MEGA-CC (COMPUTE CORE) AND MEGA-PROTO

User Manual

OVERVIEW…

- MEGA-CC (Molecular Evolutionary Genetics Analysis Computational Core) is an integrated suite of tools for statistics-based comparative analysis of molecular sequence data based on evolutionary principles (Tamura et al. 2011, Kumar et al. 2012). MEGA is used by biologists for reconstructing the evolutionary histories of species and inferring the extent and nature of selective forces shaping the evolution of genes and species.

…OVERVIEW CONTINUED

- MEGA-CC has 2 components
  - MEGA-Proto, an analysis prototyper that is used for generating analysis options files which tell megacc which analysis to run and which options to use. On Windows it is launched from the start menu. On Linux it is launched from a terminal using the *megaproto* command. On Mac it is launched from the *Applications* folder.
  - A command-line executable that performs all calculations. This executable is launched from a terminal using the *megacc* command.

DOCS AND EXAMPLE DATA FILES

- The installers for MEGA-CC copy doc files and example data files to OS-specific locations
  - For Windows users - *%HOMEPATH%\Documents\megacc*
  - For Linux users - */usr/share/megacc*
  - For Mac users - *~/Documents/megacc*
- For Linux and Mac users a man page is included with MEGA-CC. From a terminal:
  - man megacc
- The doc files and example data files are also available from the mega website:
  - http://www.megasoftware.net/webhelp7/helpfile.htm
  - http://www.megasoftware.net/examples.php

MEGA-CC INPUT FILES

- MEGA Analysis Options file
  - Specifies the calculation and desired settings.
  - Created using MEGA-Proto.
  - Has a *.mao* file extension.
- Data file (one of the following)
  - Multiple sequence alignment in MEGA or Fasta format.
  - Distance matrix in MEGA format.
  - Unaligned sequences in Fasta format (for alignment only).
- Newick tree file (required for some analyses)
- Calibrations file (for timetree analysis – but it's optional)
- Groups file (optional)

MEGA-CC OUTPUT FILES

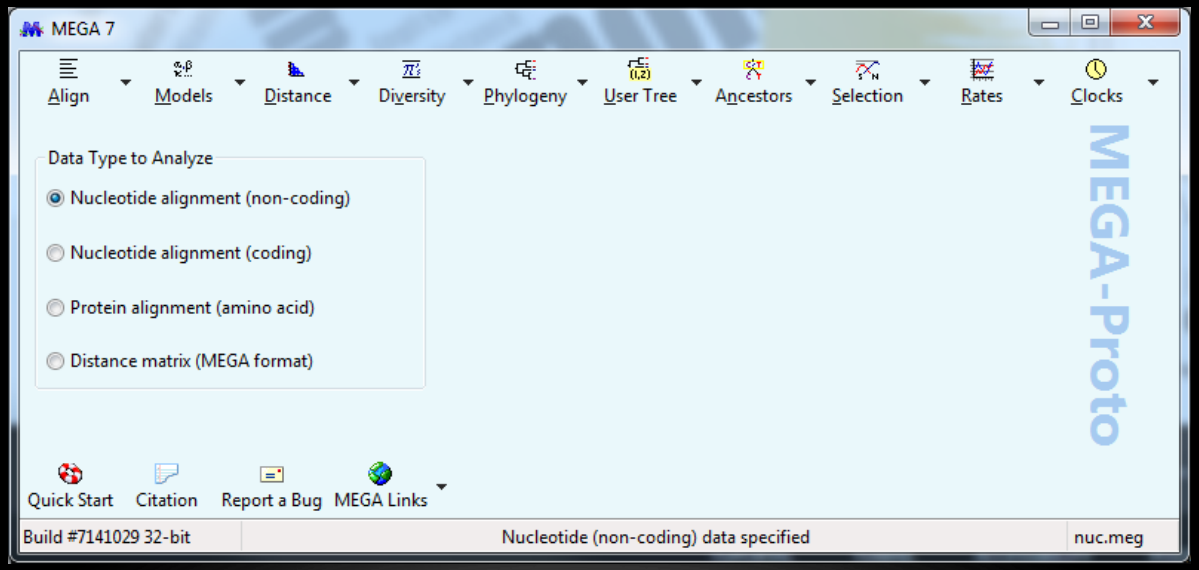
- In general, two output files are produced
  1. Calculation-specific results file (Newick file, distance matrix,…).
  2. A summary file with additional info (likelihood, SBL,…).
  - Some analyses produce additional output (bootstrap consensus tree, csv files, etc…).
- Output directory/filename
  - Default is the same location as the input data file.
  - Specify an output directory and/or file name using *-o* option.
  - If no output filename is specified, MEGA-CC will assign a unique name.
- Errors/warnings
  - If MEGA-CC produces any errors or warnings, they will be logged in the the summary file.

RUNNING MEGA-CC

- Easiest to run using command-line or batch scripts:
  - megacc –a settings.mao –d alignment.meg –o outFile
- Can also be run using custom scripts (Perl, Python, …):
  - exec('megacc –a options.mao –d alignment.meg –o outFile');
- Integrated *File Iterator* system can process multiple files without the need for using scripts (see Demo2 below)
- In addition, other applications can launch MEGA-CC:
  - status = CreateProcess("path/to/megacc…");
- To see a list of available command options, call megacc from a command-line prompt with the –h flag (Unix users can view the man page).

MEGA-PROTO (ANALYSIS PROTOTYPER)

- Has the same look and feel as the GUI edition of MEGA.
- Produces MEGA Analysis Options files.
- Has no computational capabilities.

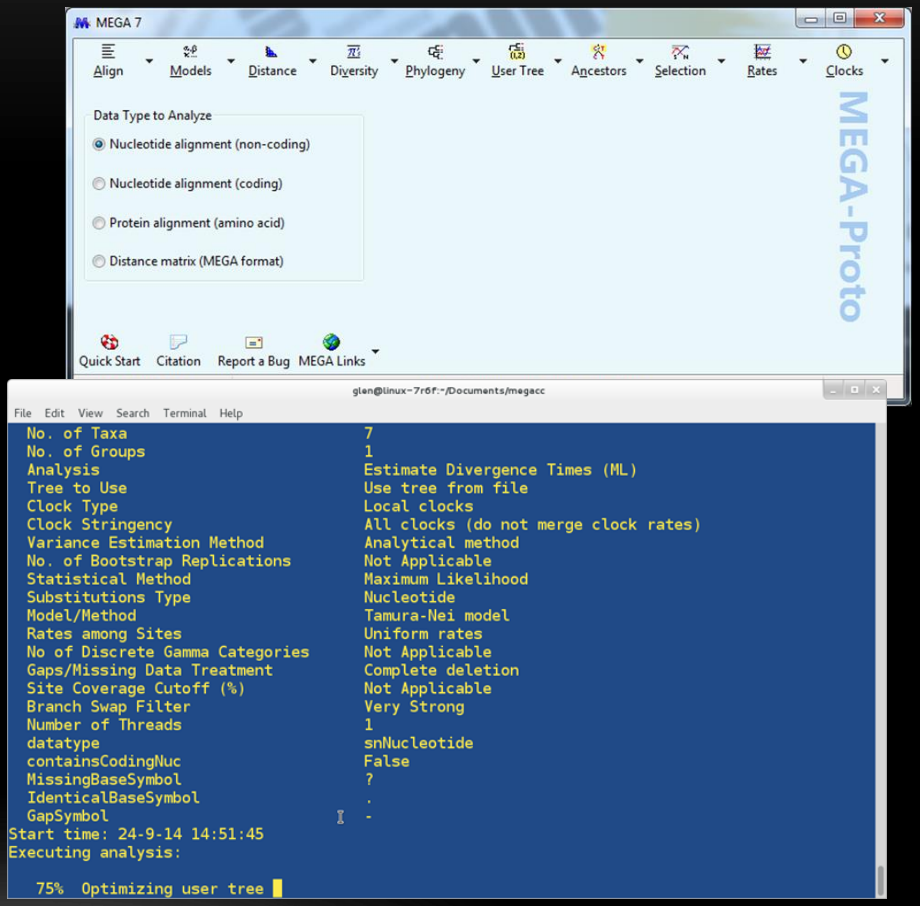


USING MEGA-PROTO

1. Select input data type.
   - Nucleotide (non-coding)
   - Nucleotide (coding)
   - Protein (amino-acid)
   - Distance matrix (MEGA format)
2. Select analysis from menu.
3. Adjust analysis settings.
4. Click *Save Settings*… and save the MEGA Analysis Options (*.mao) file to your computer.

DEMO 1

- The following example demonstrates how to create a timetree using MEGA-Proto and MEGA-CC



DEMO 1 DATA FILES

- For this demo, we will be using some of the example data files that were copied to your computer by the installer
- For Windows users, the files are located in your %HOMEPATH%\Documents\megacc directory
- For Linux users, the files are located in your */usr/share/megacc* directory
- For Mac users, the files are located in your *~/Documents/megacc* directory

DEMO 1 SETUP (LINUX ONLY)

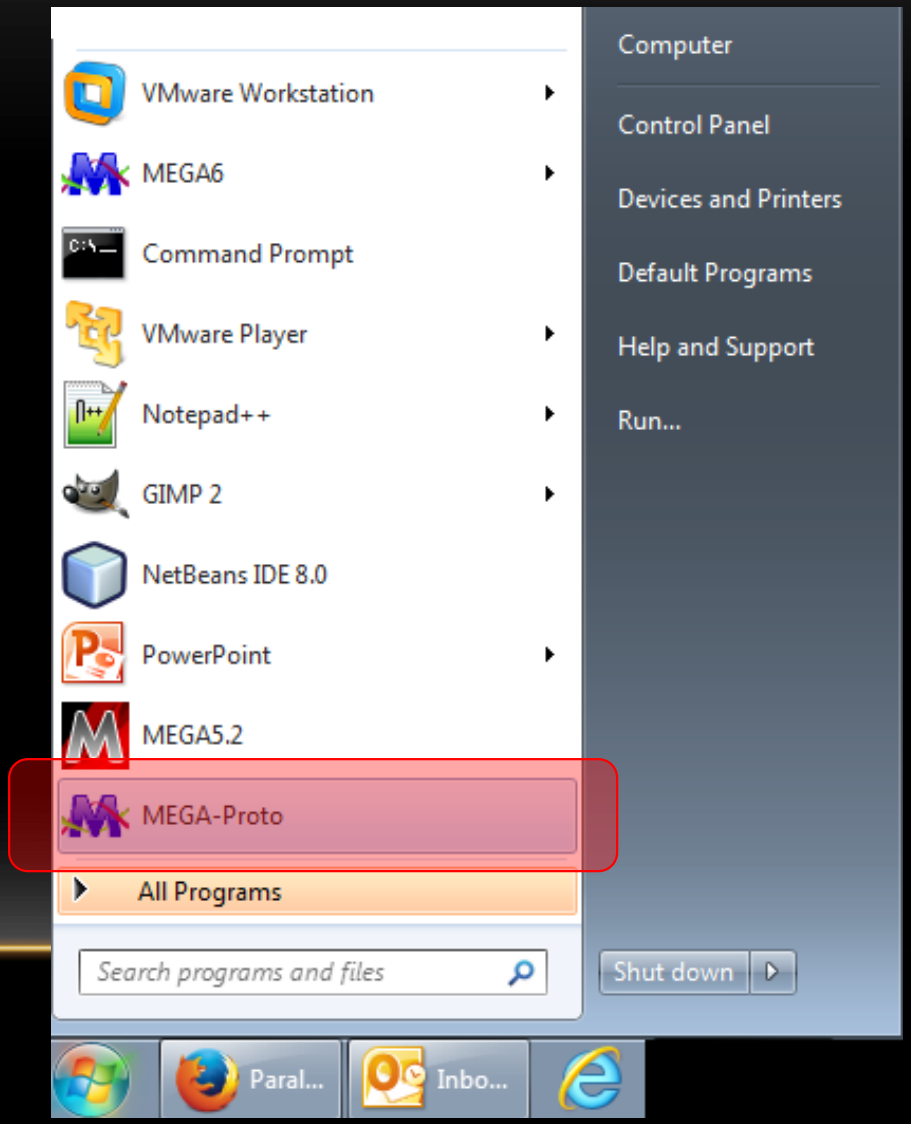
- If you are using Linux (Windows and Mac users can skip this), we want to move the example data files to a more accessible location and change ownership (currently owned by root). Execute the following commands to move the files into your Documents directory

```
glen@linux-pslv:~/Documents
File  Edit  View  Search  Terminal  Help
glen@linux-pslv:~> cd ~/Documents/
glen@linux-pslv:~/Documents> sudo cp -r /usr/share/megacc/ megacc
glen@linux-pslv:~/Documents> sudo chown -R glen:users megacc
glen@linux-pslv:~/Documents> ls -alh
total 16K
drwxr-xr-x  4 glen users 4.0K Sep 25 08:58 .
drwxr-xr-x 36 glen users 4.0K Sep 25 08:53 ..
drwxr-xr-x  6 glen users 4.0K Sep 25 08:42 lazarusProjects
drwxr-xr-x  4 glen users 4.0K Sep 25 08:58 megacc
glen@linux-pslv:~/Documents>
```

Be sure to use your own user and group names

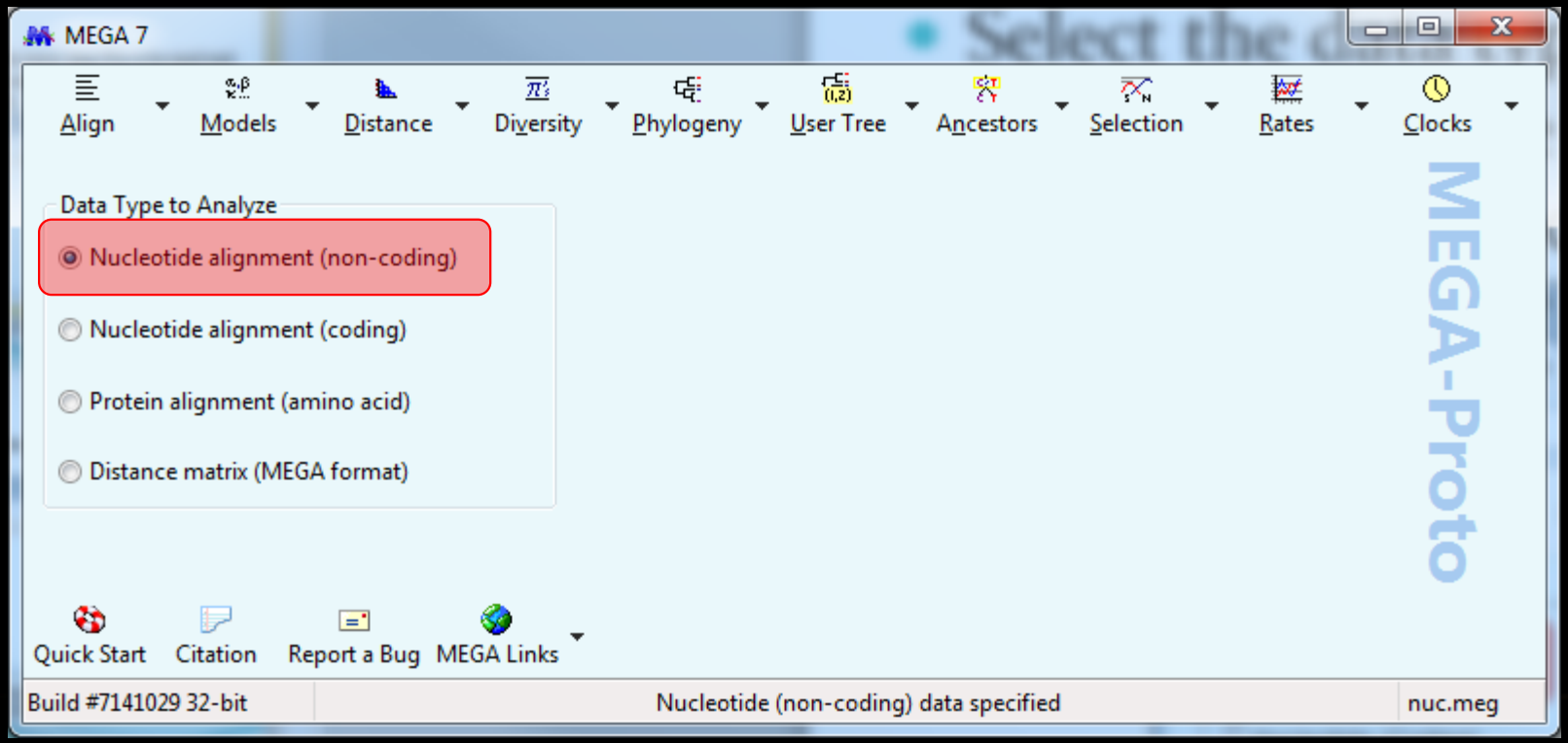
STEP 1

- Windows - Open MEGA-Proto by selecting MEGA-Proto from the Start Menu.
- Linux – Open MEGA-Proto by entering the megaproto command in a terminal window.
- Mac – Open MEGA-Proto by double-clikcing it in your Applications folder.



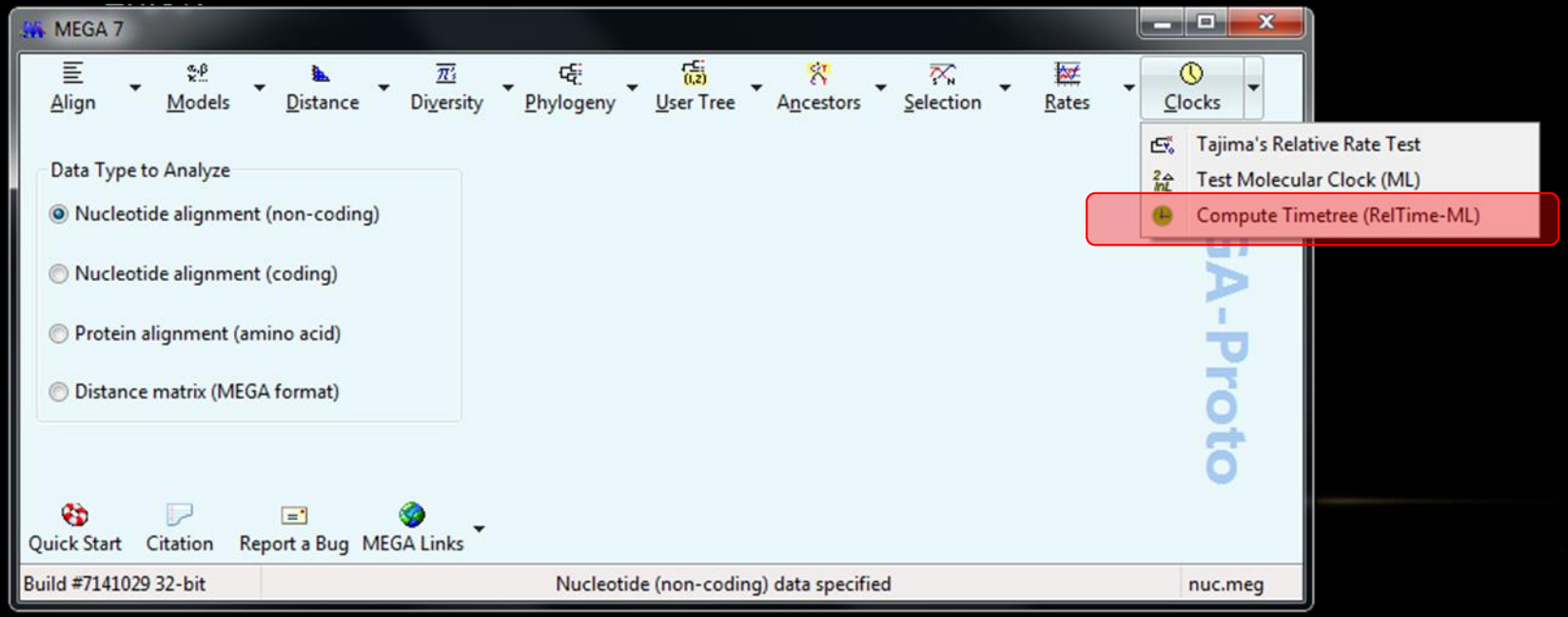
STEP 2

- Select the data type of the input data file to be analyzed. For this demo, we will accept the default setting - Nucleotide (non-coding).



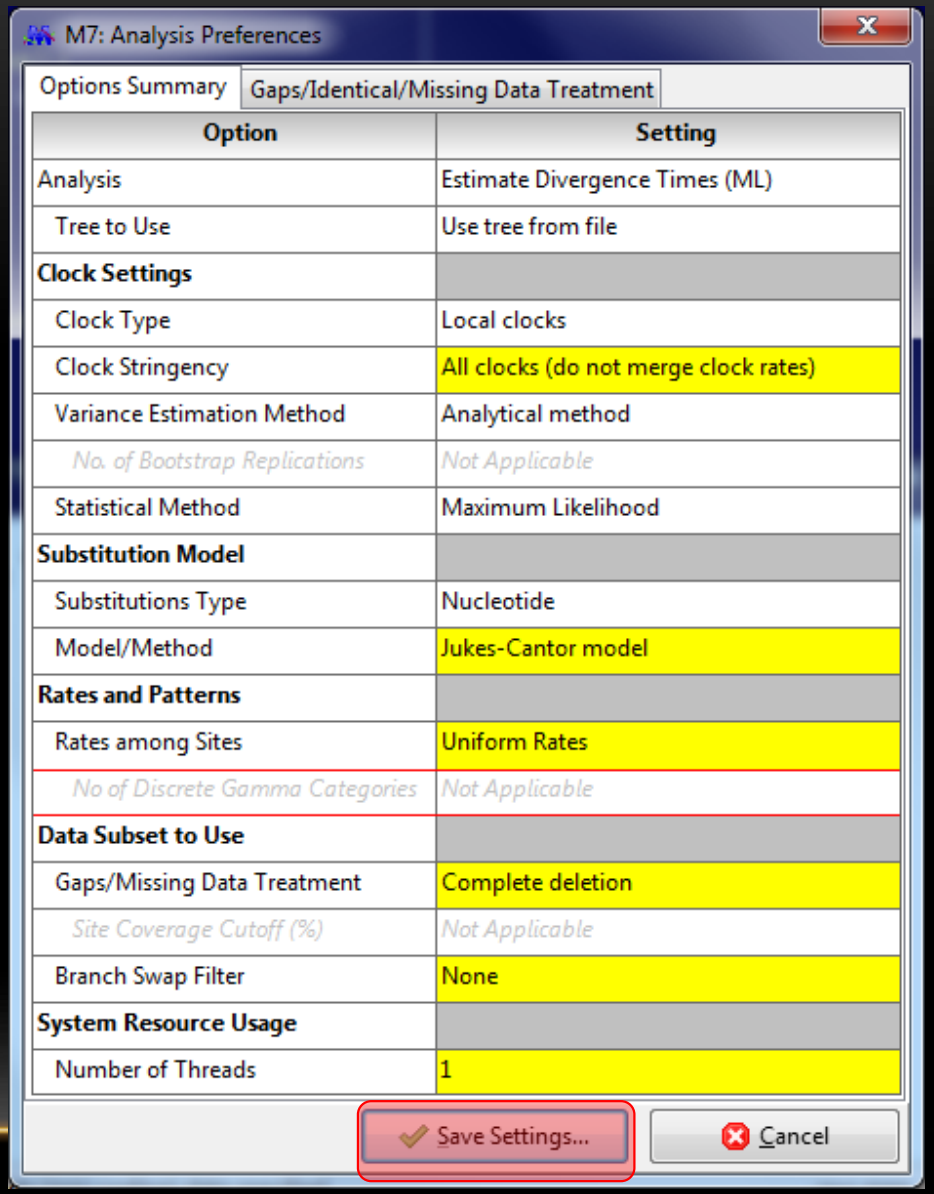
STEP 3

- Select *Compute Timetree (Reltime ML)* from the *Clocks* menu.



STEP 4

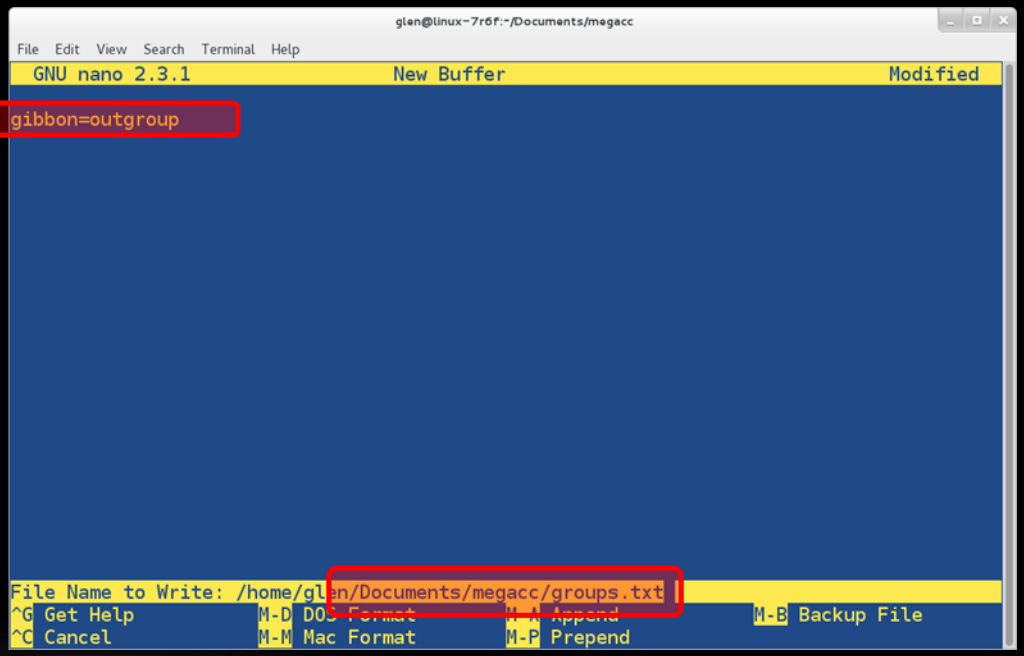
- Adjust the analysis preferences to match those shown.
- Click the *Save Settings…* button and save the analysis options file as *demoSettings.mao* in the megacc\examples directory.



Option	Setting
Analysis	Estimate Divergence Times (ML)
Tree to Use	Use tree from file
Clock Settings	
Clock Type	Local clocks
Clock Stringency	All clocks (do not merge clock rates)
Variance Estimation Method	Analytical method
No. of Bootstrap Replications	*Not Applicable*
Statistical Method	Maximum Likelihood
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	Jukes-Cantor model
Rates and Patterns	
Rates among Sites	Uniform Rates
No of Discrete Gamma Categories	*Not Applicable*
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Branch Swap Filter	None
System Resource Usage	
Number of Threads	1

STEP 5

- The timetree analysis requires that we specify an outgroup. To do so, create a text file and add the line ‘gibbon=outgroup’. Save this file as *groups.txt* in the *megacc* directory.



STEP 6

- Open a command terminal and navigate to ~/Documents/megacc using the cd command
- Execute the analysis by calling megacc as follows:

```
glen@linux-pslv:~/Documents/megacc
File  Edit  View  Search  Terminal  Help
glen@linux-pslv:~/Documents> cd ~/Documents/megacc
glen@linux-pslv:~/Documents/megacc> megacc -a demo1settings.mao -d Examples/mtCDNA.meg -t Exam
ples/mtCDNA.nwk -g groups.txt -c Examples/mtCDNACalibration.txt -o timeTreeDemo
```

The divergence time calibration file is optional for the timetree analysis

STEP 7

- The analysis will be launched and progress updates will be displayed in the command prompt window.

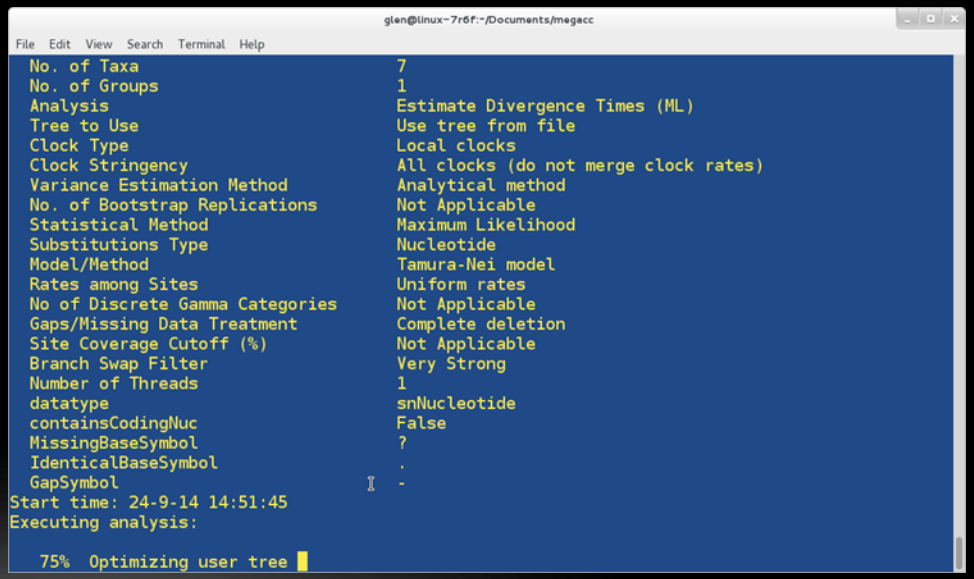
```
glen@linux-7r6f:~/Documents/megacc
File  Edit  View  Search  Terminal  Help
  No. of Taxa                          7
  No. of Groups                        1
  Analysis                             Estimate Divergence Times (ML)
  Tree to Use                          Use tree from file
  Clock Type                           Local clocks
  Clock Stringency                     All clocks (do not merge clock rates)
  Variance Estimation Method           Analytical method
  No. of Bootstrap Replications        Not Applicable
  Statistical Method                   Maximum Likelihood
  Substitutions Type                   Nucleotide
  Model/Method                         Tamura-Nei model
  Rates among Sites                    Uniform rates
  No of Discrete Gamma Categories      Not Applicable
  Gaps/Missing Data Treatment          Complete deletion
  Site Coverage Cutoff (%)             Not Applicable
  Branch Swap Filter                   Very Strong
  Number of Threads                    1
  datatype                             snNucleotide
  containsCodingNuc                    False
  MissingBaseSymbol                    ?
  IdenticalBaseSymbol                  .
  GapSymbol                            -
Start time: 24-9-14 14:51:45
Executing analysis:

   75%  Optimizing user tree
```

FINALLY

- The analysis will produce several output files in the directory megacc\examples\M7CC_Out
  - mtCDNA-xxxx_exactTimes.nwk
    - This Newick file gives the timetree scaled according to the estimated divergence times.
  - mtCDNA-xxxx_relTimes.nwk
    - This Newick file gives the timetree scaled according to the estimated relative divergence times.
  - mtCDNA-xxxx.txt
    - This text file gives a more detailed representation of the timetree, including relative times, exact times, evolutionary rates, and divergence time std errors.
  - mtCDNA-xxxx_summary.txt
    - This file gives analysis information such as the log likelihood value of the Maximum Likelihood tree, ts/tv ratio, etc…

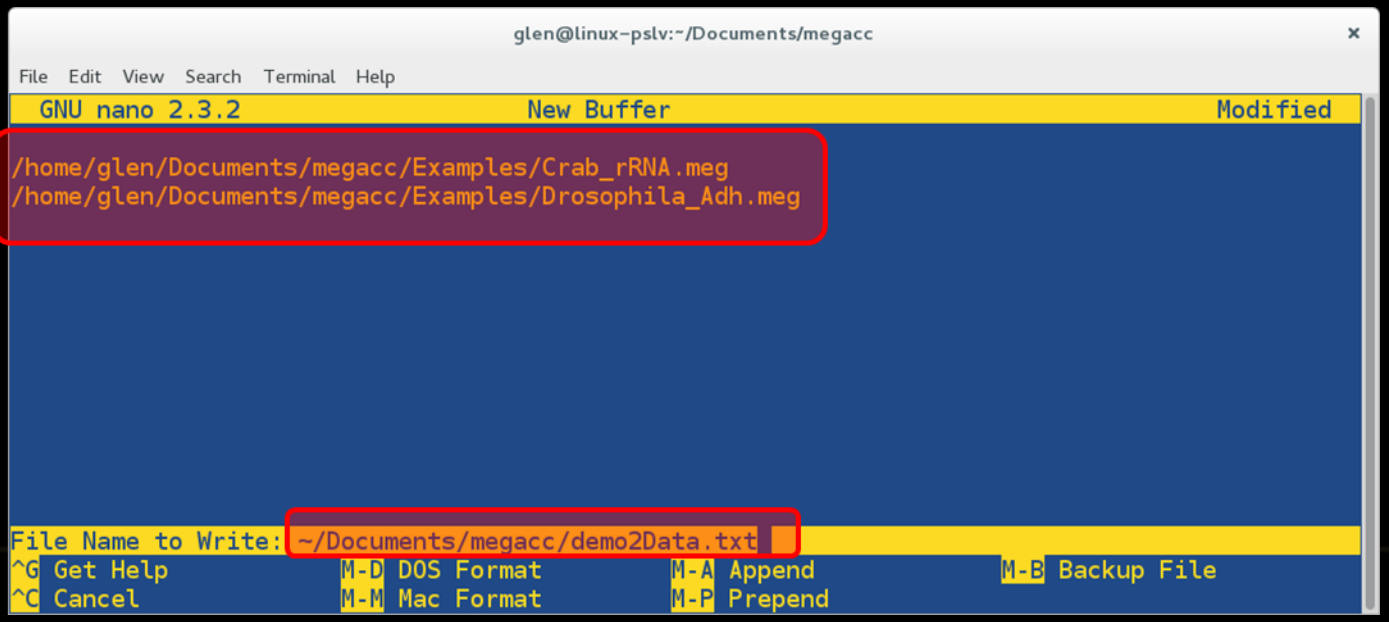
DEMO 2

- The following example demonstrates how to use the File Iterator system in MEGA-CC to process multiple input data files using a single analysis options file.



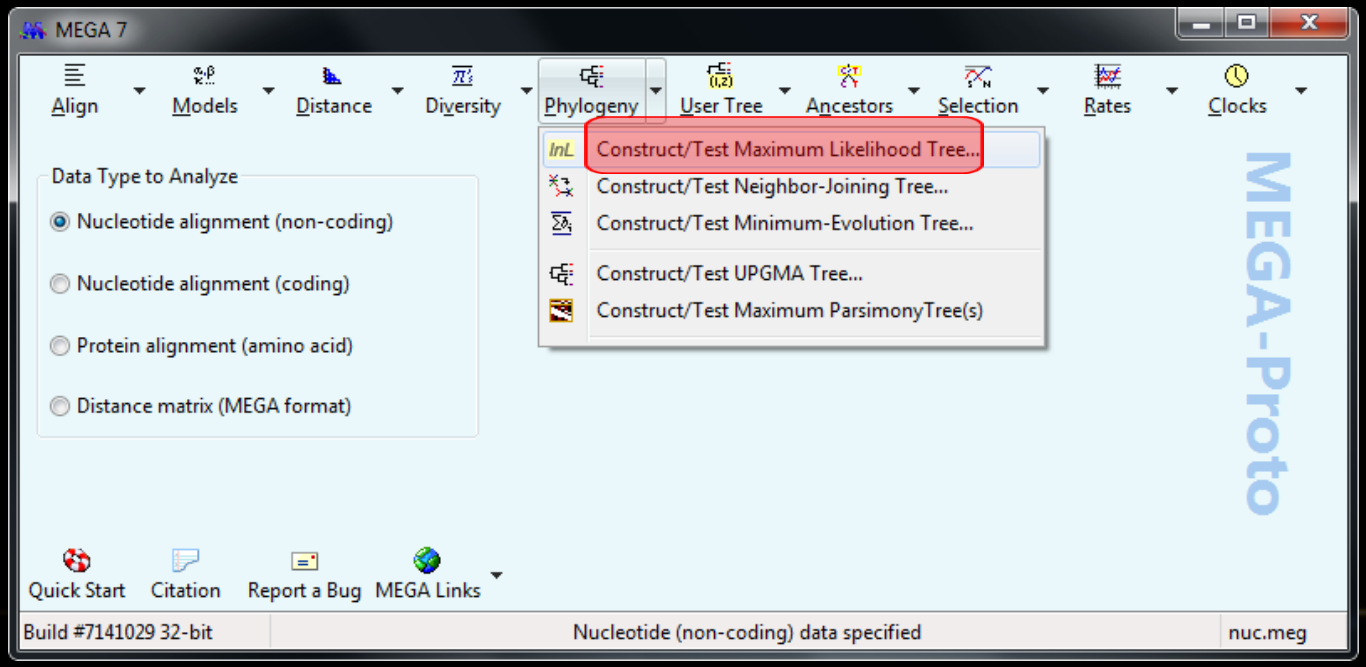
STEP 1

- Create a text file named demo2Data.txt which we will use to specify multiple alignment files for ML phylogeny inference. Save this file in the *megacc* directory.
- In this file, add the full paths to the Crab_rRNA.meg and Drosophila_Adh.meg example files.



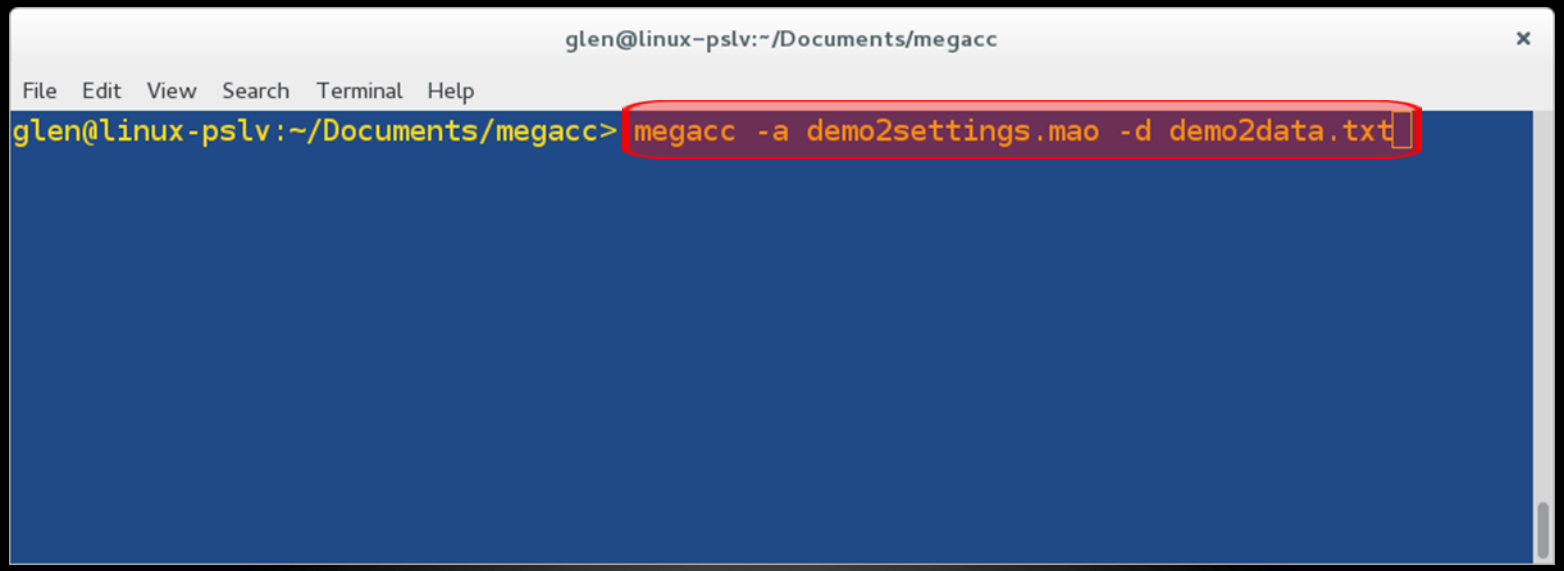
STEP 2

- Using MEGA-Proto, create a .mao file for ML phylogeny construction with the default settings and save the file to ~/Documents/megacc/demo2settings.mao



STEP 3

- From a command-line prompt, call MEGA-CC as below (note that we don't specify an output name):



STEP 4

- The analyses will be launched sequentially and progress updates will be displayed in the command prompt window.

```
glen@linux-pslv:~/Documents/megacc
File  Edit  View  Search  Terminal  Help

    0%  Organizing sequence information                          g
    0%   25-9-14 09:55:36
Using the following analysis options:
  No. of Taxa                          11
  Analysis                             Phylogeny Reconstruction
  Statistical Method                   Maximum Likelihood
  Test of Phylogeny                    None
  No. of Bootstrap Replications        Not Applicable
  Substitutions Type                   Nucleotide
  Model/Method                         Tamura-Nei model
  Rates among Sites                    Uniform rates
  No of Discrete Gamma Categories      Not Applicable
  Gaps/Missing Data Treatment          Complete deletion
  Site Coverage Cutoff (%)             Not Applicable
  ML Heuristic Method                  Nearest-Neighbor-Interchange (NNI)
  Initial Tree for ML                  Make initial tree automatically (Default -
 NJ/BioNJ)
  Branch Swap Filter                   Very Strong
  Number of Threads                    1
  datatype                             snNucleotide
  containsCodingNuc                    False
  MissingBaseSymbol                    ?
  IdenticalBaseSymbol                  .
  GapSymbol                            -
  Select Codon Positions               Select Codon Positions=1st, 2nd, 3rd, Non-
Coding
Start time: 25-9-14 09:55:36
Executing analysis:

       100% Analysis Complete
glen@linux-pslv:~/Documents/megacc>
```

FINALLY

- The analysis will produce output files for each input data file
- In this example, the same analysis options were used for each alignment file
- Enjoy!

MEGA-CC DEVELOPMENT TEAM

Koichiro Tamura

- Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Hachioji, Tokyo, Japan
- Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

Glen Stecher

- Institute for Genomics and Evolutionary Medicine, Temple University
- Center for Evolutionary Medicine and Informatics, Arizona State University

Sudhir Kumar

- Institute for Genomics and Evolutionary Medicine, Temple University
- Department of Biology, Temple University
- Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia