Departament de Llenguatges i Sistemes Informatics
UNIVERSITAT POLITÈCNICA DE CATALUNYA

M-GCAT User Guide

Todd J Treangen[1*] , Xavier Messeguer[1]

[1]ALGGEN, Algorithmics and Genetics Group, Department of Software, Technical University of Catalonia, Barcelona, Spain

Email: treangen@lsi.upc.edu; peypoch@lsi.upc.edu;

[*]Corresponding author

Contents

List of Figures

Release Notes

Version 1.30 was released on March 1st, 2008. The main features/changes/fixes in version 1.30 include:

- Tabbed run support for performing multiple comparisons in the viewer interface. Runs can be individually saved, renamed, and deleted. A maximum of 10 runs at a time can be used.
- New "Run log" tab, organizing the input/output for each Run.
- Buttons for easy activation/deactivation of windows inside of each workspace.
- Blastall query interface has been modified to facilitate BLAST searches on the database create of the input sequences. Genome positions can be directly specified, or sequence can be copy & pasted into the input window. Common BLAST parameters (such as e-value cutoff and word size) can be easily configured.
- Blastall queries are now visually displayed and can be inspected interactively.

1 Introduction to M-GCAT

M-GCAT is an acronym for **M**ultiple **G**enome **C**omparison and **A**lignment **T**ool, pronounced \. 'em,je,kat\. . In addition, the nucleotides in the program name were arranged in homage to the birthplace of the program, at the Technical University of `CAT`alonia (http://www.upc.cat). M-GCAT was originally developed to provide an interactive graphical user interface for efficient multiple genome comparisons based on maximal unique matches (MUMs) and has evolved into a multiple genome alignment tool. M-GCAT is based around a novel suffix graph data structure used to rapidly search for common unique matches on both strands, allowing M-GCAT to simultaneously compare and build alignment frameworks for hundreds of microbial-sized genomes via an anchored alignment approach. While optimized for genome comparisons involving closely-related prokaryote species, M-GCAT is also able to efficiently compare metazoan chromosomes by partitioning the input chromosomes & genomes into several smaller parts. M-GCAT doesn't require the order of the alignment anchors to be collinear, and therefore is able to detect large-scale rearrangements such as translocations and inversions, in addition to large-scale insertions and deletions. In summary, M-GCAT is a useful starting and ending point for generating multiple genome comparisons among closely-related prokaryote species.

2 Installation

A compiled and packaged version of M-GCAT is available for Windows(32bit), Linux(32bit), Solaris Sparc(64bit), and Mac OS X.

2.1 Windows 32bit

For Windows no Python installation is necessary, simply unzip the provided files and run.

2.2 Other platforms

-For all other platforms, a working Python installation(2.3 or newer) and TCL/TK (8.3 or newer) is necessary.

2.3 Required software

Where to obtain the required software:

M-GCAT : http://alggen.lsi.upc.es/recerca/align/mgcat/intro-mgcat.html

MUSCLE [1]: http://www.drive5.com/muscle

Python [2]: http://www.python.org/download

TCL/TK [3]: http://www.tcl.tk/

2.4 System Requirements

M-GCAT was written in C++ and Python, and has been compiled in Windows, Linux, Mac OS X, and Solaris, and is readily portable to other platforms. M-GCAT requires a minimum of 10 megabytes of disk space for installation, and it is recommended to have at least 1024 MB RAM available when comparing genomes.

[1] Version 3.6 or newer
[2] Version 2.3 or newer
[3] Version 8.3 or newer

3 Quick start

3.1 Running M-GCAT in Windows 98,2000,XP,Vista

1. To start the main M-GCAT application, run the program executable 'viewer.exe'.

2. To run M-GCAT from the command line, run `mgcat.exe <inifile>`, where inifile is the automatically generated mgcat.ini or other properly formatted parameter file.

3.2 Running M-GCAT in other platforms

1. To start the main M-GCAT application, use python to call viewer.pyc via `>python viewer.pyc`.

   (a) Note: if you see a 'Bad magic number error' it is likely that you have downloaded an incorrect version of the required python files. Please see the M-GCAT web site to download the appropriate version, either 2.3.x, 2.4.x, or 2.5.x. If you are interested in running the GUI on a earlier version of Python (version 2.1.x or 2.2.x), please contact the developers.

2. To run M-GCAT from the command line, run `>mgcat <inifile>`, where *inifile* is a required argument of the provided mgcat.ini or any other properly formatted parameter file. The parameter file can be automatically generated from inside of the GUI in the File menu.

After starting the main application, you will be able to either generate a new comparison between multiple genomes, or be able to view saved output from a previous M-GCAT run.

3.3 Generating a new multiple genome comparison

To generate a new comparison:

1. Go to the parameters page, then in the *Sequence* parameter section select *Add Sequences* (Figure 1). Select a minimum of 2 sequences to be compared. Note, M-GCAT accepts only FASTA formatted DNA sequences as input. This does not include multi-FASTA formatted files.

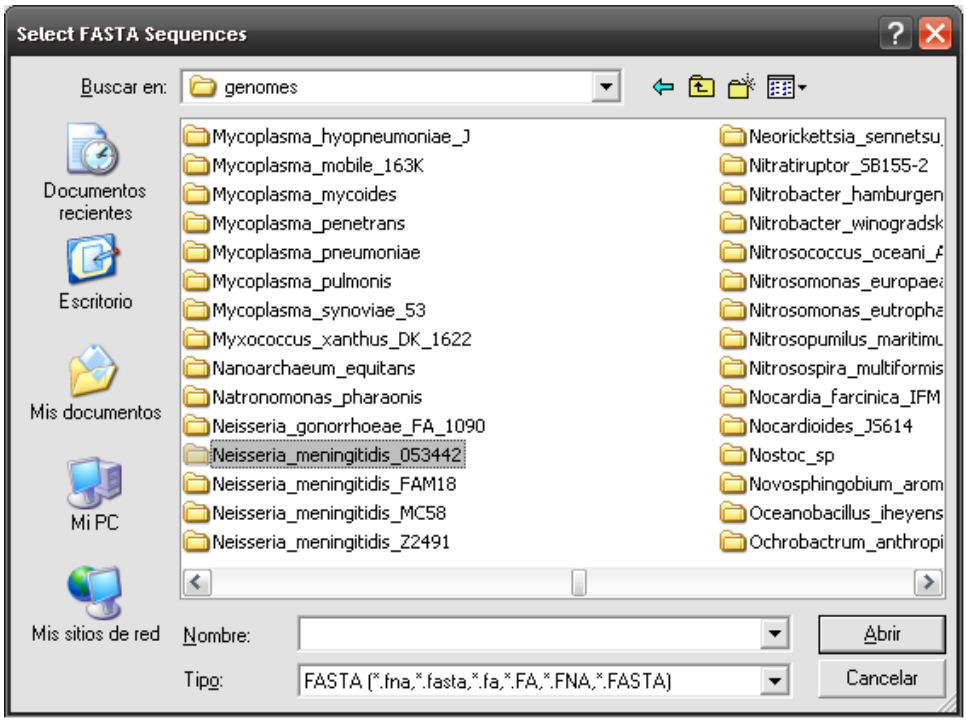


Figure 1: Selecting the input sequences

2. Next, in the *MUM Parameters* section, configure the parameters (see Figure 2):

   (a) **Min MUM Length***: This will determine the the minimum allowable size for a Maximal Unique Match (MUM) within the genome region during the recursive MUM search. That is to say, as searchable genome regions between the initial MUM anchors become smaller and smaller, so should this value.

   (b) **Min Anchor Length***: The minimum allowable size for the initial set of MUM anchors found among all genomes.

   (c) **Random MUM Length**: Random MUM length is the maximum length of MUMs that can be considered statistically insignificant with respect to the genomes being compared. All MUMs less than this length and which meet the random criteria will be removed.

   (d) Note: The entry fields for **Min MUM Length** and **Min Anchor Length** allow for equations and expressions to be entered, as well as integer values. The default value

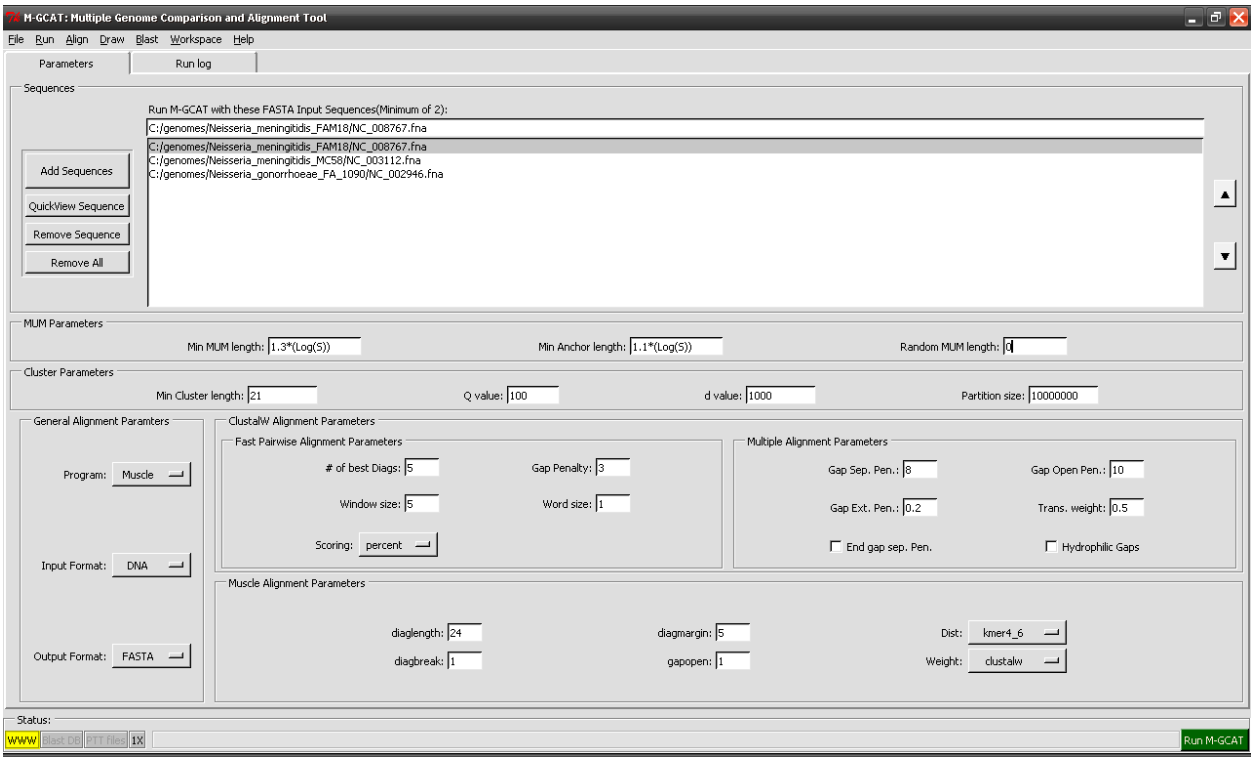


Figure 2: Configuring the input parameters

for both use a 'Log(S)' expression which translates to a log base 4 on the length of the smallest region/sequence.

3. Then, in the *Cluster Parameters* section, configure the parameters:

   (a) **Min Cluster Length**: The value is the minimum allowable length of a Cluster w.r.t the length of the sum of the lengths of the multi-MUMs. The idea is allow larger, collinear Clusters to have more influence on the global comparison framework. By increasing the **Min Cluster Length** value the number of non-collinear Clusters will be reduced and previously separate Clusters will be joined together into a single Cluster.

   (b) **Q**: This value is the minimum allowable length of a region where will perform a search for new MUMs. Decreasing this value will generally generate more mums.

   (c) **D**: This value is the maximum allowable distance between any two MUMs in a cluster.

Increasing this value will generally increase the number of MUM Clusters and decrease the area of the genomes able to be aligned with MUSCLE[4].

(d) **Partition size**: This parameter is used to partition large genomes in order to reduce memory usage. For example, a Partition value of 1000 would split a genome with 10,000 bps into 10 parts, and thus reduce memory usage by a factor of 9 or 10. Decreasing this value generally decreases memory usage and increases running time.

4. Finally, configure the alignment parameters to define the alignment characteristics.

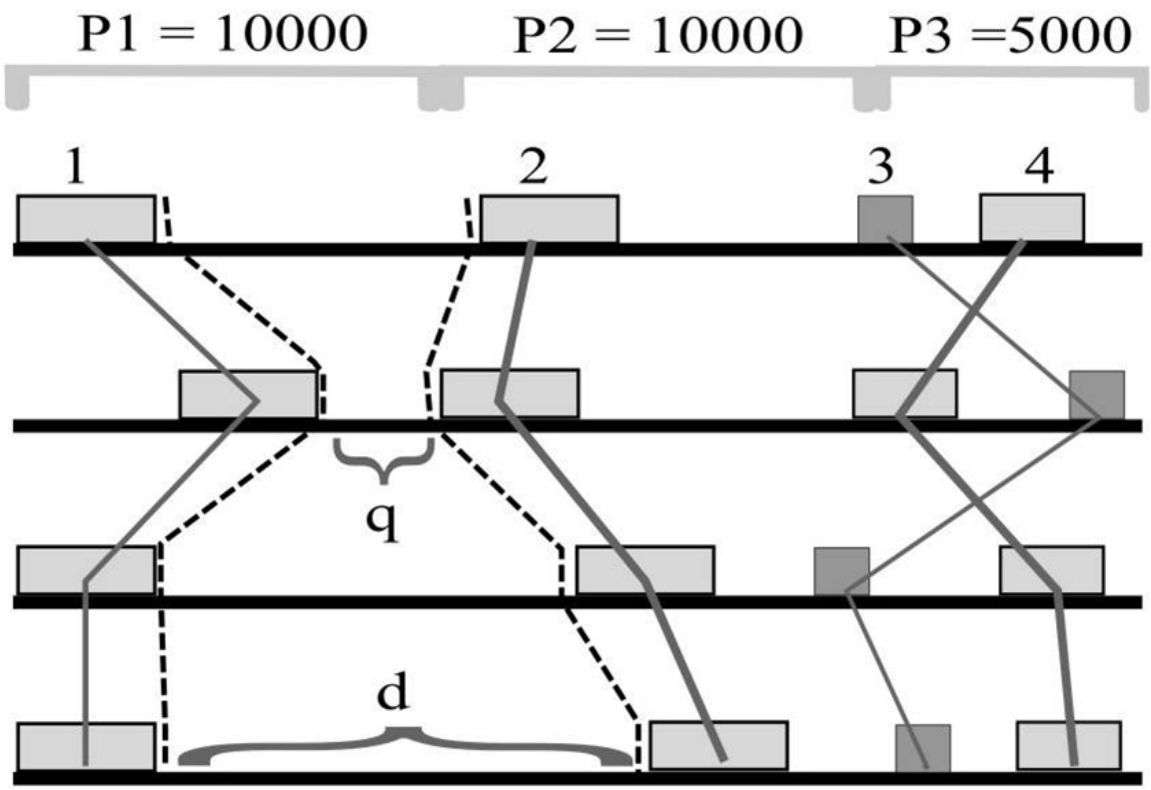


Figure 3: The Partition size parameter in this example is set to 10,000 nt, which consequently partitions the reference genome into three smaller parts, P1, P2, and P3. The parameter Q is minimum allowable distance between any two collinear MUMs to perform a recursive MUM search. Finally, D is the maximum distance that can separate any adjacent MUMs to join them into a MUM cluster.

Now that M-GCAT's parameters are configured, it can be executed from the Run menu or from the Run M-GCAT button located in the lower right corner. Both will call M-GCAT with the configured parameters and the status bar at the bottom of the screen will display *Running M-GCAT*. A command prompt will appear updating the user with the execution times required for each step in the comparison process (see Figures 3.3). When finished, *Finished Running M-GCAT* will appear in the status bar. It is possible to view the results of the analysis by selecting the new "Run #N" tab. Run numbers are increased sequentially, starting at 1. The most recent run is placed in a tab furthest to the right. By default, no more than 10 runs can be used at a time, but runs can be saved and removed. Selecting the "Run #N" will display alignment framework among all genomes consistent with the previously configured parameters.

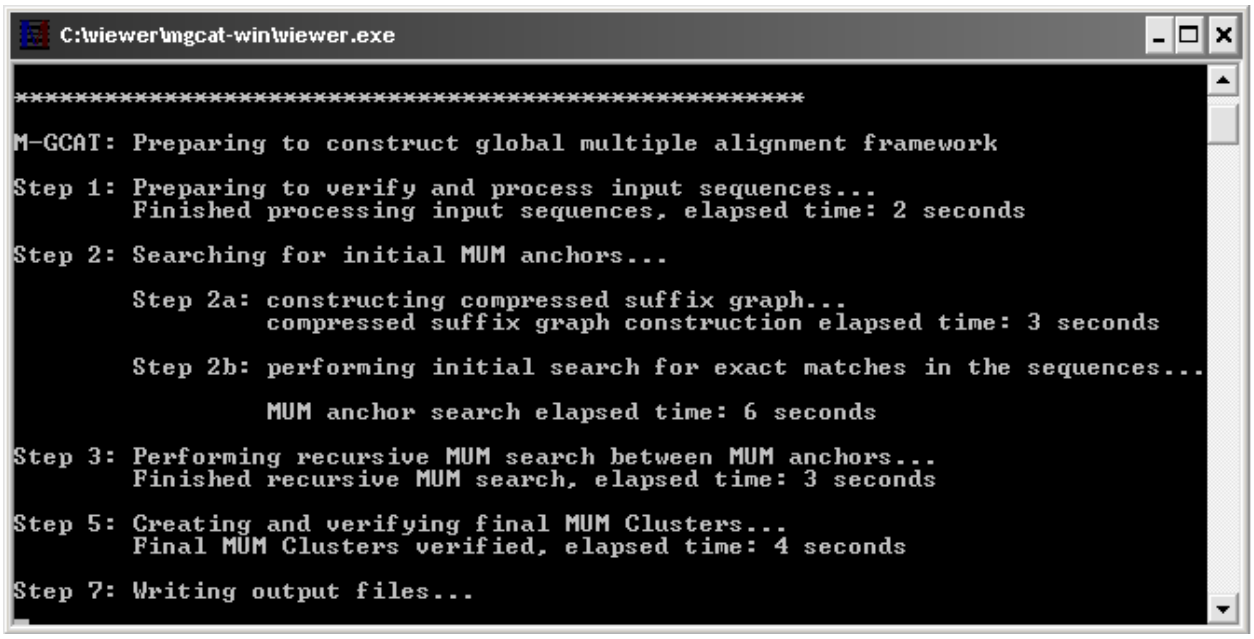


Figure 4: Command-line output while running.

There are six workspaces, each equipped with configurable features and options, designed to provide a distinct working environment based on each interactive task. The main workspace is the Gene viewer workspace in which any selected region can be aligned, displayed with gene information or sent as a NCBI-BLASTN web query with the results incorporated inside of the user interface.

The gene information is provided by the PTT files of NCBI.

1. **Gene viewer workspace**: this is the default workspace inside the graphical user interface of M-GCAT. The topmost window displays the multi-MUM clusters found between these two sequences, which is the global framework that will be used to build the alignment. All regions can be aligned using MUSCLE[4], and when finished the information is stored for future reference. The quality of the alignment is scored and displayed visually, ranking from low identity (light yellow) to high identity (dark red).The bottommost window is the gene map, and is derived from a PTT file that corresponds to each sequence. The PTT file can be edited by the user to update any existing gene annotations, as well as creating any new annotations. Individual genes can be selected and any relevant information for a selected gene is displayed in the window adjacent to the gene map window.

2. **MUM Workspace**: Contains two windows used for displaying a visual representation of multi-MUMs found among all sequences, along with any relevant information. Each multi-MUM can be selected to view its length, start and end positions in the bottom window.

3. **Cluster Workspace**: Contains two windows used for displaying all of the multi-MUM clusters found among all sequences, along with any relevant information. Each cluster can be selected to view its length, start and end positions in the bottom window. Additionally, the clusters can be lined up and traced with the mouse movement.

4. **MUM & Cluster Workspace**: Joins all of the information in the MUM Workspace and Cluster Workspace into one. In this mode, the zoom and movement can be put in sync so that the relationships between the multi-MUMs, multi-MUM clusters, and the d value can be easily studied.

5. **BLAST Workspace**: Configures the windows to display the results from *Blast-¿Blast for pattern in all genomes....* The BLAST hits returned by blastall are displayed visually in black along the sequences, and can be selected and individually inspected.

6. **Alignment viewer Workspace**: The Alignment viewer Workspace joins the Cluster Workspace with an additional window containing the alignment results from the resulting MUSCLE[4] alignment if the selected cluster has been aligned. If it has not been previously aligned, a new alignment can be performed by selecting *Align => Align selected region* from the Main Menu Bar. Once selected, a progress bar at the bottom of window will be activated until finished.

Additionally, in *File => Save RUN as...* (see Figure 5) you can save the output from M-GCAT for viewing at a later time. In the *Save As* window, enter the directory and file name you would like to use to identify the output for the current genome comparison, and all of the output files will be saved accordingly.

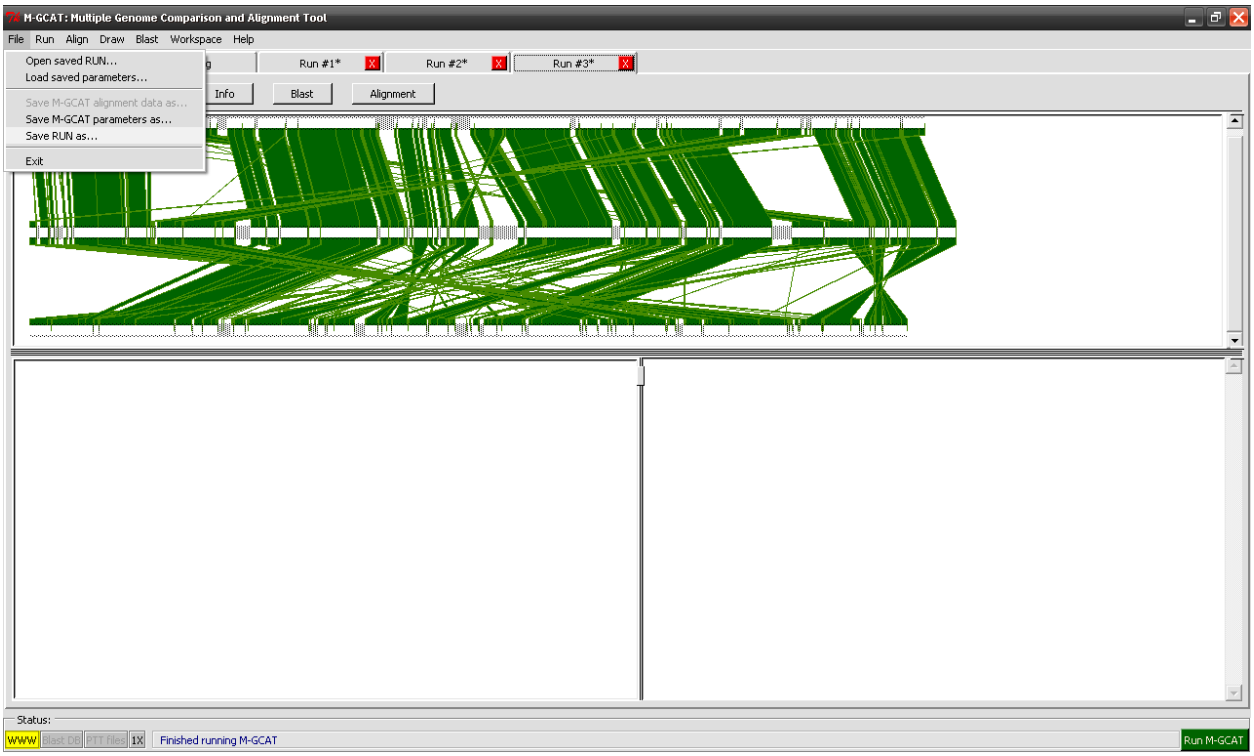


Figure 5: Saving a RUN

3.4 M-GCAT Output Files

The viewer will generate four output files for each successful run of M-GCAT for the given genome set. The output files are:

- *filename*.MUMS: contains all multi-MUMs found
- *filename*.MGCAT: contains all MUMs, Clusters and Regions found.
- *filename*.LOG: contains the information displayed in 'M-GCAT Summary'.
- *filename*.MLN: contains any aligned Clusters/Regions saved in the Mauve[2] alignment format.
- *filename*.XMFA[3]: contains the entire alignment saved in XMFA format.
- *filename*.ALIGN: contains the entire alignment saved in M-GCAT's Cluster based Alignment.
- *filename*.UNALIGN: contains ALL regions that will NOT be included in the final multiple alignment.

To view previously saved output, use the following steps:

1. In the Main Menu Bar, select File-¿Open...
2. Browse to the location of the previously saved .MGCAT file and open it.
3. Once opened, view the results of the saved output by selecting.

3.5 Viewing the LOG file

To view more details on each run, select the Run log tab 6. This will list information relevant to the genomes such as size and name, composition, as well as MUM and MUM cluster details, organized by run.

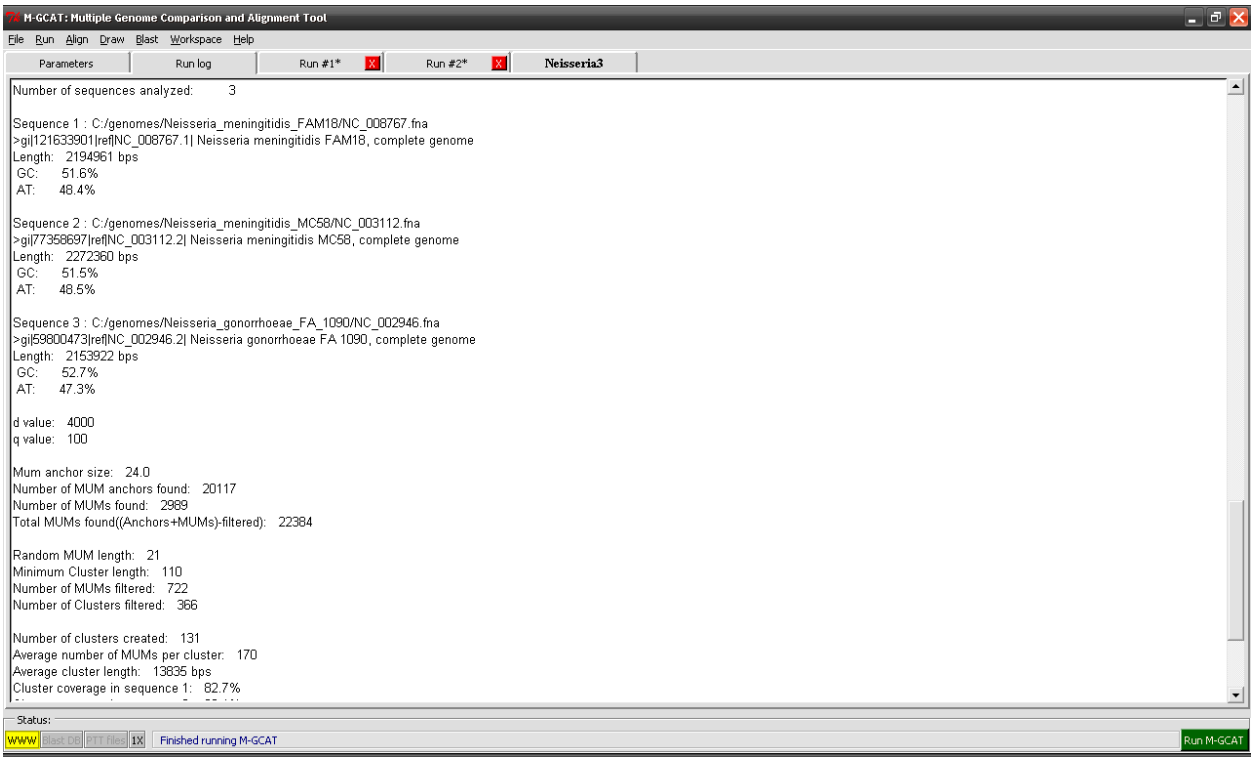


Figure 6: Viewing the LOG file

4 Aligning genomic regions inside of M-GCAT

We will now describe how to align genome regions from within the M-GCAT interactive viewer. All genome alignments are performed with the MUSCLE [1] alignment software, which has been demonstrated to have excellent speed and high accuracy. Some of the MUSCLE alignment parameters can be configured from the M-GCAT Parameter page, while the remaining MUSCLE parameters are configured automatically for the user. Since the design of M-GCAT has focused on aligning closely-related prokaryote genomes that have diverse and mosaic gene repertoires, highly conserved syntenic regions (grouped by MUM Clusters) can be individually aligned and manually inspected to ensure high alignment quality. Region by region alignment can be accomplished by highlighting any given MUM Cluster by selecting it with left mouse-button and then selecting *Align => Align selected region* (see Figure 7a). This will perform a MUSCLE alignment on the selected MUM Cluster, and output on the progress of the MUSCLE alignment will be provided in the command line window (see Figure 7b). Once finished, the alignment will be shown in the window immediate below the MUM Cluster global alignment framework window. If not visible, the window can be activated by selecting the *Alignment* tab inside the current workspace (see Figure 7c).

5 Performing a BLAST on the input sequences

It is also possible to find all regions of local similarity between a query sequence and the sequence database created on the set of input genomes using `blastall` [2]. Select *Blast => Blast for pattern in all genomes...* to activate the BLAST search interface that has been integrated into the M-GCAT interface (see Figure 8a). For each run, the BLAST database will be created the first time a BLAST search is performed. Inside of the BLAST window, specific positions in each genome can be retrieved for the query sequence, or it can be directly constructed or copy & pasted by the user. Common BLAST parameters can be configured at the bottom of the window, such as *e-value cutoff* and *Word size* (see Figure 8b). Once the query sequence and parameters have been configured, click 'OK' to run the BLAST query on the input genome database. When finished, the BLAST results will be displayed in a separate window inside of the current run

workspace. If it doesn't readily appear, click on the *Blast* tab to view the results. Each individual BLAST hit is displayed as a black rectangle along each genome (see Figure 8c). The hits can be activated with the left mouse button, and if they are inside of a MUM Cluster, a green polygon representing the corresponding MUM Cluster will be drawn behind the BLAST hit. Additionally, after clicking on a BLAST hit position and sequence information will be displayed in the *Alignment* window.

6 Viewing M-GCAT alignments inside of Mauve

M-GCAT's closest relative is the robust genome comparison software Mauve [3]. While several differences exist between the two programs, they both accomplish global genome alignment using an recursive MUM anchoring approach. We thus provide multiple alignment output compatible with Mauve, allowing users to view alignments within the Mauve interface that have been generated using M-GCAT (see Figure 10).

7 Contact

For any problems, bugs, doubts, suggestions, comments, etc: treangen at lsi dot upc dot edu

8 How to cite

The original M-GCAT paper was published as:

Todd Treangen and Xavier Messeguer. 2006. M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinformatics 2006, 7:433.

9 Acknowledgments

This work has been supported by the by the project LogicTools (TIN2004-03382) funded by the Spanish Ministry of Science and Technology and the EU program FEDER and AGAUR Train-

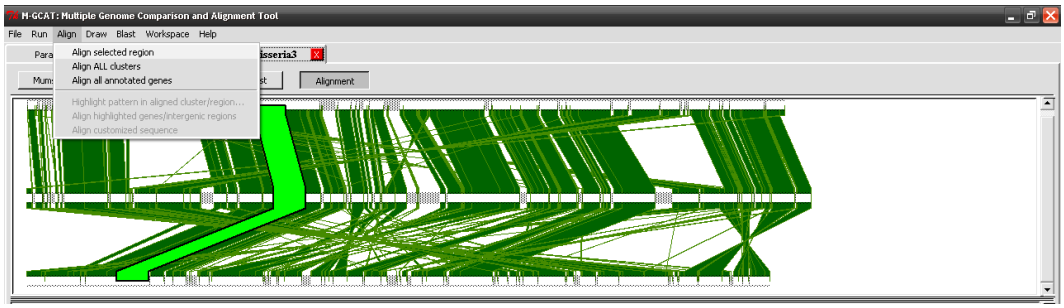
ing Grant FI-IQUC-2005. We would like to thank Aaron Darling for his insightful suggestions throughout the development of M-GCAT.

References

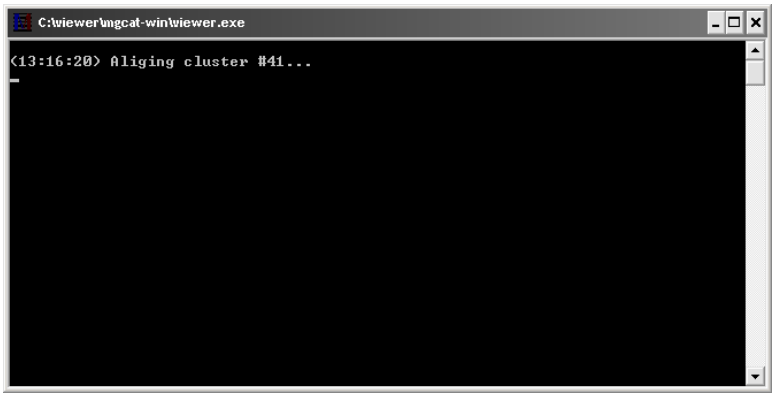
1. Edgar R: **MUSCLE: Multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res.* 2004, **32**(5).

2. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Journal of Molecular Biology* 1990, **215**(3):403–410.

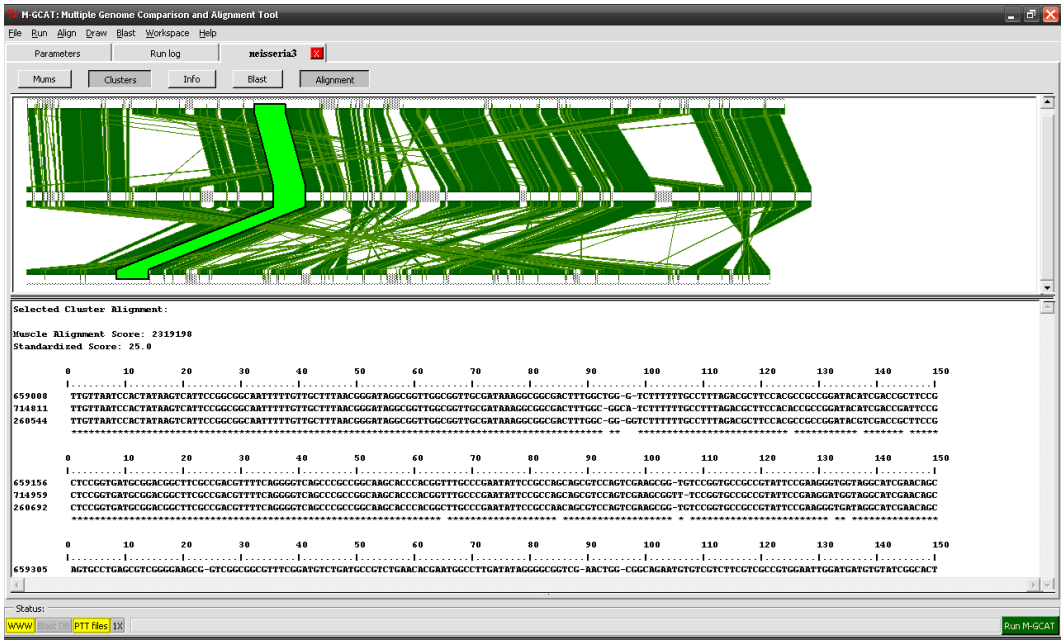
3. Darling A, Mau B, Blattner F, Perna N: **Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements**. *Genome Res.* 2004, **14**:1394–1403.



(a) Choose a MUM Cluster to align with left mouse-button and then select *Align => Align selected region*

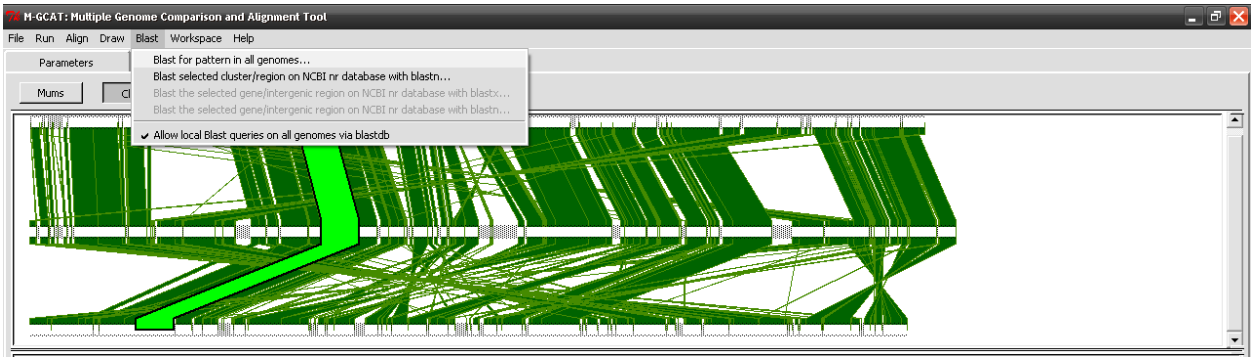


(b) MUSCLE command line output while aligning a region.

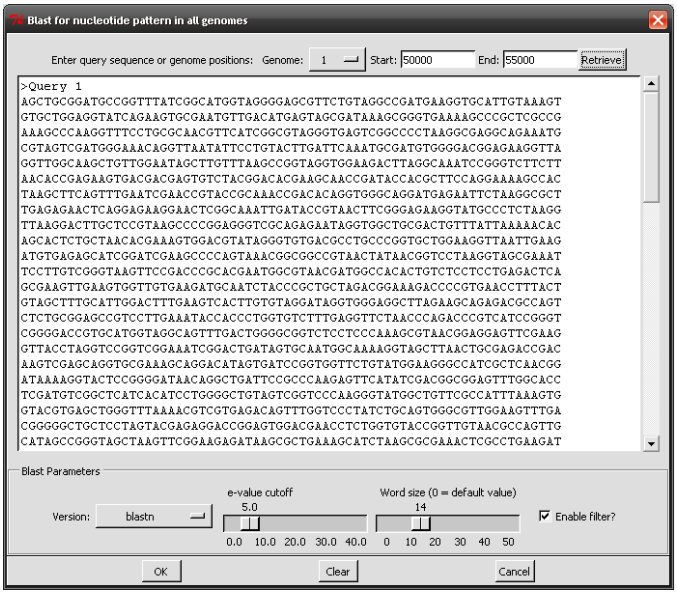


(c) Viewing the resulting multiple alignment.

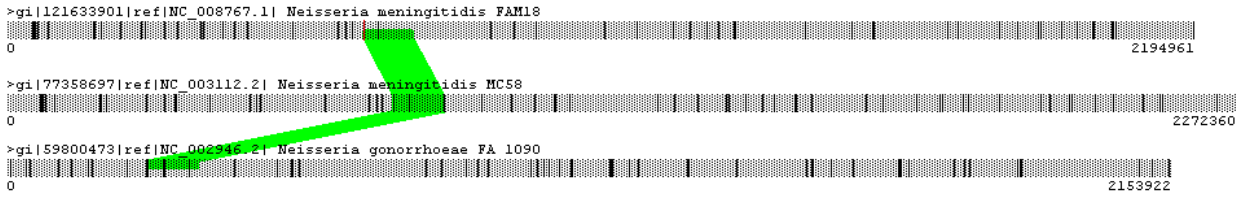
Figure 7: Aligning a selected region in the genomes.



(a) Select *Blast => Blast for pattern in all genomes...* to activate the BLAST search interface.



(b) The BLAST query window



(c) Viewing the BLAST results.

Figure 8: Performing a BLAST on the input sequences.

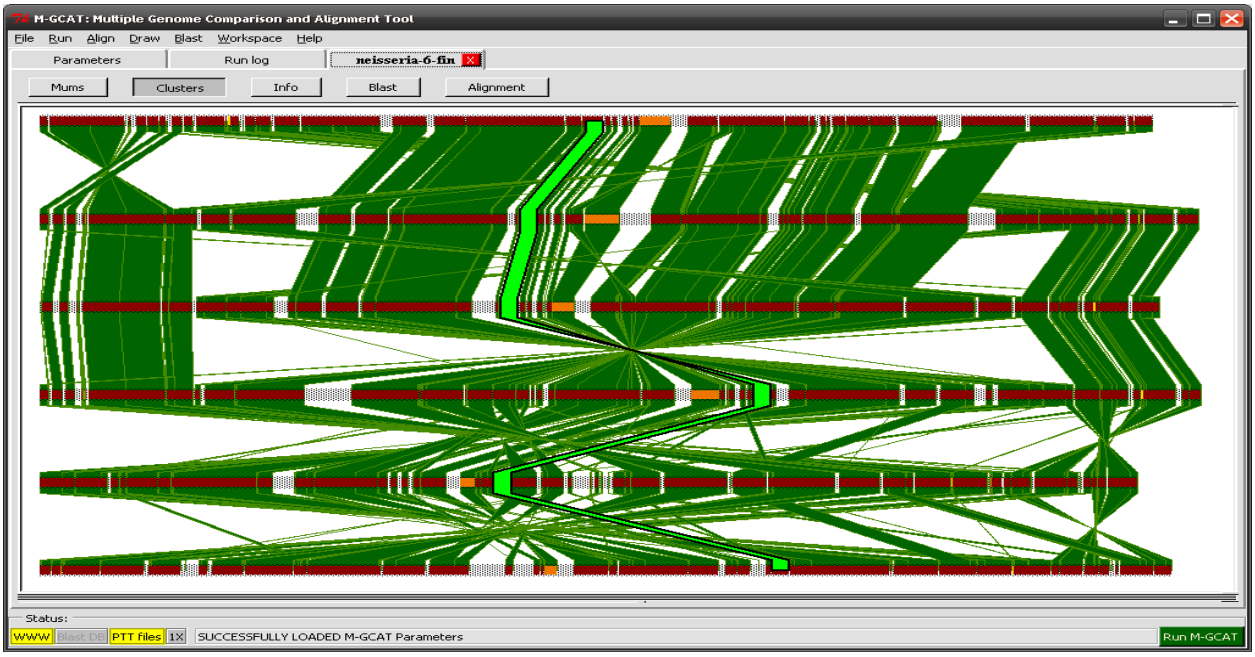


Figure 9: M-GCAT genome alignment of six Neisserial genomes visualized within the M-GCAT interactive viewer.

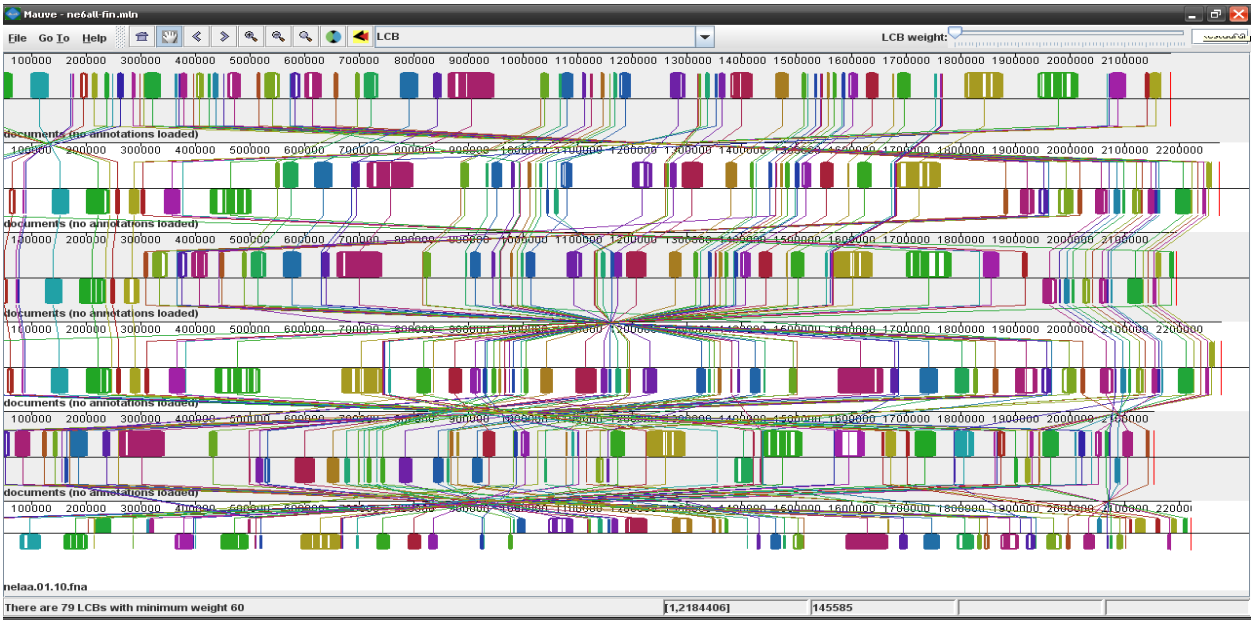


Figure 10: M-GCAT genome alignment of six Neisserial genomes viewed from inside of the Mauve interactive viewer.