be analyzed, additional to an analysis of the taxonomical content.

The third tabbed pane titled the *Files pane* can be used to setup the location of files. The first two items are used to specify the location of the input files to be read, namely the BLAST file and the reads file. The third item is used to specify the location of the new RMA file. This pane provides two options. The *Max number of matches per read* file specifies how many matches per read to save in the RMA file. A small value will reduce the size of the RMA file, but may exclude some important matches. By default, the 100 highest scoring matches per read are save.

The fourth tabbed pane titled the *LCA Parameters pane* contains all items of the `Parameters` dialog which allows one to set the parameters used by the LCA algorithm. Because re-computation of an analysis can take quite long on a very large dataset, it is recommended to set these values at this stage.

The last tabbed pane titled the *Advanced Options pane* controls how MEGAN attempts to identify the taxon associated with a given BLAST hit. By default, MEGAN looks for the name of a taxon in the header line of the subject sequence, which is the fastest option.

The `Set Synonyms File` button can be used to load a file of customized synonyms to help identify taxa, e.g. *human* for *homo sapiens*. Each line of a *synonyms file* should contain two strings, separated by a tab, the synonym followed by the NCBI taxon name or id. The `Use Synonyms` check box item is used to turn the use of synonyms on and off.

The `Load GI-Lookup File` button can be used to load a file that maps *GI accession* numbers to taxon ids. Due to the large size of this lookup table, the file is preprocessed so as contain this data in a binary format suitable for direct access so that MEGAN does not need to read in the whole table. This file should be used when importing matches that do not contain the names of taxa in a text format. To use this feature, please download the file `gi_taxid_nucl.zip` or the file `gi_taxid_prot.zip` from the MEGAN website and then unzip the file. Please note that the unzipped file `gi_taxid_nucl.bin` or `gi_taxid_prot.bin` is over one gigabyte in size. The `Use GI Lookup` check box item is used to turn the use of this feature on or off.

MEGAN supports three different *text storage policies*. Select `Save in main file` to have all reads and BLAST matches embedded in the computed RMA file. This provides best portability of files. If the `Save in separate file` button is selected, then all reads and matches are stored in a separate *RMAZ* file. In this case, the RMA file will be much smaller and can be used independently of the RMAZ file, unless one wants to access the reads or matches, in which case the RMAZ file will be asked for. Finally, if the `Don't save` is selected, reads and matches are not stored explicitly. If they are requested by the user, then the program will obtain them from the original files. This mode leads to the smallest RMA files and shortest computation time, but is less portable.

17 Inspector Window

The `Inspector` Window can be used to inspect the alignments that are the basis of the assignment of reads to taxa. It can be opened either using the `Window→Inspector Window...` menu item or by right-clicking on a taxon and then selecting the `Inspect` popup item. This window displays data hierarchically using a data tree. The root node of this tree represents the current input file. This window can only be opened when data has been loaded into the program.

Any taxon added to the window, either by right-clicking a taxon and then selecting the `Inspect`

popup item in the main viewer, or by using the `Options→Show Taxon` item, is shown at a second level below the root. Clicking on such a *taxon node* will open a new level of nodes, each *read node* representing a read that has been assigned to the named taxon. Clicking on a read node will then open a new level of nodes, each such *read hit node* representing an alignment of the given read to a sequence associated with some taxon. Finally, double-clicking on a read hit node will display the actual BLAST alignment provided to deduce the relationship.

17.1 Inspector Menus

Here we describe those menu items that are different from the main window.

17.2 The Inspector Edit Menu

The `Edit` menu contains the following item:

- The `Edit→Show Taxon...` item: Show the named taxon and all reads assigned to it.

17.3 The Inspector Options Menu

The `Inspector Options` menu contains the following items:

- The `Options→Collapse` item: Collapse the selected nodes, or all, if none selected.
- The `Options→Expand` item: Ecxpand the selected nodes, or all, if none selected.
- The `Options→Ignore Hit` item: Ignore all selected hits.
- The `Options→Use Hit` item: Use all selected hits.
- The `Options→Use All Hits` item: Use all hits.

18 Microbial Attributes Window

The `Microbial Attributes Window` can be used to analyze the composition of microbial taxa (Bacteria and Archeae) and their various physiological features. Taxa have to be assigned with at least one read to be considered. The classification and its nomenclature is based on the NCBI's prokaryotic attribute table (derived from: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The window can be opened using the `Window→Microbial Attributes Window...` menu item when data has been loaded into the program.

19 Rarefaction Window

The `Rarefaction` Window be used to compute and draw a species rarefaction plot. This operates by repeatedly sampling subsets from a set of reads and computing the number of leaves to which

taxa have been assigned. This analysis uses the current leaves of the taxonomy, in other words collapsing or uncollapsing nodes will lead to a different result.

20 Taxon Chart Window

The `Taxon Chart Window` can be used to visualize the abundance distribution of the taxa as pie, bar, line chart, heatmap or word cloud. It can be opened using the `Window→Chart...` menu item. If nodes of the dataset have been selected in the main MEGAN window, they will be displayed directly in the chart. To change the taxa shown in the chart window, select them in the main window and then press the `sync` button.

21 SEED Chart Window

The `SEED Chart Window` can be used to visualize the abundance distribution of the SEED classes as pie, bar, line chart, heatmap or word cloud. It can be opened using the `Window→Chart...` menu item. If nodes of the dataset have been selected in the `SEED` window, they will be displayed directly in the chart. To change the nodes shown in the chart window, select them in the main window and then press the `sync` button.

22 KEGG Chart Window

The `KEGG Chart Window` can be used to visualize the abundance distribution of the KEGG classes as pie, bar, line chart, heatmap or word cloud. It can be opened using the `Window→Chart...` menu item. If nodes of the dataset have been selected in the `KEGG` window, they will be displayed directly in the chart. To change the nodes shown in the chart window, select them in the main window and then press the `sync` button.

23 Alignment Viewer

The `Alignment Viewer` can be used to compute and visualize a multiple sequence alignment of all reads that have significant matches to a reference sequences associated with a given taxon, SEED class or KEGG class. It can be opened using the `Window→Alignment Viewer...` menu item. Here is an overview of the menus available in this viewer and those menu items that do not appear in the main viewer. See [7] for details.

The File menu contains the following items:

- The `File→Save Alignment...` item: Save alignment to a file.
- The `File→Save Consensus...` item: Save consensus to a file.

The Edit menu contains the following items:

- The `Edit→Copy Alignment` item: Copy selected part of the alignment.
- The `Edit→Copy Consensus` item: Copy selected consensus sequence to clipboard.
- The `Edit→Copy Reference` item: Copy selected reference sequence to clipboard.
- The `Edit→Translate...` item: Translate DNA or cDNA sequence.

The Options menu contains the following items:

- The `Options→Show As Alignment` item: Show as alignment.
- The `Options→Show As Mapping` item: Show alignment as mapping.
- The `Options→Unsorted` item: Do not sort sequences.
- The `Options→Sort By Names` item: Sort sequences by names.
- The `Options→Sort By Start` item: Sort sequences by start positions.
- The `Options→Move Up` item: Move selected sequences up.
- The `Options→Move Down` item: Move selected sequences down.
- The `Options→Sort By Similarity` item: Sort sequences by pairwise similarity.
- The `Options→Set Amino Acid Colors...` item: Set the color scheme for amino acids.
- The `Options→Color Matches` item: Color letters that match the reference sequence.
- The `Options→Color Mismatches` item: Color letters that do not match the reference sequence.
- The `Options→Chart Diversity` item: Opens a chart showing a "diversity analysis" that aims at estimating the number of distinct genomic sequences corresponding to a given gene. Using a sliding window (default length 25) along the reference sequence, the program records the total number n of reads that cover the window and the number k of such reads that have distinct sequences over the window. These are then depicted in a scatter plot. Using a simple function based on Michaelis-Menten kinetics [15], the program plots a curve for the data that is used to estimate the number of different genomes involved. See [7] for details.

The Layout menu contains the following items:

- The `Layout→Show Insertions` item: Show insertions in reads.
- The `Layout→Contract Gaps` item: Contract all columns consisting only of gaps.
- The `Layout→Show Nucleotides` item: Show nucleotides in alignment.
- The `Layout→Show Amino Acids` item: Show amino-acids in alignment.
- The `Layout→Show Reference` item: Show reference sequence.
- The `Layout→Show Consensus` item: Show consensus sequence.
- The `Layout→Show Unaligned` item: Show the unaligned prefix and suffix of reads.

24 Find Toolbar

The `Find` toolbar can be opened using the `Edit→Find...` item. Its purpose is to find taxa, genes or other strings in a window. Use the following check boxes to parameterize the search:

- If the `Whole words only` item is selected, then only taxa or reads matching the complete query string will be returned.
- If the`Case sensitive` item is selected, then the case of letters is distinguished in comparisons.
- If the `Regular Expression` item is selected, then the query is interpreted as a Java regular expression.

Press the `Close`, `Find First` or `Find Next` buttons to close the toolbar, or find the first, or next occurrence of the query, respectively. Press the `Find All` button to find all occurrences of the query.

Press the `From File` button to load a set of queries, one per line, from a file.

25 Format Dialog

The `Format` dialog is opened using the `Edit→Format...` item. This is used to change the font, color, size and line width of all selected nodes and edges. Also, it can be used to turn labels on and off.

26 Message Window

The `Message` window is opened using the `Window→Message Window...` item. The program writes all messages to this window. The window contains the usual File and Edit menu items.

27 Parameters Dialog

The`Parameters` dialog is used to control the parameters of the LCA-assignment algorithm. It can be invoked by selecting `Options→Change LCA Parameters...` . The dialog options are:

- The `Min Support` item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result. Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support (version 3.10 onward, earlier versions counted the read as *unassigned*). By default, the minimum number of reads required for a taxon to appear in the result is 5.
- The `Min Score` item can be used to set a minimum threshold for the bit score of hits. Any hit in the input data set that scores less than the given threshold is ignored.

- The `Top Percentage` item can be used to set a threshold for the maximum percentage by which the score of a hit may fall below the best score achieved for a given read. Any hit that falls below this threshold is discarded.
- The `Win Score` item can be used to try and separate matches due to sequence identity and ones due to homology. If a win score is set, then, for a given read, if any match exceeds the win score, only matches exceeding the win score ("winners") are used to place the given read. The hope is that secondary, homology-induced matches are discarded in the presence of stronger primary matches. The `Min Complexity` item can be used to identify low complexity reads. These are placed on a special *Low Complexity* node. To turn this filter off, set the value to 0. A value of 0.3 catches most low complexity short reads.
- The `Paired Reads` item can be used to turn paired-read awareness of MEGAN on and off. In paired-read mode, MEGAN utilities read-pairing information to enhance the taxonomic assignment of reads.
- The `Use Percent Identity Filter` item can be used to turn on an additional filter for assigning reads to a specific taxonomic level. When this is active, the percent identity of a match must exceed the given value of percent identity to be assigned at the given rank: Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%, Phylum 80%. This should only be used when analyzing 16S rRNA sequences.

28 Compare Dialog

The `Compare` dialog is opened using the `Options→Compare...` item. This dialog provides a list of currently open datasets. To construct a comparison, select at least two different datasets and then press "ok". Select `Use absolute counts`, if you want the comparison the original counts of reads for each dataset. Select `Normalize over all reads`, if you want all counts to be normalized such that each dataset has 100,000 reads. Select `Ignore all unassigned reads`, if you want all reads assigned to the three special nodes labeled 'Not Assigned', 'No Hits' and 'Low Complexity' (if present) to be ignored. To change the order in which the datasets appear in the comparison, use the `Move up` and `Move down` buttons. To change the order of datasets or their names, as they appear in the window, select the `Options→Reorder or Rename...` item.

29 Extractor Dialog

This provides an alternative to the `Export→Reads...` item which allows to save reads from different taxa to files whose names contain the taxon name.

The `Extractor` dialog is opened using the `File→Extract Reads...` item. The dialog is used to extract all reads assigned to selected nodes. For any selected nodes, all reads assigned to it, or to *any node below* it in the hierarchy, are saved to a file.

Use the Browse button to specify the output directory. As the MacOS X dialog does not support the selection of a dialog, select any file inside the desired target directory. Specify the file name for output in the `File name` field. If the name contains `%t`, then the program will produce one output

file per node, and the name of the file is generated by replacing `%t` by the node name. Otherwise, all reads are written to one file.

30 Export Image Dialog

The `Export Image` dialog is opened using the `File→Export Image...` item. This dialog is used to save a picture of the current tree in a number of different formats, see Section 32.6.

The format is chosen from a menu. There are two radio buttons `Save whole image` to save the whole image, and `Save visible image` to save only the part of the image that is currently visible in the main viewer. If the chosen format is EPS, then selecting the `Convert text to graphics` check box will request the program to render all text as graphics, rather than fonts.

Pressing the apply button will open a standard file save dialog to determine where to save the graphics file.

31 About Window

The `About` Window is opened using the `Window→About...` item. It shows the program's splash screen.

32 File Formats

MEGAN uses its own file formats to store the data describing the result of a sequence comparison computation between a file of DNA reads and a database of reference sequences, such as computed by BLASTX, BLASTP or BLASTN [1].

32.1 RMA Files

Files ending in `.rma` are in a compressed binary format called RMA (read-match archive), which is a new open format that we will describe in a separate document. MEGAN 1 used a text format (files ending on `.megan` or `.meg`), which are now deprecated and will not be supported by futher versions of the program. By convention, we use the suffix `.megan` for MEGAN text files and `.rma` for binary read-match archive files.

With MEGAN 4, we have introduced a new version of the RMA format, internally known as RMA 2. This format is more flexible, as it does not necessarily need to contain all reads and matches. Moreover, it has better locality and thus updating it is much faster. MEGAN 4 can read RMA files generated by earlier versions of MEGAN, but not vice-versa.

A *RMA file* is generated using the `File→Import From BLAST...` menu item from a BLAST file and a *read file* (or from multiple files, if the reads are spread across multiple files). Depending on which of the three possible text storage policies is chosen, the RMA file may contain all reads

and matches in a compressed form, or these may be stored in a separate RMAZ file, or otherwise only links to there locations in the original reads and match files are stored.

In the first case, the size of such a file is around 10-20% of the size of the original input files and is thus usually smaller than the file that one obtains by simply compressing the BLAST file. The file is indexed and thus provides MEGAN with fast access to data stored in it. The reads and matches can be extracted from the file and so the MEGAN file provides a means of keeping all reads, BLAST matches and analysis in one document.

RMA is an open format which we will describe in a separate document.

32.2 The Text File Summary Format

As of version 4, the *MEGAN text file* format has been completely rewritten to accommodate SEED and KEGG classifications.

A MEGAN file starts with a number of header lines, each starting with a . These lines can occur in any order. This is best illustrated by an example:

```
 1 @Creator        MEGAN (version 4.0alpha20, built 14 Oct 2010)
 2 @CreationDate   Wed Oct 27 17:10:52 CEST 2010
 3 @ContentType    Summary4
 4 @Names  155_PE_1_fixed-paired   ecoli-testrun-2000-nr
 5 @Uids   1288068180866   1288190195887
 6 @Sizes  51246   2000
 7@TotalReads     200000
 8 @Collapse       SEED    2000041
 9 @Algorithm      Taxonomy        tree-from-summary
10 @Parameters     normalized_to=100000
11 @NodeStyle      KEGG    piechart
```

The first and second lines are optional descriptions of who generated the file when. The third line identifies the format as Summary4, indicating that this is a summary file in the format introduced with MEGAN 4. The fourth line lists the names of all datasets that are represented by this file. In this case there are two. Line 5 of this example lists the unique identifier numbers associated with the datasets, if any. Line 6 lists the original sizes of the datasets. Line 7 lists the total number of reads. This is not necessary the sum of the original sizes. Line 8 specifies, for the SEED classification, which nodes are to be collapsed in the visual representation of the classification. The keyword SEED can be replaced by TAXONOMY or KEGG for the other classifications. Line 9 contains the name of the algorithm used to compute a classification. The second word here is a keyword to identify which classification is meant. Line 10 lists parameters of the computation used to generate the file. Line 11 specifies the style used to draw nodes in a given classification, in this case KEGG.

The main body of a MEGAN text file contains multiple lines as follows:, in any order:

```
TAX     199310  0       1250
TAX     1       271     100
```

```
TAX     28216   35
TAX     32523   8
TAX     2       8336    1350
KEGG    7716    12
KEGG    3859    2
KEGG    7714    2       100
SEED    54      6       50
```

The general format is *classification*, *count-1*, *count-2*, ..., *count-n*. Here, *classification* is either `TAX` for taxonomy, `SEED` or `KEGG`. This is followed by a number indicating a class in the given classification. In the case of taxonomy, this is the NCBI taxonid. This is followed by up to n numbers, where n is the number of datasets mentioned in the header, indicating how many reads in the 1-st, 2-nd etc dataset were assigned to the given class.

Because this format is also embedded in RMA files to provide a summary of the data, when opening an RMA file generated by an earlier version of MEGAN, the program automatically updates the summary to the new format. As a consequence, any RMA file that has been opened by MEGAN 4 cannot later be opened by an earlier version of the program.

32.3 Required Syntax of BLAST Files

MEGAN imports data from a *BLAST file*. MEGAN can parse BLAST files in standard or XML format obtained using the BLAST output option `-m 0` or `-m 7`, respectively. MEGAN can also parse tabular format (BLAST output option `-m 8`). For this to work, the subject field must either contain taxon names or GI accession numbers. In the latter case, please use the `Load GI-Lookup File` button to load a GI lookup file. Alternatively, a nine column may be supplied that contains either the taxon name or taxon Id associated with the database sequence. The program also scans the subject field for RefSeq identifiers to determine the associated gene.

MEGAN can read *gzipped BLAST files*.

For human readable format, any *BLASTX file* or *BLASTP file* is expected to adhere to the format shown in Figure 1. Any *BLASTN file* is expected to adhere to the format shown in Figure 2.

32.4 How MEGAN Parses Taxon Names

MEGAN uses the following algorithm to determine the taxon from the header line of a reference sequence. If the string consists only of an integer, then this is interpreted as a taxon id. Otherwise, if `Use Synonyms` is turned on, then MEGAN attempts to match an entry in the given synonyms file. The longest matching synonym is used to determine the taxon. Otherwise, if `Use GI Lookup` is turned on, then MEGAN searches for an occurrence of the string `gi|` followed by a number and tries to use the number as a GI accession to determine the taxon.

Otherwise, if the header line contains a semi-colon, then MEGAN assumes that a list of taxon names is given, e.g. `Bacteria;Proteobacteria; Alpha proteobacteria`, as present, for example, in the Silva database. In this case, MEGAN uses the right-most name to determine the taxon id.

Otherwise, if the header line contains the text `/TAXON_ID=`, then MEGAN will attempt to read

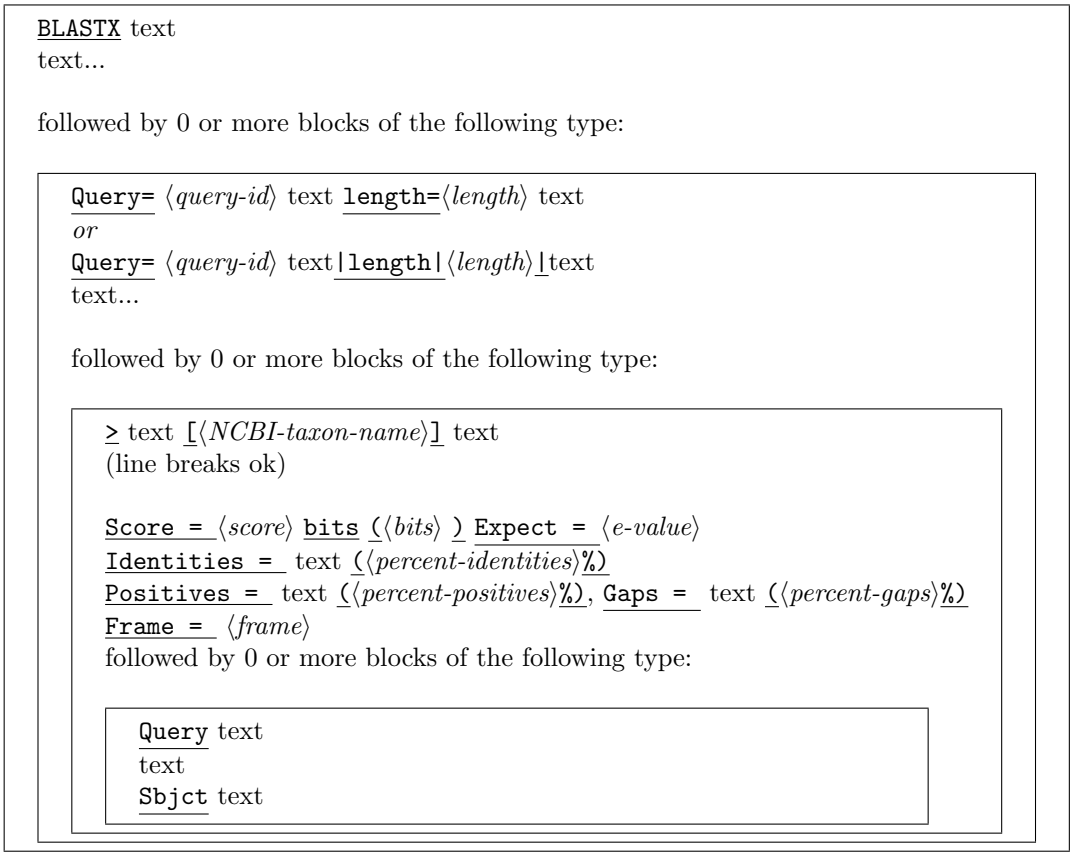

Figure 1: The required structure of a BLASTX file. Labels shown as `label` are tokens that must occur verbatim in the file. Labels shown as ⟨*label*⟩ are values that are read into the program. The first word in the file must be `BLASTX`. The header line starting with `Query =`, which is taken from the Fasta header of the query sequence (a read), must start with a one word unique identifier for the read and must also contain a statement containing the length of the read, in the format `length=`⟨*length*⟩, or as `|length|`⟨*length*⟩`|`. Another important feature is that the comment line of the database sequence must contain a NCBI-taxon name. If names are not contained in the comment lines, then the accession lookup support must be used. Finally, the `Gaps=` statement is optimal.

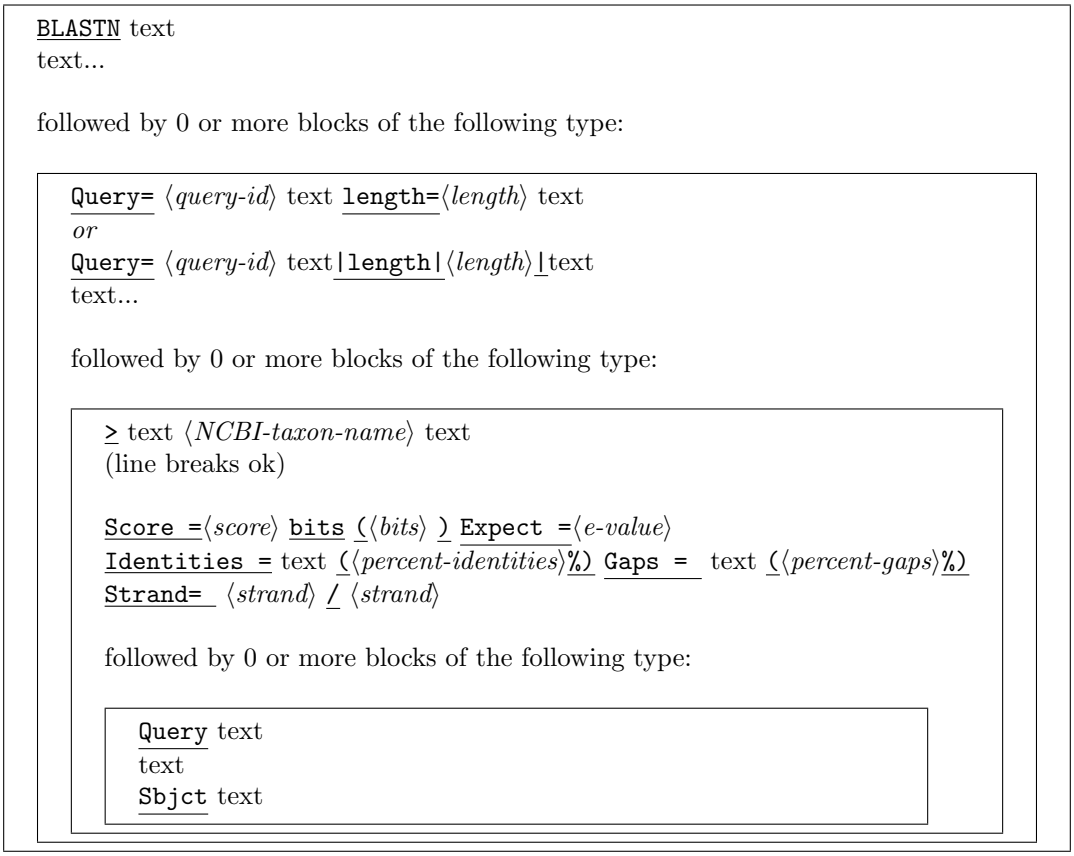

Figure 2: The required structure of a BLASTN file. Labels shown as `label` are tokens that must occur verbatim in the file. Labels shown as ⟨*label*⟩ are values that are read into the program. The first word in the file must be `BLASTN`. The header line starting with `Query=`, which is taken from the Fasta header of the query sequence (a read), must start with a one word unique identifier for the read and must also contain a statement containing the length of the read, in the format `length=`⟨*length*⟩. Another important feature is that the comment line of the database sequence must contain a NCBI-taxon name. If names are not contained in the comment lines, then the accession lookup support must be used.

a taxon id following the text. This syntax is used in BLAST files obtained from the CAMERA website.

Otherwise, MEGAN searches for all pairs of disjoint square brackets and attempts to parse the strings between such brackets to obtain a set of taxon ids. The taxon id for the match is then set to the LCA of the ids. (In the NCBI-NR database, names of taxa are placed between square brackets.)

Otherwise, MEGAN searches for maximal and non-overlapping substrings that can be mapped onto an NCBI taxon id. Again, the taxon id of the match is set to the LCA.

Otherwise, the taxon is set to `unknown''`.

32.5 Required Format of Read Files

Reads from sequencing are assume to be provided in multi-FastA format in a *reads file*. The first word of a FastA header is assumed to be the read-id. The remaining text of the FastA header must contain the length of the read either as `length=`*number*, or as `|length|`*length*—.

32.6 Graphics Formats

The following graphics formats are supported:

- BMP, "Bitmap".
- EPS, "Encapsulated PostScript", vector format.
- GIF, "Graphics Interchange Format".
- JPEG, "Joint Photographic Experts Group".
- PDF, "Portable Document Format", vector format.
- PNG, "Portable Network Graphics".
- SVG, "Scalable Vector Graphics", vector format.

32.7 CSV File Format

MEGAN supports importing data from other programs in a comma-separated format from a *CSV file*, using the `File→Import CSV Format...` menu item. The input file must be a text file in which either all lines each contain two strings that are separated by a comma. or all lines each contain three strings separated by commas.

Importing read assignments If each line of the CSV file contains two strings separated by a comma, then the first string will be intepreted as a taxon name or taxon id and the second string will be intepreted as an integer specifying the number of reads assigned to the named taxon.

MEGAN will assume that this is the result of some analysis and thus will produce a summary file from it and will simply display it on the NCBI taxonomy with no further analysis.

For example, assume that you have performed a metagenome analysis using some other method and have obtained the following result:

- Gammaproteobacteria: 55 reads
- Mollicutes: 400 reads
- Escherichia coli K-12: 42 reads
- Unknown: 100 reads

To import this data into MEGAN so as to visualize the taxonomical assignments, produce the following CSV file:

```
Gammaproteobacteria, 55
Mollicutes, 400
Escherichia coli K-12, 42
Not assigned, 100
```

MEGAN will draw a tree with four nodes, one for each of the named taxa.

Importing read matches Otherwise, if each line of the CSV file contains three strings separated by a comma, the first string will be interpreted as a read id, the second one as a taxon name or id and the third one will be interpreted as a bit score for this assignment. MEGAN will assume that this data describes a collection of reads and their matches. This data will be analysed using the LCA algorithm and the result will be displayed on the NCBI taxonomy.

For example, assume that you have done a database search using some other method than BLAST and have obtained the following result:

- The read r01 matches *Escherichia coli CFT073* with a bitscore of 100,
- The read r01 matches *Escherichia coli K-12* with a bitscore of 110, and
- The read r01 matches *Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67* with a bitscore of 120.
- The read r02 matches *Caldicellulosiruptor saccharolyticus DSM 8903* with a bitscore of 90.

To import this data into MEGAN so as to analyze is using the LCA algorithm, produce the following CSV file:

```
r01, Escherichia coli CFT073, 100
r01, Escherichia coli K-12, 110
r01,  Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67,120
r02, Caldicellulosiruptor saccharolyticus DSM 8903, 90
```

MEGAN can also import SEED or KEGG counts. In addition, MEGAN is able to map entries consisting of a RefSeq Id and counts to KEGG or SEED.

32.8 Tree and Map Format

The NCBI taxonomy is loaded by MEGAN at startup. It is contained in a *NCBI tree file* in the standard Newick tree format. The mapping from taxon-IDs to taxon names is loaded by MEGAN at startup. It is contained contained in a *NCBI mapping file* in a line based format in which each has three entries: taxon-ID, taxon name and then a number indicating the size of the genome, or -1, if the size is unknown.

33 Command-Line Options

MEGAN has the following *command-line* options:

```
Program usage:
+g <switch>          (default=true): Use GUI
-f <String>          (default=""): MEGAN file (separate multiple files using '|')
-fs <String>         (default=""): Synonyms file
-fg <String>         (default=""): GI lookup file
-p <String>          (default="/Users/huson/Library/Preferences/Megan.def"): Properties file
-m <int>             (default=0): minimum score
+w <switch>          (default=true): show message window
-x <String>          (default=""): Execute this command at startup
-E <switch>          (default=false): Quit if exception thrown in non-gui mode
-V <switch>          (default=false): show version string
-S <switch>          (default=false): silent mode
-d <switch>          (default=false): debug mode
+s <switch>          (default=true): show startup splash screen
-h <switch>          (default=false): Show usage
```

Launching the program with option `+g` will make the program run in non-GUI *command-line mode* , first excuting any command given with the `-x` option and then reading additional commands from standard input.

Please be aware that the command-line version of the program uses the same *properties file* as the interactive version. So, any *preferences* set using the interactive version of the program will also apply to the command-line version of the program. It this is not desired, then please use the `-p` option to supply a different properties file.

Another important thing to note is that the command-parser operates in a line-by-line fashion. When processing commands in a given line, the parser makes note of required updates to the taxonomy and data-structures. These updates are not executed until all commands in the current input line have been processed. For example, if you want to open and MEGAN file and then to save a picture of the taxonomical analysis in a PDF file, then the two commands should be entered on separate lines because otherwise the taxonomy will be drawn before the data from the MEGAN file has been processed. Here is an example of the correct way to produce a picture of a taxonomic analysis:

```
open file='/Users/huson/data/megan/x.rma'
exportimage file='/Users/huson/data/megan/x.pdf'format=PDF replace=true
quit
```

Alternatively, the `update` command can be used to explicitly force MEGAN to update all data-structures, in this case the commands show appear together on one line, e.g.:

```
open file='x.rma';update;exportimage file='x.pdf' format=PDF replace=true;
```

As described below, the `update` command takes a number of different parameters that can be used to determine exactly what type of update is required.

All commands supplied using the `-x` command-line option are parsed as if they were contained in one line. So, here the `update` command must be used to ensure that commands are completed when necessary. To open a file, print the taxonomical analysis and then close the file using the `-x` option, enter the following:

```
-x "open file='x'; update;exportimage file='x.pdf' format=PDF replace=true; quit;"
```

To save an image of a SEED analysis, after loading a file, we must then open the SEED viewer, change the command context to the seed viewer, then configure the size of the window and how much of the tree to uncollapse. Then we save the image file. Here is an example:

```
open file='/Users/huson/data/megan/x/x.rma'
show window=seedviewer
set context=seedviewer
set windowsize=1000 x 1000
select nodes=all
uncollapse subtrees
exportimage file='/Users/huson/data/megan/x/x.pdf'format=PDF replace=true
quit
```

34 Command-Line Commands

Command processing has been completely rewritten for MEGAN 4. Each type of window that can be opened by MEGAN has its own command interpreter. Initially, on startup the program will open a Main window and all commands piped to the program will be executed using the command interpreter associated with the main window. The main window provides a number of commands for opening other windows. For example, the command `show window=seedviewer;` will open the SEED classification viewer. To pipe commands to the SEED viewer, the *command context* has to be set to the SEED viewer, by entering `set context=seedviewer;`. After entering this command, all subsequent commands are handled by the interpreter associated with the SEED viewer. To obtain a list of all commands available for the current interpreter, enter `help;`. In obtain help on a particular command, for example on *export*, enter `help export;`. All command description lines that contain the word "export" (case insensitive) will be listed.

In the following we list all commands available in the Main viewer. Other viewers support many of these commands, too, but also other, viewer-specific ones. To determine which commands are available for a given window, run MEGAN in GUI mode, open the window of interest and then select the Window→Command-Line Syntax... item to obtain a listing of all commands available for the given window. Here are the commands that are available in the Main viewer:

```
Available commands (context=mainviewer):
File menu:
new; - Open a new empty document
open file=<filename> [readonly={false|true}] [fixlinks={true|false}]; - Open a MEGAN file (ending on .rma, .meg or .megan)
import blastfile=<name>[,<name>,<name>,...] [fastafile=<name>[,<name>,<name>,...]] meganfile=<name> [maxmatches=<num>]
[minscore=<num>] [toppercent=<num>] [winscore=<num>] [minsupport=<num>]
        [mincomplexity=<num>] [useseed={true|false}] [usekegg={true|false}] [paired={false|true}
[suffix1=<string> suffix2=<string>]] [textstoragepolicy={0|1|2}]
  [blastformat={GUESS|BLASTX|BLASTN|BLASTP|BLASTXML|BLASTTAB|RDP-Assignment-Detail|RDP-Standalone|SILVA|SAM}];
- Import BLAST (or RDP or Silva or SAM) and reads files to create a new MEGAN file
save file=<filename> [summary={false|true}]; - Save current data set
exportimage file=<filename> [format={eps|svg|gif|png|jpg|pdf}] [replace={false|true}] [textasshapes={false|true}];
- Export content of window to an image file
show window=pagesetup; - Setup the page for printing
show window=print; - Print the main panel
extract what=document file=<megan-filename> [sparsefile={false|true}] [data={Taxonomy|SEED|KEGG}] [ids=<numbers...>]
[names=<names...>] [allbelow={false|true}];
- Extract all reads and matches on or below selected node(s) to a new document
extract what=reads outdir=<directory> outfile=<filename-template> [data={Taxonomy|SEED|KEGG}]
[ids=<SELECTED|numbers...>] [names=<names...>] [allbelow={false|true}]; - Extract reads for the selected nodes
import csv={reads|summary} separator={comma|tab} file=<fileName> [toppercent=<num>] [taxonomy={true|false}]
[seed={false|true}] [kegg={false|true}] [useRefSeq={false|true}] [minscore=<num>] [minsupport=<num>];
- Load data in comma-separated-values (CSV) format: READ_NAME,CLASS-NAME,SCORE or CLASS,COUNT(,COUNT...)
import format=biome file=<fileName>; - Import data from a table in BIOME format
show window=properties; - Show document properties
close; - Close the window

Export sub-menu:
export what=CSV format={readname_taxonname|readname_taxonid|readname_taxonpath|taxonname_count|taxonpath_count|taxonid_count|taxonname_readname|
taxonpath_readname|taxonid_readname|taxonname_length|taxonpath_length|taxonid_length|readname_refseqid|readname_ seedname|
readname_seedpath|seedname_count|seedpath_count|seedname_length|seedpath_length|seedname_readname|seedpath_readname|
readname_keggname|readname_keggpath|keggname_count|keggpath_count|keggname_length|keggpath_length|keggname_readname|keggpath_readname}
separator={comma|tab} file=<filename>;
- Export assignments of reads to nodes to a CSV (comma-separated values) file
export what=reads [data={Taxonomy|SEED|KEGG}] file=<filename>; - Export all reads to a text file (or only those for selected nodes, if any selected)
export what=matches [data={Taxonomy|SEED|KEGG}] file=<filename>; - Export all matches to a text file (or only those for selected nodes, if any selected)

Edit menu:
show window=formatter; - Format nodes and edges
show findtoolbar={true|false}; - Open the Find toolbar

Preferences sub-menu:
set db=<string> user=<string> password=<string>; - Set postgres database name and user authorization
set showlegend={true|false}; - Show legend identifying different datasets

Select menu:
select nodes=all; - Select all nodes
select nodes=none; - Deselect all nodes
select nodes=previous; - Select from previous window
select nodes=leaves; - Select all leaves
select nodes=internal; - Select all internal nodes
select nodes=intermediate; - Select all intermediate nodes
select nodes=subtree; - Select subtree
select nodes=subleaves; - Select allow leaves below
select nodes=invert; - Invert selection

Level sub-menu:
select rank=Kingdom; - Select Kingdom
select rank=Phylum; - Select Phylum
select rank=Class; - Select Class
select rank=Order; - Select Order
select rank=Family; - Select Family
select rank=Varietas; - Select Varietas
select rank=Genus; - Select Genus
select rank=Species_group; - Select Species_group
select rank=Subspecies; - Select Subspecies
select rank=Species; - Select Species

Options menu:
recompute [minsupport=<number>] [minscore=<number>] [toppercent=<number>] [winscore=<number>] [mincomplexity=<number>]
[pairedreads={false|true}] [useseed={false|true}] [usekegg={false|true}]; - Rerun the LCA analysis with different parameters
set totalreads=<num>; - Set the total number of reads in the analysis (will initiate recalculation of all classifications)
list summary={all|selected}; - List summary of hits for selected nodes of tree
compare mode={absolute|relative|merge}[ignore_unassigned={false|true}] [pid=<number>,...] [meganfile=<filename>,...];
     - Open compare dialog to produce a comparison of multiple datasets
set order=<number> <number>...; - Change the order of datasets in a comparison view
show window=colorpalette; - Edit the color palette used in comparison views
```

```
show webpage taxon=<name|id>; - Open NCBI Taxonomy web site in browser
inspector taxa=selected; - Inspect the read-to-taxon assignments

Taxon Disabling sub-menu:
enable taxa=all; - Enable all taxa
disable taxa={selected|<name,..>}; - disable all selected taxa or the named ones
enable taxa={selected|<name,...>}; - enable all selected taxa or the named ones
list taxa=disabled; - List all disabled taxa

Layout menu:
set autolayoutlabels={true|false}; - Layout labels
set scaleby=assigned; - Scale nodes by number of reads assigned to taxon
set scaleby=summarized; - Scale nodes by number of reads assigned to and below a taxon
set maxnoderadius=<num>; - Set the maximum node radius in pixels
set zoom=selected; - Zoom to the selection
set zoom=fit; - Contract tree vertically
set zoom=full; - Expand tree vertically
set nodedrawer=circle; - Draw data as circles
set nodedrawer=piechart; - Draw data as pie charts
set nodedrawer=heatmap; - Draw data as heat maps
set nodedrawer=barchart; - Draw nodes as bars
set drawer={Cladogram,Phylogram}; - Draw tree as cladogram with all leaves aligned right
set drawleavesonly={true|false}; - Only draw leaves

Expand/Contract sub-menu:
expand direction=horizontal; - Expand view horizontally
contract direction=horizontal; - Contract view horizontally
expand direction=vertical; - Expand view vertically
contract direction=vertical; - Contract view vertically

Highlight Differences sub-menu:
set highlightdifferences={true|false} [correction={none|bonferroni|holm_bonferroni}]; - In a comparison of exactly two
datasets, highlight statistically significant differences, using no correction
set comparison_highlight_color=<number>; - Set the pairwise comparison highlight color

Tree menu:
collapse nodes=selected; - Collapse selected nodes
collapse level=<num>; - Collapse all nodes at given depth in tree
uncollapse nodes={all|selected|subtree}; - Uncollapse selected nodes
nodelabels names={true|false}; - Display the full names of taxa
nodelabels ids={true|false}; - Display the NCBI ids of taxa
nodelabels assigned={true|false}; - Display the number of reads assigned to a taxon
nodelabels summarized={true|false}; - Display the total number of hits to a taxon and its descendants
show labels=selected; - Show labels for selected nodes
hide labels=selected; - Hide labels for selected nodes
show intermediate=<bool>; - Show intermediate labels at nodes of degree 2

Collapse At Taxonomic Level sub-menu:
collapse rank=Kingdom; - Collapse Kingdom
collapse rank=Phylum; - Collapse Phylum
collapse rank=Class; - Collapse Class
collapse rank=Order; - Collapse Order
collapse rank=Family; - Collapse Family
collapse rank=Varietas; - Collapse Varietas
collapse rank=Genus; - Collapse Genus
collapse rank=Species_group; - Collapse Species_group
collapse rank=Subspecies; - Collapse Subspecies
collapse rank=Species; - Collapse Species

Window menu:
show window=howtocite; - Show how to cite the program
show window=website; - Go to the program website
show window=register; - Show registration window
show window=message; - Open the message window
set windowsize=<width> x <height>; - Set the window size
show window=inspector; - Open inspector window
show window=mainviewer; - Brings the main viewer to the front
show window=seedviewer; - Opens the SEED Analyzer
show window=keggviewer; - Opens the KEGG Analyzer
show chart data={taxonomy|SEED|KEGG|attributes}; - Chart assigned reads
show wordCloud data={taxonomy|SEED|KEGG|attributes}; - WordCloud based on assigned reads
show window=network; - Open a network comparison window
show rarefaction data={taxonomy|seed|kegg}; - Compute a rarefaction curve
help [keyword]; - Shows syntax help for commands

Additional commands:
exportimage-old file=<filename> [format={eps|svg|gif|png|jpg|pdf}] [replace={false|true}] [textasshapes={false|true}];
- Export content of window to an image file
list assignments; - List the number of reads assigned to each level of the taxonomy
load colorfile=<filename>;- Load dataset colors from a file (format: one RGB color per line)
load gi2taxfile=<filename>; - Load the GI mapping file gi_taxid_nucl.bin, downloaded from the MEGAN website
load synonymsfile=<filename>; - Load a file of taxon-name synonyms
load treefile=<filename> [mapfile=<filename>; - Load the taxonomy .tre and .map files (e.g. ncbi.tre and ncbi.map)
mp-analyzer what={lca-ranks|compare} infile=<filename> outfile=<filename>; - Compute the rank at which the LCA is found for each mate-pair, or preprocess comparison
quit; - Quit the program
replacelinks [old=<filename> new=<filename>] [...]; - Replace links to source files
select ids=<ids...>; - Select the nodes for the given ids
select name=<names...>; - Select the named nodes
```

```
set context=<window-name>; - Choose command context, i.e. the window that should parse the subsequent commands
set dir=<directory> - Set the current directory
set margin [left=<number>] [right=<number>] [bottom=<number>] [top=<number>]; - Set margins used in tree visualization
set proxy=<string> port=<number> user=<string> password=<string>; - Set proxy credentials
set scaleby=none; - Do not scale nodes
set usekegg={true|false}; - Turn KEGG analysis on or off
set usepercentidentity={false|true}; - Adjust assignment based on best percent identity of matches, using the following minimum requirements:
Species 97%, Genus 95%, Family 90%, Order 85%, Class 80%, Phylum 75%
set useseed={true|false}; - Turn SEED analysis on or off
setprop <name>=<value>; - Set a property
show chart=taxavsseed; - Chart taxa vs SEED
show histogram taxonid=<num>; - Shows the distribution of matches for a given taxon
show window=about; - About MEGAN and the authors
show window=checkforupdate; - Check for an update of the program
show window=cogs; - Open COG window
show window=comparisonstats; - Open dialog to produce a statistical comparison of two datasets
show window=fixlinks; - Fix missing links to source BLAST and reads files
show window=webservice; - Open metagenomic files from the MEGAN-DB website
tofront; - Bring window to front
update [reprocess={false|true}] [reset={false|true}] [reinduce={false|true}]; - Update data. If nothing specified, assumes reinduce=true
version; - Show version info
```

35 Examples

Example files can be downloaded from the MEGAN website.

36 Using More Memory

The MEGAN installer allows you to specify the amount of MEGAN that the program can use. We recommend at least 2 GB on a 64-bit machine and recommend 8 GB on a desktop.

MEGAN is a memory-hungry application. When importing BLAST files, we recommend that you use a machine that allows you to run MEGAN with at least 4 GB of main memory. Using less memory will work, but Java will be forced to perform frequent garbage collection, which will slow the program down. Also, because the program is i/o intensive, it is best to have all files on local disks, as this will increase the speed of the program.

To run MEGAN with more than 2GB under MacOS X on an intel Mac, edit the file `/Applications/MEGAN/MEGAN.app/Contents/Info.plist` as follows: Find the lines

```
<key>VMOptions</key>
<string>-server -Xms2000M -Xmx2000M </string><!-- I4J_INSERT_VMOPTIONS -->
```

and replace them by:

```
<key>VMOptions</key>
<string>-server -Xms2000M -Xmx8000M </string><!-- I4J_INSERT_VMOPTIONS -->
```

to run using 8 gigabytes, for example.

To run MEGAN with more than 2GB on a 64-bit unix/linux system, open the file ⟨*installation-dir*⟩`/MEGAN.vmoptions` in a text editor. Find the current memory specification (e.g. `-Xmx1600M`) and replace it by `-Xmx8G` to run with 8 gigabytes of memory, say.

37 Acknowledgments

This program uses the following Java libaries: BrowserLauncher2-10rc4, Jama-1.0.2, MRJAdapter, axis, batik, colt, h2, jcommon-1.0.16, and postgresql-9.0-801.jdbc4. These libaries, and their licenses, are located in the `jars` folder of the MEGAN installation directory.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[2] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. Genbank. *Nucleic Acids Res*, 1(33):D34–38, 2005.

[3] J. Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, Antonio G. Pena, Julia K. Goodrich, Jeffrey I. Gordon, Gavin A. Huttley, Scott T. Kelley, Dan Knights, Jeremy E. Koenig, Ruth E. Ley, Catherine A. Lozupone, Daniel McDonald, Brian D. Muegge, Meg Pirrung, Jens Reeder, Joel R. Sevinsky, Peter J. Turnbaugh, William A. Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, April 2010.

[4] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Research*, 37(suppl 1):D141–D145, January 2009.

[5] Elizabeth M. Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, 2010(1):pdb.prot5368+, January 2010.

[6] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–386, March 2007.

[7] Daniel H. Huson and Chao Xie. Reference-guided multiple sequence alignment of metagenomic data. Under review, 2012.

[8] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map (SAM) format and SAMtool. *Bioinformatics*, 25:2078–9, 2009.

[10] S. Mitra, J.A. Gilbert, D. Field, and D.H. Huson. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J*, 2010. doi:10.1038/ismej.2010.51.

[11] Suparna Mitra, Max Schubach, and Daniel H Huson. Short clones or long clones? a simulation study on the use of paired reads in metagenomics. *BMC Bioinformatics*, 11(Suppl 1):S12+, 2010.

[12] Ross Overbeek, Tadhg Begley, Ralph M Butler, Jomuna V Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards, Michael Fonstein, Ed D Frank, Svetlana Gerdes, Elizabeth M Glass, Alexander Goesmann, Andrew Hanson, Dirk Iwata-Reuyl, Roy Jensen, Neema Jamshidi, Lutz Krause, Michael Kubal, Niels Larsen, Burkhard Linke, Alice C McHardy, Folker Meyer, Heiko Neuweger, Gary Olsen, Robert Olson, Andrei Osterman, Vasiliy Portnoy, Gordon D Pusch, Dmitry A Rodionov, Christian Rückert, Jason Steiner, Rick Stevens, Ines Thiele, Olga Vassieva, Yuzhen Ye, Olga Zagnitko, and Veronika Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005.

[13] Hendrik N Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D E Macphee, Bernard Buigues, Alexei Tikhonov, Daniel H Huson, Lynn P Tomsho, Alexander Auch, Markus Rampp, Webb Miller, and Stephan C Schuster. Metagenomics to paleogenomics: large-scale sequencing of mammoth dna. *Science*, 311(5759):392–394, Jan 2006.

[14] E. Pruesse, C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F.O. Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nuc. Acids Res.*, 35(21):7188–7196, 2007.

[15] Wikipedia. Michaelis-Menten kinetics. http://en.wikipedia.org/wiki/Michaelis--Menten_kinetics, 2012.

[16] Yongan Zhao, Haixu Tang, and Yuzhen Ye. Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, 2012.

Index