MOLE-BLAST

A tool for clustering multiple sequences with their database neighbors

https://www.ncbi.nlm.nih.gov/moleblast/moleblast.cgi

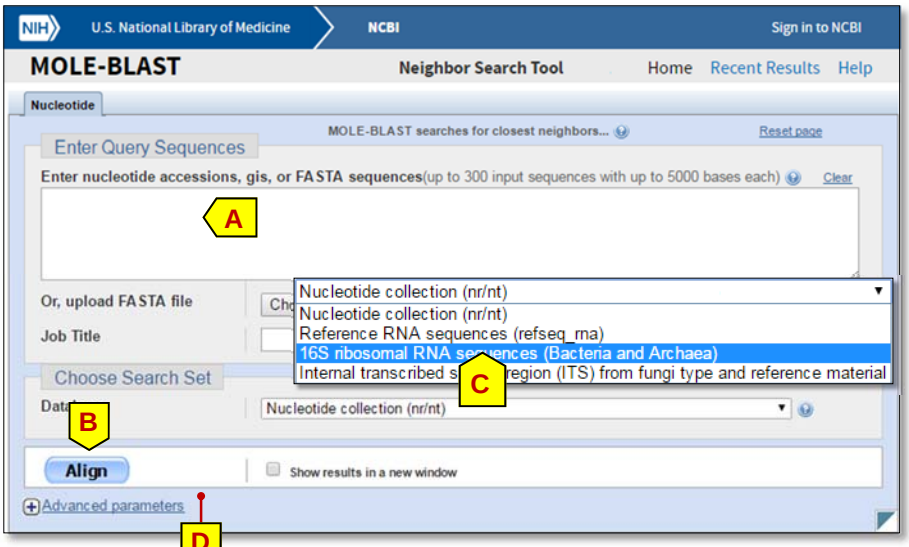
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Overview

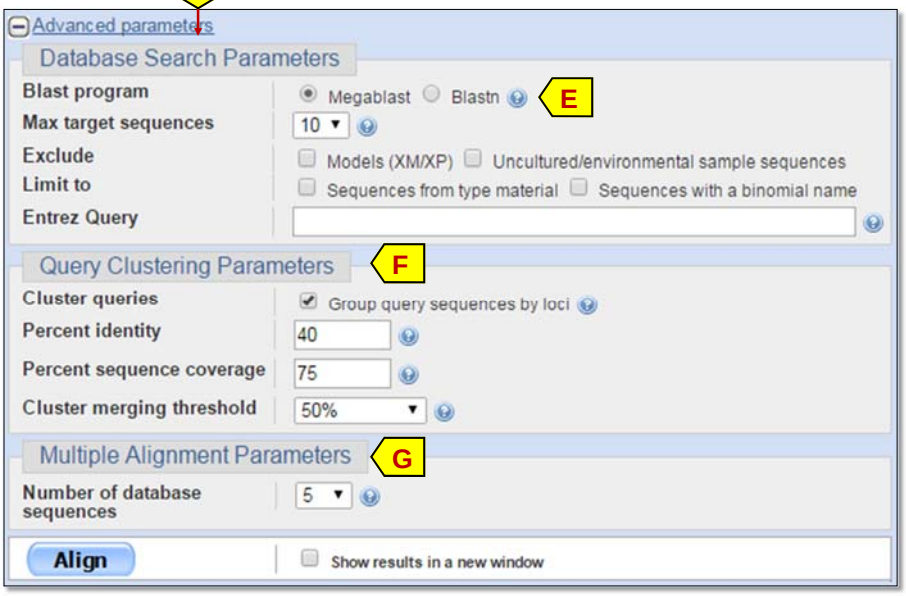
Large scale sequencing of unknown or environmental samples is a powerful method in biological researchto help identify organisms present on the skin, in the gut, in soil sample, or in harsh environments such as toxic sites and natural eco-systems, to provide important insight on human health, bioremediation, global and local ecosystem processes. MOLE-BLAST is a tool designed cluster input sequences and place them in taxonomic context to help identify these sequences and the organisms they were isolated from. It works by taking multiple nucleotide sequences as input, searching them against a source database to find their best matches. The input query sequences along with their newly identified data-base matches are then subjected to multiple sequence alignment and the result is used to construct a phylogenetic tree. By placing input query sequences under the context of known sequences, MOLE-BLAST helps establish the taxonomic relationship to those known sequences and significantly increases the reliability of the identification process.

Access

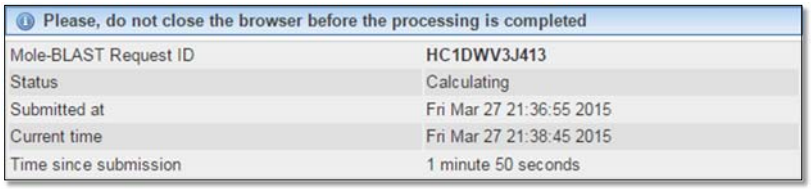
MOLE-BLAST is available from the "Specialized BLAST" section of the BLAST homepage (blast.ncbi.nlm.nih.gov). The search page is at blast.ncbi.nlm.nih.gov/moleblast/moleblast.cgi.



The top input box (**A**) of this page accepts a group of query nucleotide sequences in FASTA format or a list of NCBI database identifiers. Clicking the "Align" button (**B**) starts the search with default settings with the neighboring sequences coming from NT nucleotide database. MOLE-BLAST also provides specialized databases including the 16S reference sequences from bacteria and archaea (**C**). Additional parameters for the search are available in the expandable "Advanced parameters" section (**D**). The "Database Search Parameters" section (**E**) controls the search sensitivity - megablast for faster searches, blastn for more sensitive searches, and number of matches to keep for the BLAST search step. It also allows database subset selection for more focused search. The "Query Clustering Parameters" section (**F**) sets thresholds for separating clusters into loci. These settings may separate sequences from different genes if they are present in the sample. The "Multiple Alignment Parameters" section (**G**) determines how many database matches found in the BLAST search step should be used in multiple alignment to construct the phylogenetic tree.

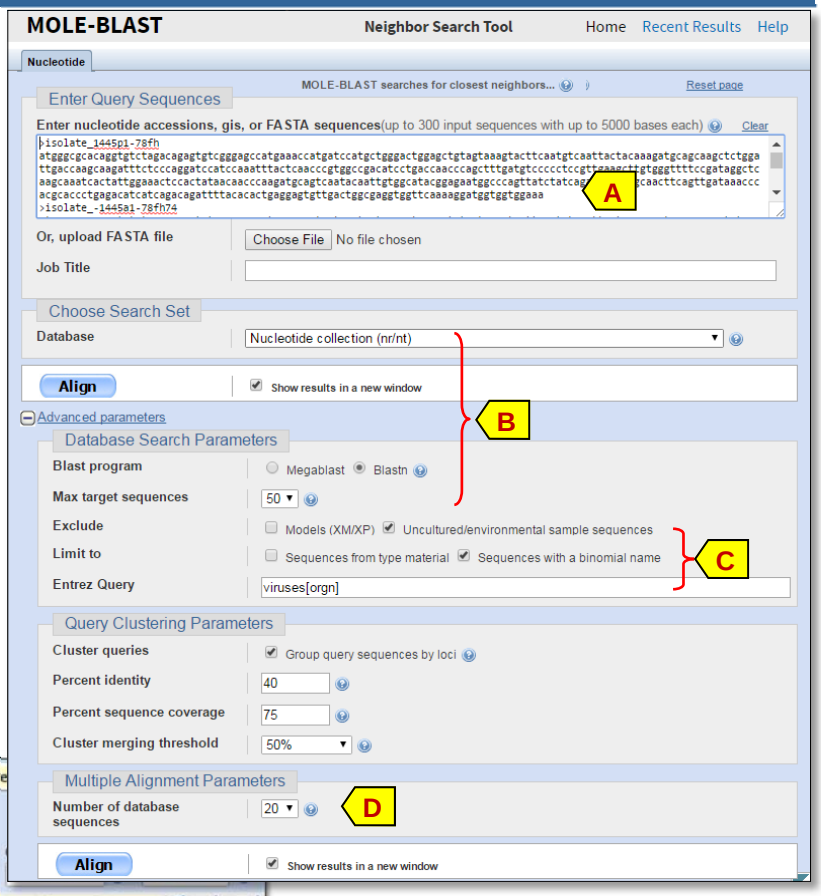


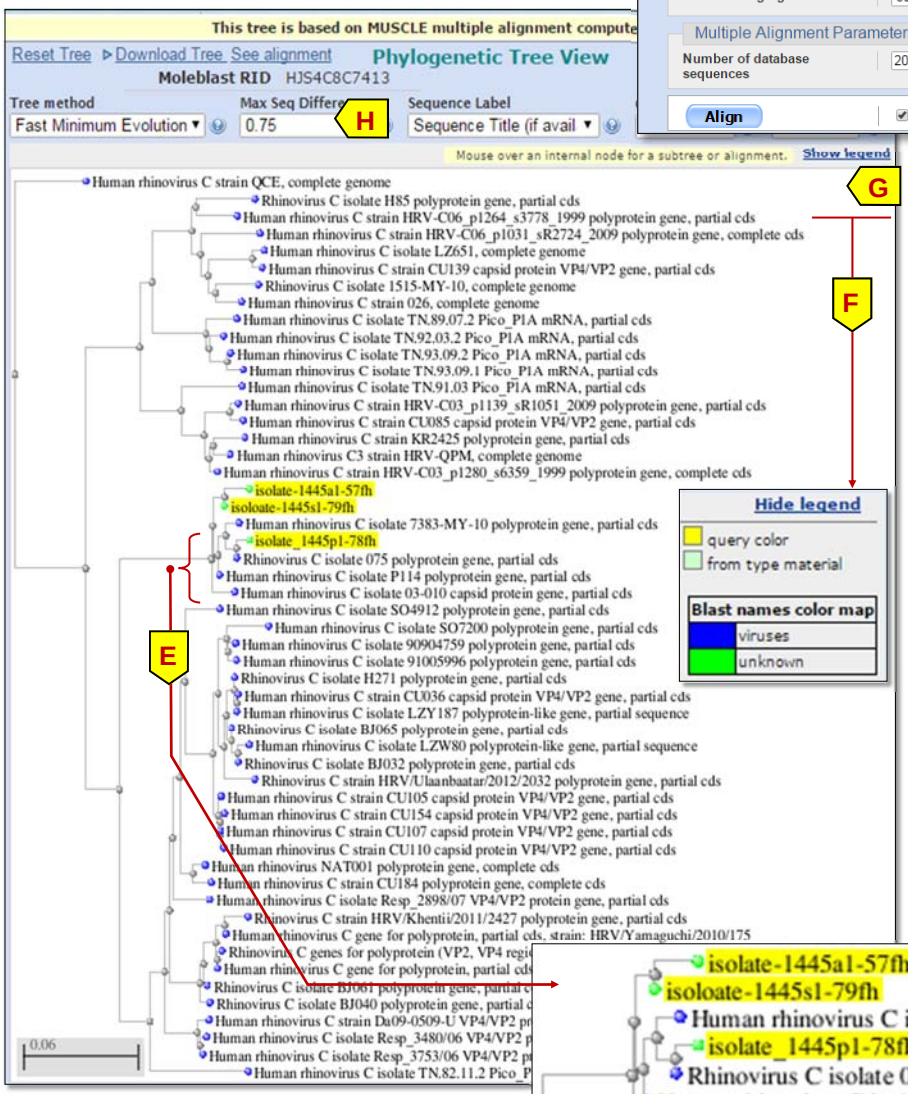
Pushing the "Align" button initiates the search and brings up the interim screen (**H**). **Leave this browser window open until the search is completed to avoid loosing the assigned Request ID (RID, I) and the results.** Search results are save for 24 hours and can be retrieved using the assigned RID and the input box under the "Recent Results" tab.



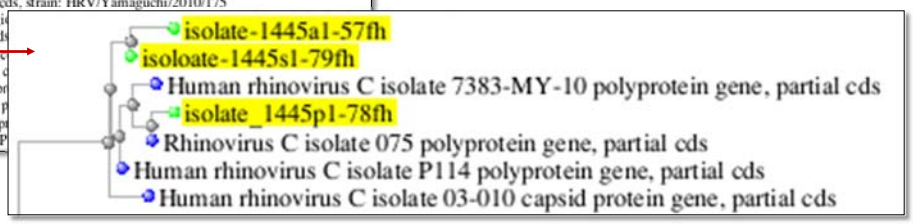
Example Searches

The first example is to establish the phylogeny of a set sequences from a clinical viral samples (**A**) using MOLE-BLAST. The search uses the NT database for its broad organism coverage, the Blastn program for increased sensitivity with “Max target sequences” set to 50 to increase the database reference matches saved. The “Show results in a new window” option is checked so input settings are available for subsequent adjustment and resubmission if necessary (**B**). Additional filtering of the database is done to ensure the quality of matches by excluding “Uncultured/environmental sample sequences,” selecting only sequences with “binomial” scientific names, and only database sequences from viruses are searched through the organism Entrez Query (**C**). The “Query Clustering Parameters” are left unchanged. The top 20 database sequences (**D**) out of the 50 saved in the BLAST steps are requested in the multiple sequence alignment and tree construction.





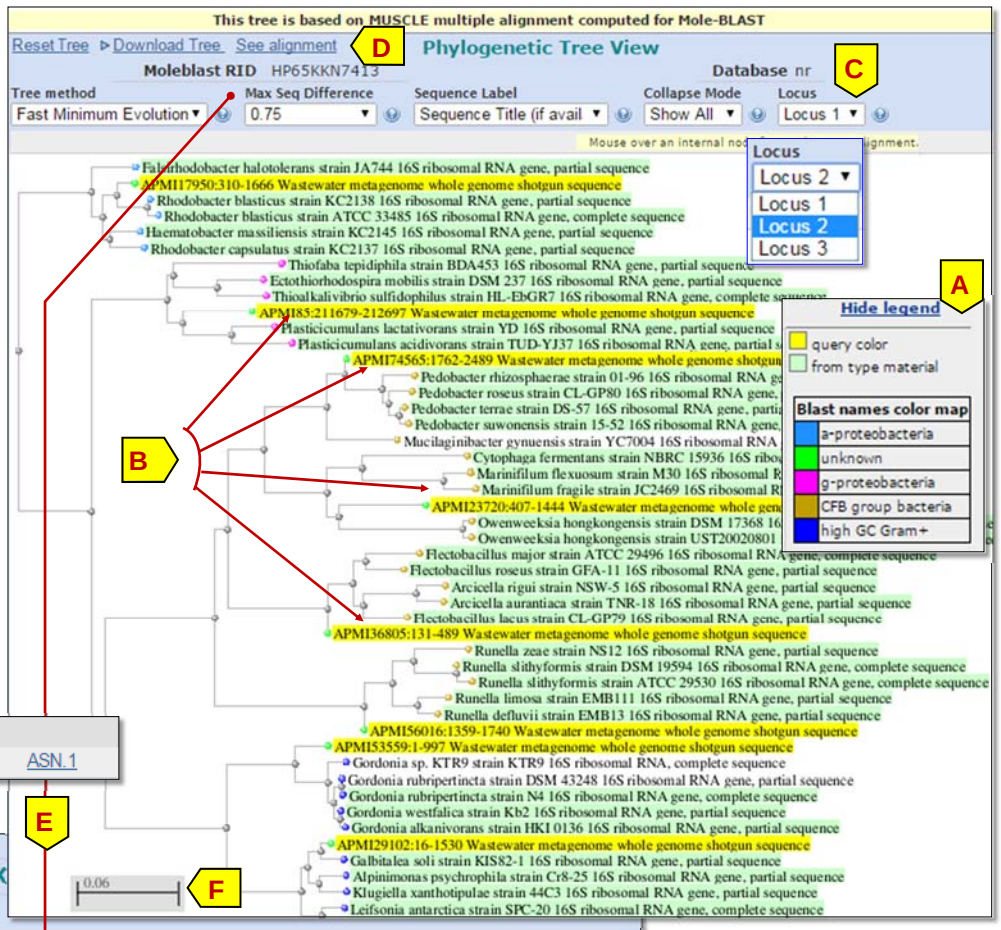
As shown to the left, MOLE-BLAST displays the result in TreeView format by default, with the query highlighted in yellow (**E**). The sequence color scheme indicated by the legend (**F**). The tree display can be adjusted by pull-down options at the top (**G**, from left to right): by selecting different methods, maximal sequence differences, different labels for sequences, collapsing or showing nodes, and showing different loci if the input queries are grouped into separate loci that cannot be merged. Links at the top of the page provide other functions (**H**, from left to right): clicking “Reset Tree” resets an adjusted tree display back to default; clicking the “Download Tree” displays link to tree files in different formats; and clicking the “See alignment” link displays the multiple sequence alignment, where the alignment can also be downloaded via a link at the top (not shown). The results indicate that the three sample are Human Rhinovirus C, closely related to previously reported isolates named as 7384-MY-10 and 075 (**G**).

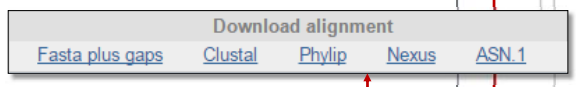


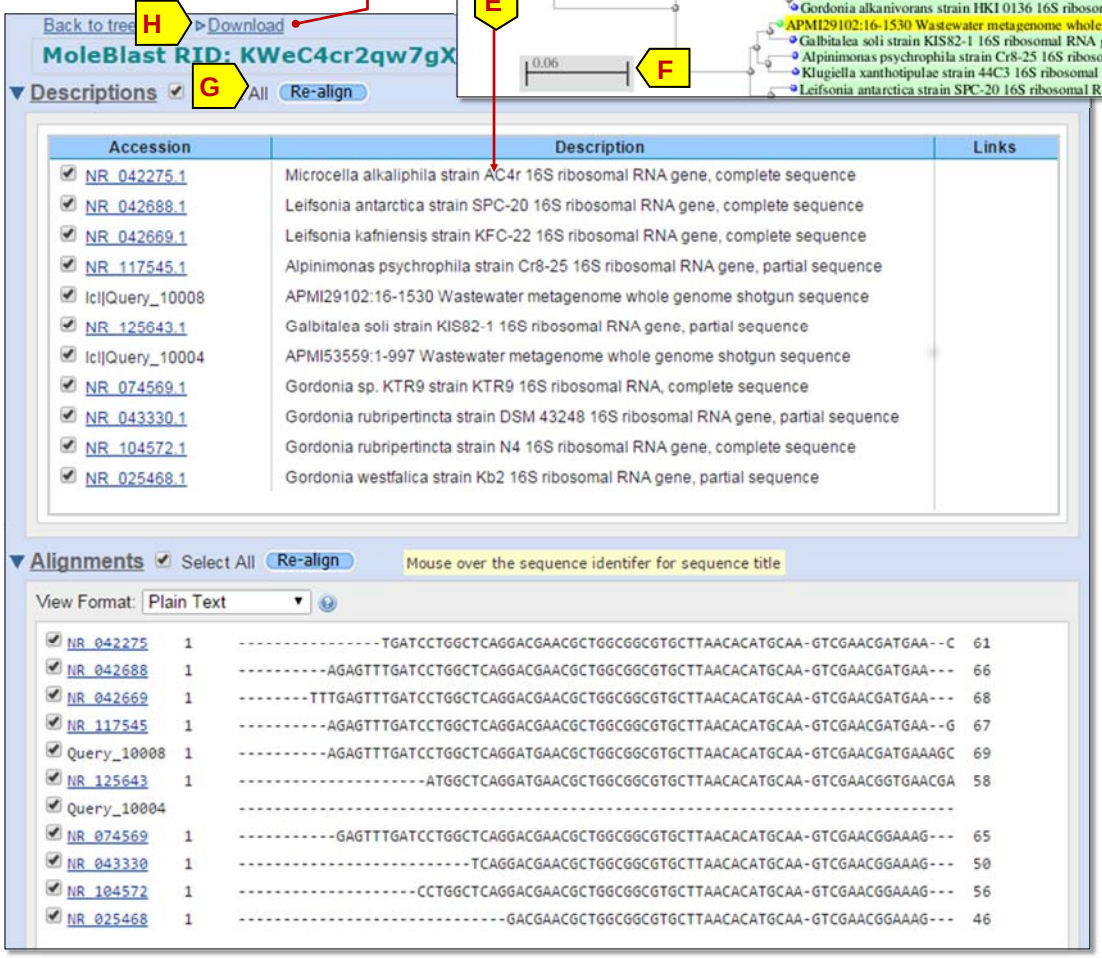
Example Searches (cont.)

The second example concerns identification of organisms present in wastewater samples, through 16S RNA sequences. Input sequences are uploaded through "Choose File" button and curated 16S RNA database is selected with other settings left at default (not shown).

In the resulting tree page (right), sequence titles are highlighted by background color to indicate if they are from type culture, with colored dot to indicate their major taxonomic group assignment (bacterial class), respectively. Color keys are listed in the Legend (**A**) shown to the right of the tree. The displayed group "Locus 1" contains eight of the eleven input queries (**B**). Results for other queries not displayed initially are available through the Locus pull down menu (**C**). Clicking the "See alignment" (**D**) link retrieves the underlying multiple alignment for the tree (**E**). The unit of the tree (**F**) represents expected number of changes per 100 bases. The tree is rooted at the middle of the longest edge.





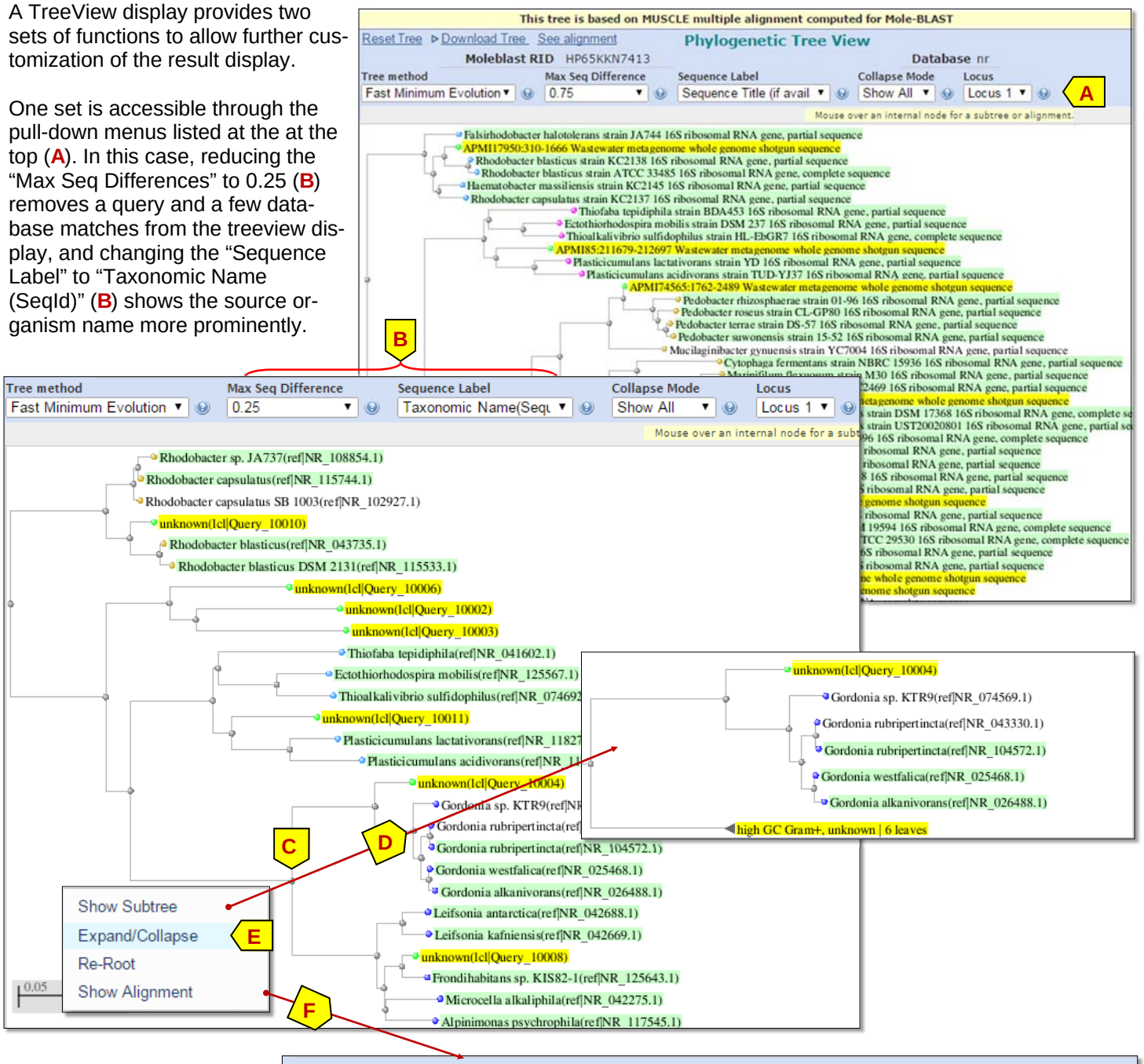


In the alignment view, the Descriptions section lists the sequence titles, and the Alignments section shows the details of aligned sequences (both shown in part to the left). The alignment can be redone using the "Re-align" button (**G**), after deselecting certain undesired entries using their checkboxes. The alignment file can be downloaded by clicking the "Download" link (**H**), which opens a section listing available formats.

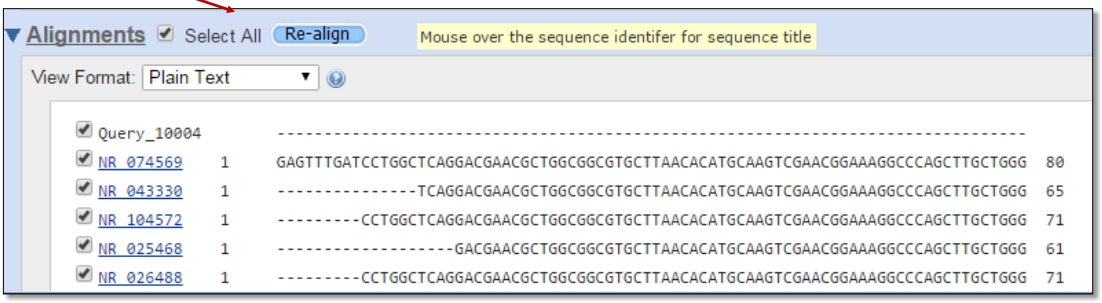
Functions Provided by TreeView Display

A TreeView display provides two sets of functions to allow further customization of the result display.

One set is accessible through the pull-down menus listed at the at the top (**A**). In this case, reducing the "Max Seq Differences" to 0.25 (**B**) removes a query and a few database matches from the treeview display, and changing the "Sequence Label" to "Taxonomic Name (SeqId)" (**B**) shows the source organism name more prominently.



Another set of functions is accessible through the context menus within the tree display (**C**). There are activated upon hovering a non-leaf (internal) node of the tree. Selecting "Show Subtree" option (**D**) shows only that selected tree branch in a zoomed-in view. Selecting the "Expand/Collapse" option (**E**) collapses the branch under that node. Selecting the "Show Alignment" option (**F**) changes the display to show the alignment for that branch.



Technical Assistance

Please send questions and feedback on MOLE-BLAST to the blast-help group at: blast-help@ncbi.nlm.nih.gov