Introduction

The greengenes web application provides access to the current and comprehensive chimera checked 16S rDNA prokaryotic multiple sequence alignment (prokMSA) for browsing, blasting, probing, and downloading. The data and tools presented by greengenes can assist the researcher in choosing phylogenetically specific probes, interpreting microarray results, and aligning/annotating novel sequences. If you are an ARB user, you can use greengenes to keep your own local database current.

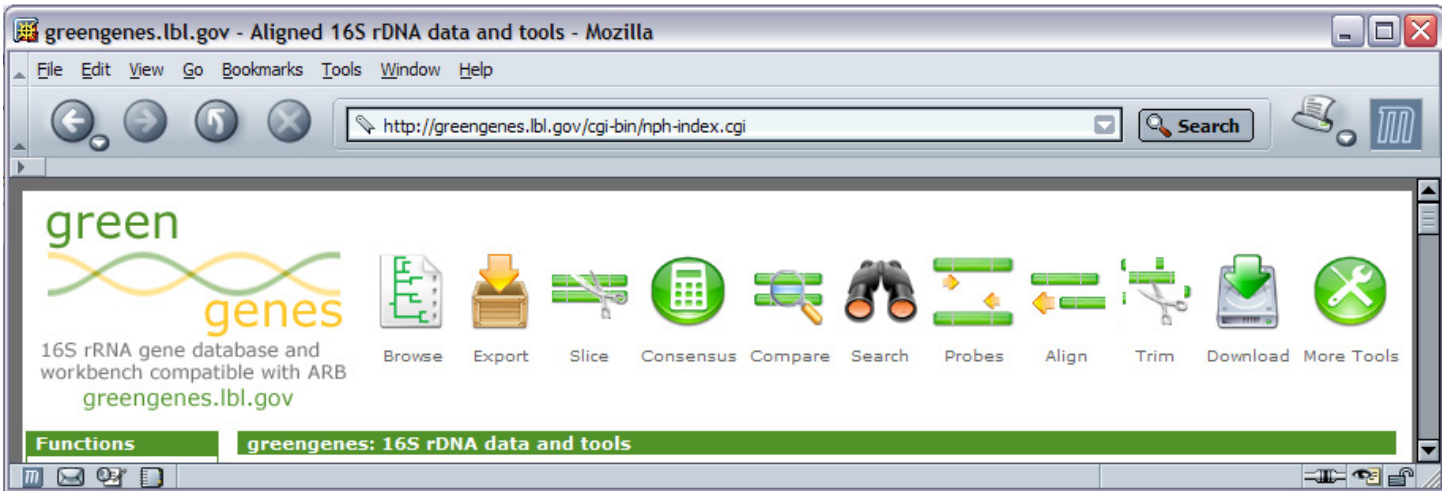
Greengenes is best viewed using recent browser releases - Internet Explorer (v6.0), Mozilla (v1.7.8), Firefox (v1.0.6), Safari (v2.0.1). Cookies should be enabled in order to utilize the 'My Interest List' function.

ProkMSA database

The prokMSA (or **prok**aryotic **m**ultiple **s**equence **a**lignment) database is a comprehensive multiple sequence alignment which includes all publicly available 16S rRNA gene sequences of substantial length (>1250 nucleotides). Chimeras are screened and removed from the dataset using Bellerophon 2 (Bel2) developed by Thomas Huber and Phil Hugenholtz. The resulting dataset is aligned using the NAST alignment algorithm as described in DeSantis et al. 2003 Bioinformatics, 19:1461-1468.

Greengenes tools

Greengenes provides a suite of tools designed to aid the researcher maintain up to date 16S rRNA gene databases and phylogenetic trees, in addition to providing tools for sequence quality control, primer and probe design and 16S microarray analysis.

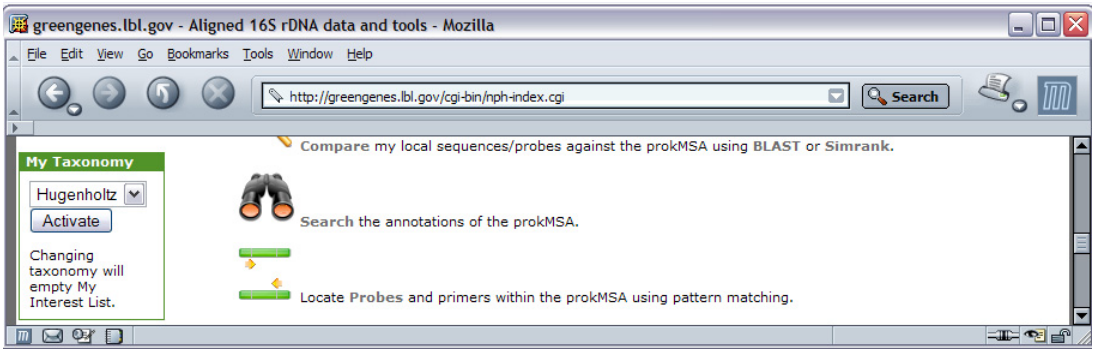


Browse

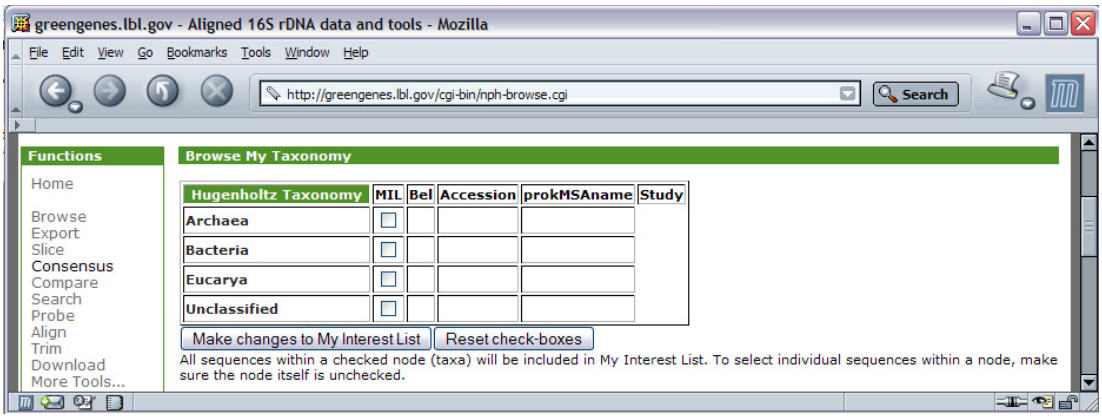
Use the browse function to navigate the complete prokaryotic 16S rRNA database according to six different taxonomies, RDP (ribosomal database project – bacteria only), NCBI (National Center for Biotechnology Information), G2-chip (taxonomy used in our high density phylogenetic 16S microarray), Pace (Norman Pace's Microbial Taxonomy), Ludwig (Wolfgang Ludwig's ARB taxonomy) and Hugenholtz (Phil Hugenholtz's modified Bergey's Taxonomy).

Tip: Click on taxonomic group to expand and browse. Click the check boxes and click 'Make changes to My Interest List' to add sequences to 'My Interest List'.

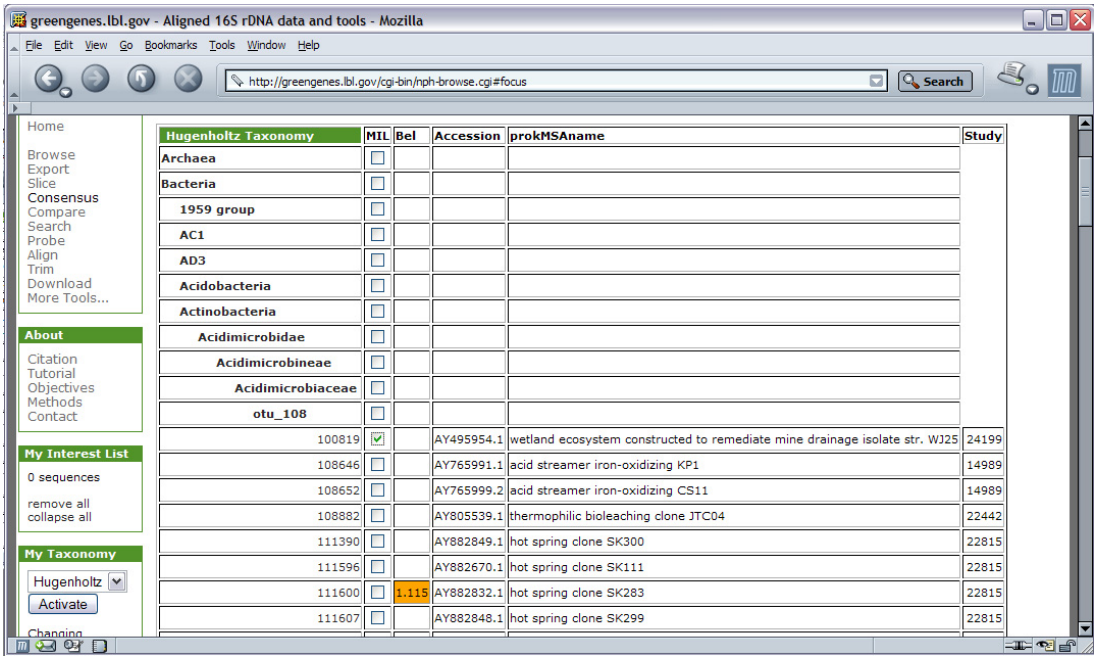
1. Choose taxonomy format in left panel. Hugenholtz, RDP, NCBI, Pace, Ludwig, G2_chip, then click ‘Activate’.



2. Then select browse beneath the "Functions" menu or on top icon bar.



3. Click on Domain of interest to expand e.g. Bacteria and continue to node of interest by clicking to expand



4. At any point, click on check box to add sequences below node to ‘My Interest List’ (MIL).
5. Update ‘My Interest List’ by clicking on Make changes to My Interest List or on Reset check-boxes to

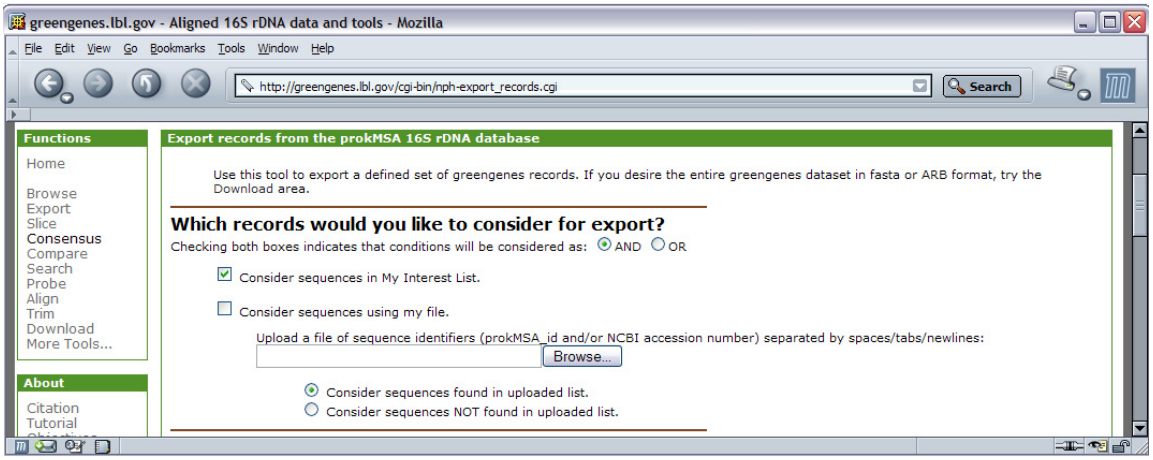
reset last checked boxes. You can remove all items in 'My Interest List' by selecting 'remove all' in the left panel under 'My Interest List'.

6. At this point, sequences in 'My Interest List' are available for other tools.
7. Note the column titled Bel, this shows the Bellerophon-2 divergence ratio. In our experience, ratios greater than 1.1 are a good indicator of a chimeric sequence.

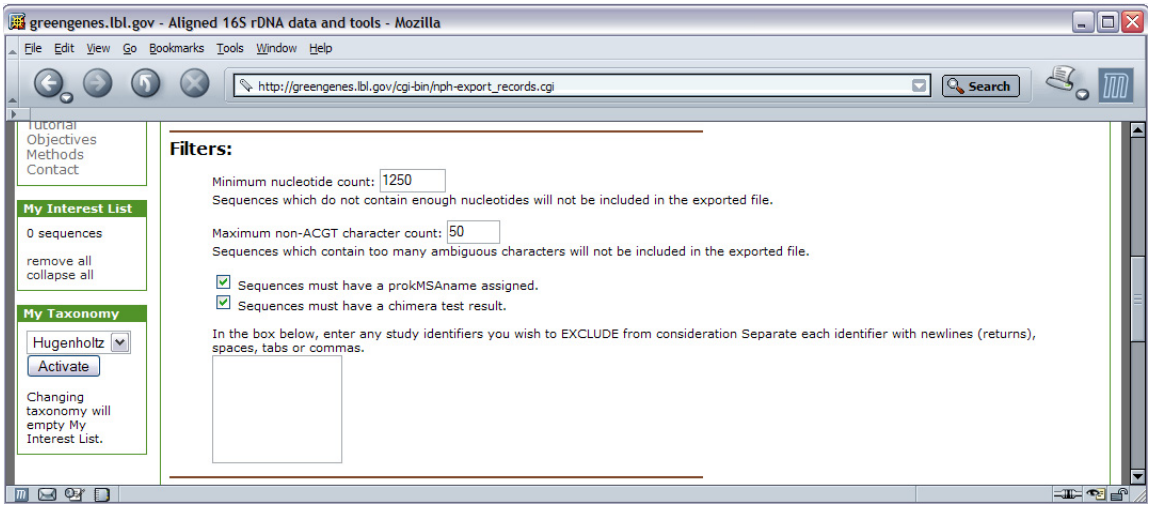
Tip: Clicking on an expanded node will collapse it again without altering selections within that node. This makes browsing more manageable.

Export

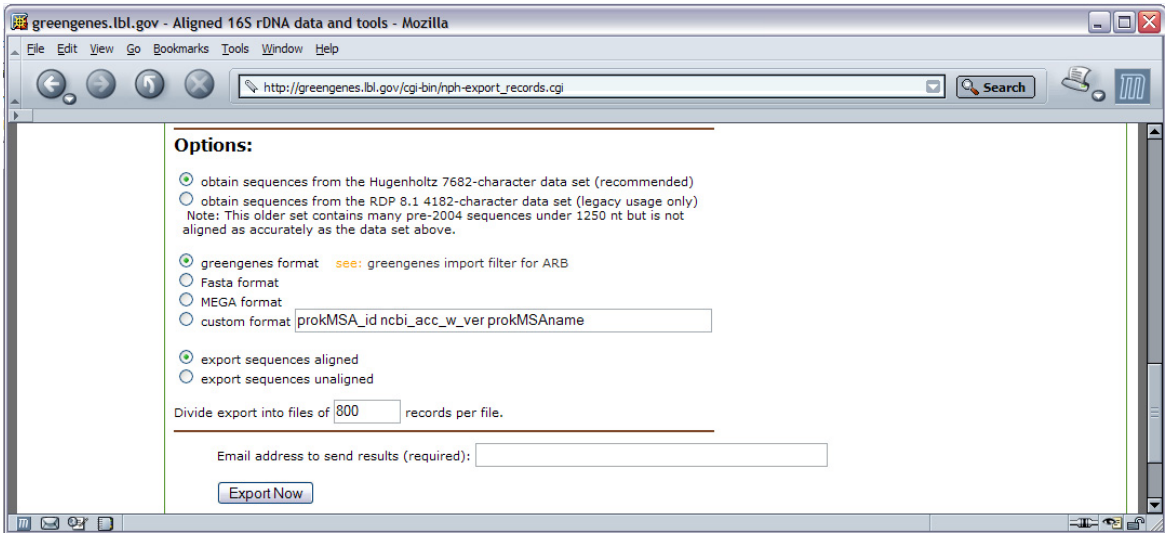
Export allows the selective download of specific sequences or groups of sequences using accession numbers, prokMSA ids or other unique identifiers. Users can chose to export sequences in 'My Interest List' or upload a list of identifiers in a text file for retrieval in aligned or unaligned formats. Output formats include 'Greengenes', 'Fasta' and 'MEGA'.



1. Select 'Export' from 'Functions' menu or top icon bar.
2. Chose sequences to export from database. You can export sequences in 'My Interest List' or chose sequences by uploading a file containing identifiers such as accession numbers.
3. Then chose filters for export e.g. number of nucleotides in a sequence, base call quality, chimera test data or specific studies you wish to exclude.



4. Now decide which dataset to export. Chose from the 7,682 character Hugenholtz dataset (recommended) or the old RDP8.1 4,182 character dataset.



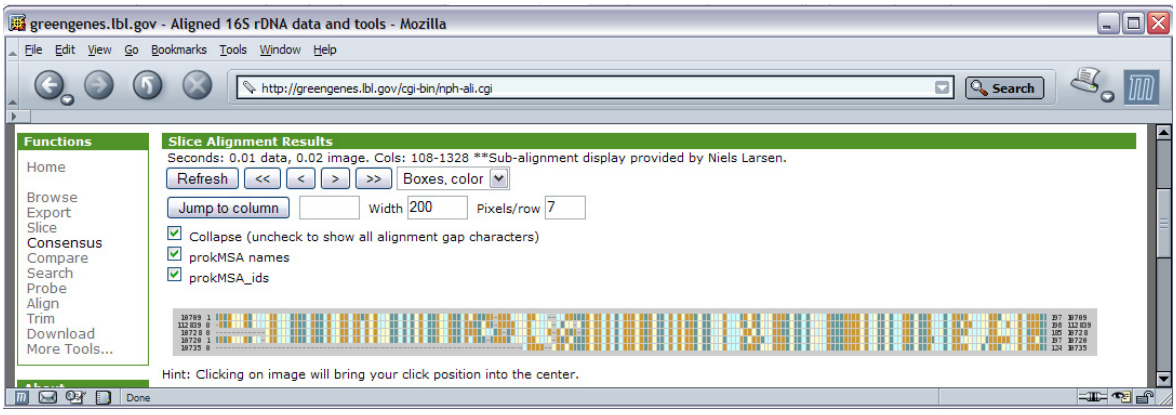
5. Chose the output format. Select from greengenes format (an ARB import filter is available), Fasta format or MEGA format. Custom format is for advanced users only.
6. Then decide if you want sequences aligned or unaligned and how many records to add to a single file to be delivered to you e-mail address.

Tip: This function can be used to update your existing database. Upload a list of accession numbers already in your database and chose 'Consider sequences **NOT** found in uploaded list', the returned file will contain all new sequences since your last update.

Slice

The slice tool allows the user to view sections of the prokMSA alignment. Sequences displayed in the alignment are those in "My Interest List'. Toggle between character (atgc) and color displays.

1. To view portions of the prokMSA alignment, slice by slice, you must first have your sequences of interest in 'My Interest List'.
2. Then select slice from the 'Function' menu or top icon bar.
3. The sequences in 'My Interest List' will be displayed.



4. Toggle display between colored alignments or character alignments by choosing from the 'Boxes, color' drop-down menu. Hit 'Refresh' after any changes to update
5. Chose to display names and/or ids.
6. Jump from one slice to another by column position.

Tip: Clicking on the image will center the alignment view.

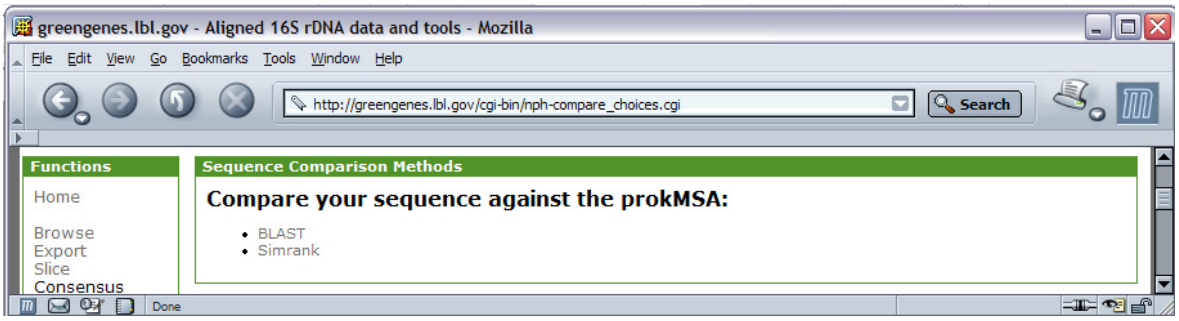
Consensus

Once functional (soon) this tool will allow users to view and export a consensus sequence using all sequences below a hierarchical position in the chosen taxonomy.

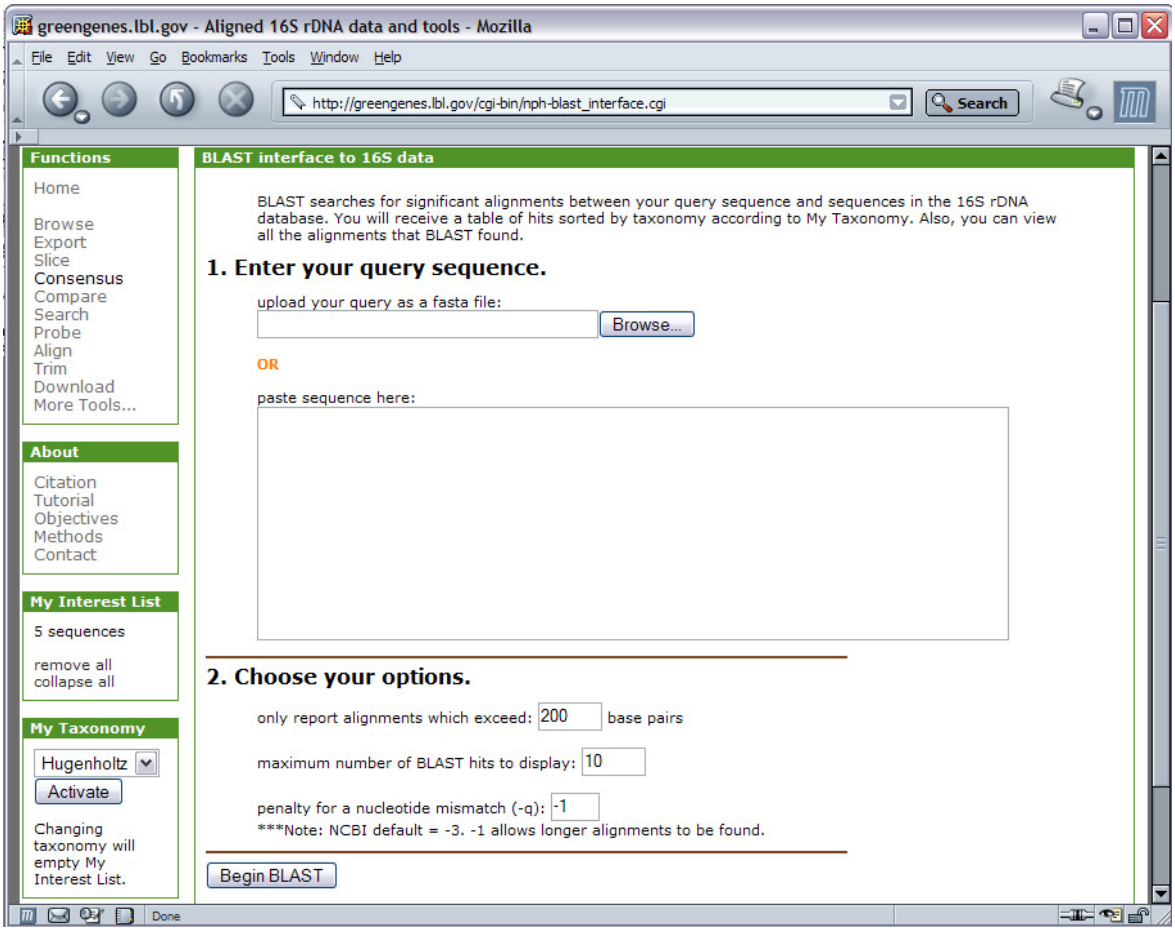
UNDER DEVELOPMENT

Compare

Compare your sequence to sequences in the prokMSA using SimRank or BLAST.



1. Select either BLAST or SimRank



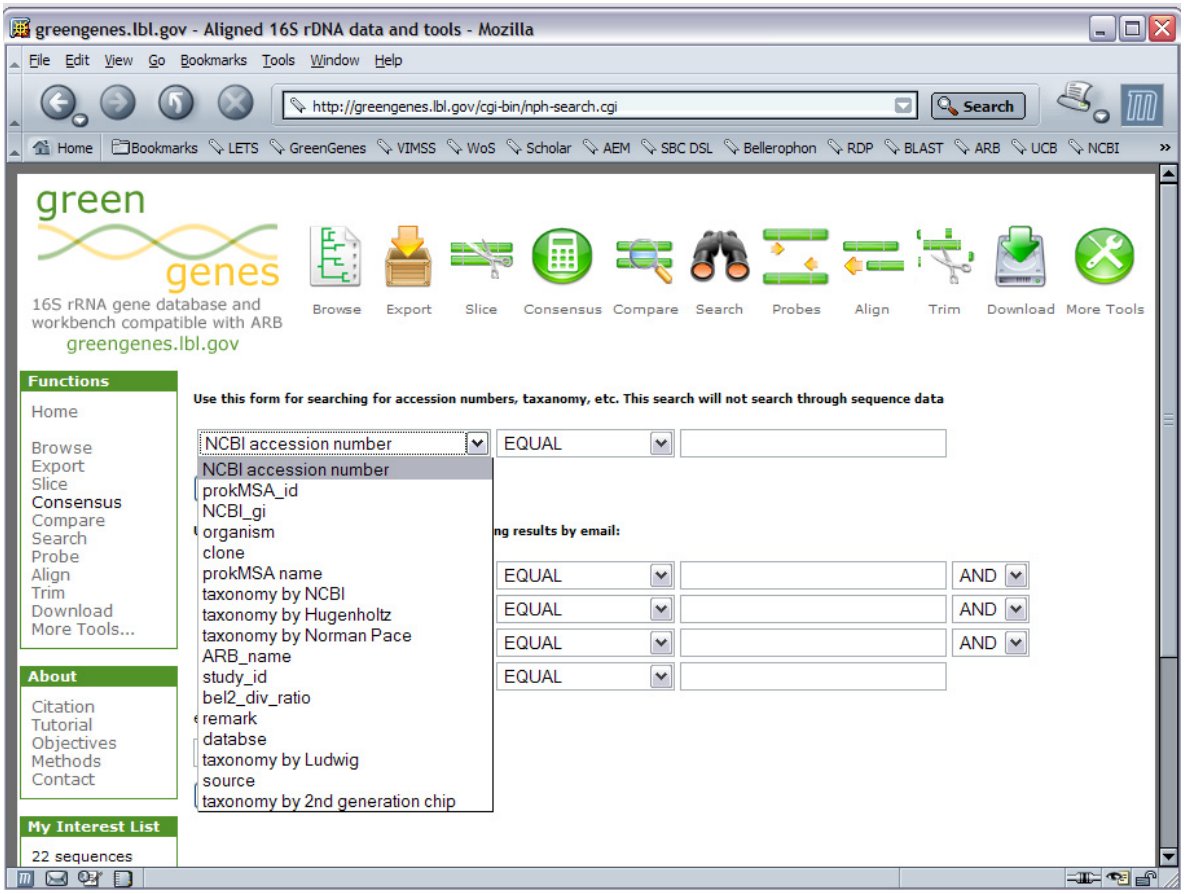
2. For BLAST or SimRank, upload (Fasta format) or paste a single sequence (this function does not currently support multiple sequence files).
3. Chose your BLAST/SimRank options and hit 'Begin BLAST' or 'Begin SimRank'.

Tip: Simrank finds the similarity between query sequence A and database sequence B in terms

of the number of unique 7-mers that they share, divided by the smallest total unique 7-mer count in either A or B.

Search

Search for sequence records by querying multiple fields such as accession number, isolation source, study id, organism name and many others.

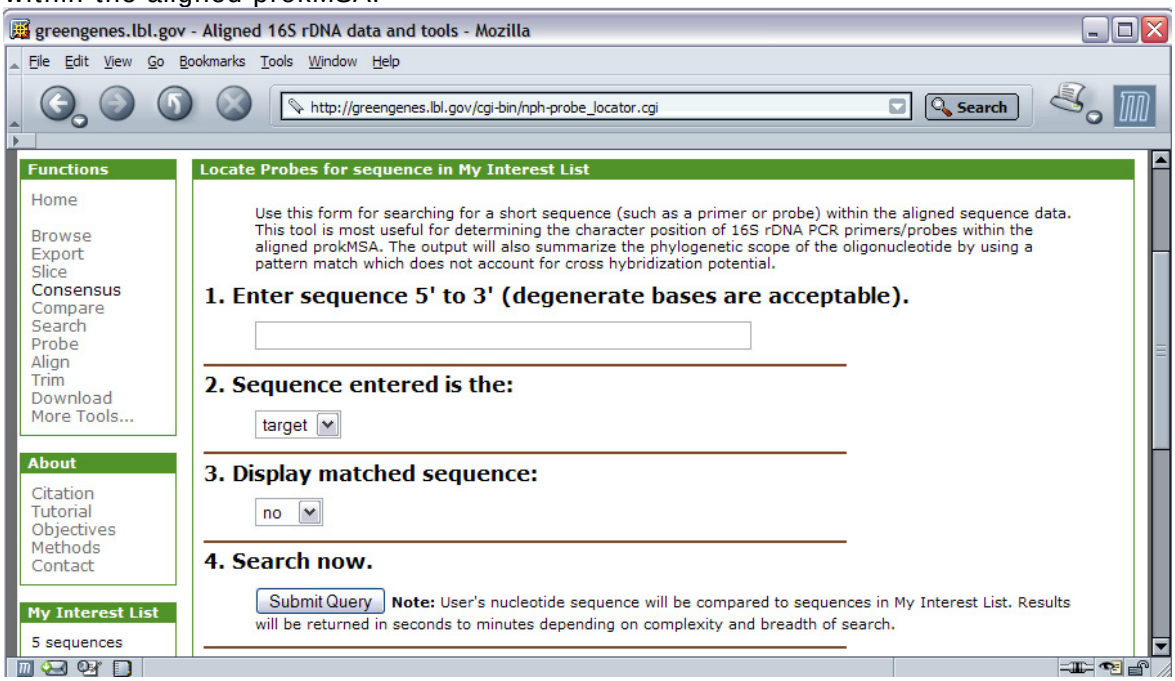


1. Chose from multiple search fields and combine searches.
2. Complex search results can be delivered to your e-mail inbox.

Tip: This search will not search through sequence data. Results will be organized according to ‘My Taxonomy’. Multiple search patterns are treated as "AND".

Probes

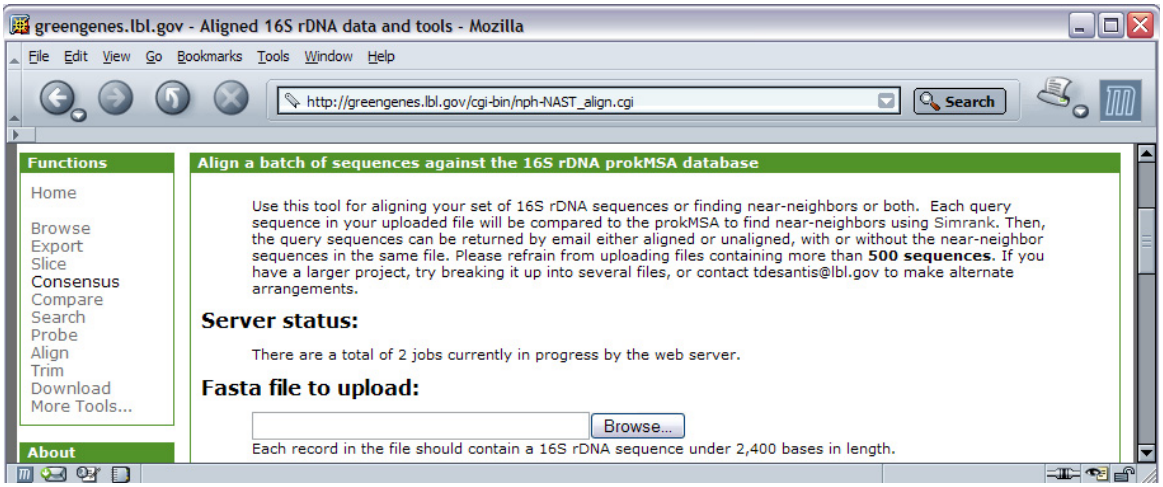
Use this tool for searching for a short sequence (such as a primer or probe) within the aligned sequence data. It is most useful for determining the character position of primers /probes within the aligned prokMSA.



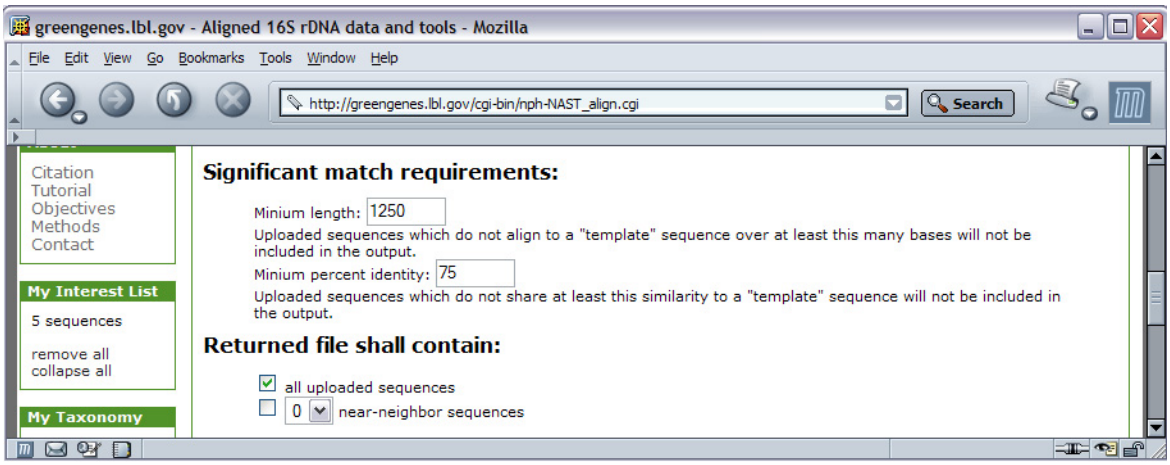
Tip: User's nucleotide sequence will be compared to sequences in 'My Interest List'.
Note: The output will also summarize the phylogenetic scope of the oligonucleotide using a pattern match not accounting for cross hybridization potential.

Align

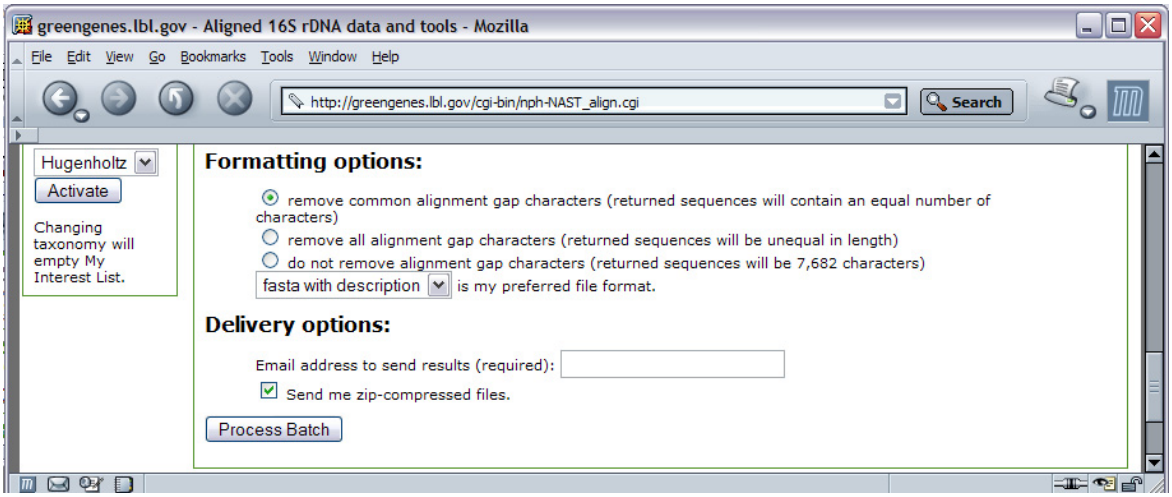
Use this tool for aligning your set of 16S rDNA sequences or finding near-neighbors or both. Each query sequence in your uploaded file will be compared to the prokMSA to find near-neighbors using Simrank. Then, the query sequences can be returned by email either aligned or unaligned, with or without the near-neighbor sequences in the same file.



1. First upload your Fasta file containing your sequences for alignment. Please limit the number of sequences to less than 500 per file.



2. Then select your match requirements.
3. You can choose to have returned data contain aligned sequences of your uploaded sequences AND up to 5 near-neighbor sequences.



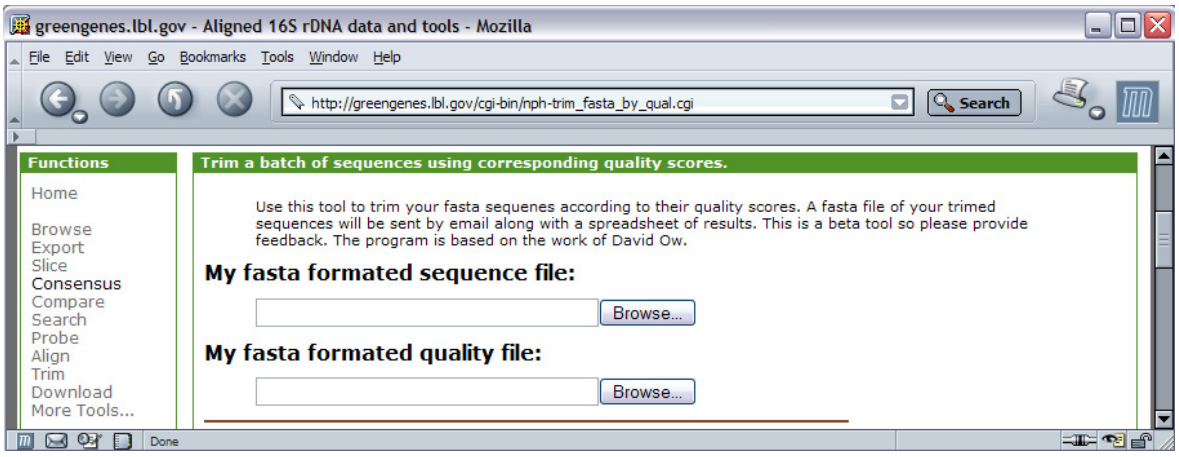
4. Select your formatting options. If you are using the greengenes.arb database do not remove common alignment gap characters. See How to use Greengenes with ARB in this tutorial.
5. Enter your e-mail address for result delivery and select 'Send me zip-compressed files' if you submitted a large (200-500 sequence) Fasta file.
6. Click 'Process Batch'.
7. A grey background screen will appear and alignment progress will be output. To kill the alignment job, simply close your browser window.

Tip: This tool is ideal for finding near-neighbors and obtaining an alignment file containing both query sequences and neighbors for fast tree generation. It is also very useful for aligning sequences for easy importation into a greengenes.arb-based ARB database.

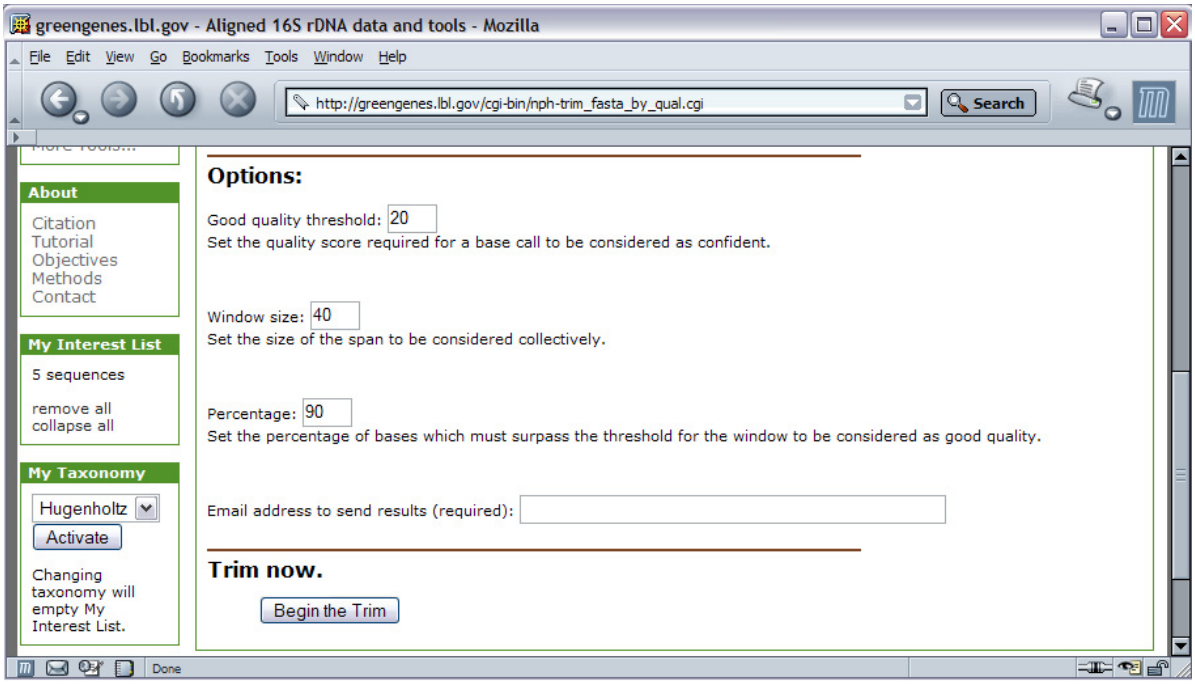
Note: Please limit number of sequences in upload to 500 or less.

Trim

Use this tool to trim your Fasta sequences according to their Phred quality scores. A Fasta file of your trimmed sequences will be sent by email along with a spreadsheet of results.



1. Upload your Fasta formatted sequence file, this may contain multiple Fasta formatted sequences.
2. Upload your Fasta formatted sequence quality file obtained from Phred.



3. Now select the quality threshold. This is the Phred quality (Q) value with which you would like to trim. The default is Q20. This indicates that there is a less than 1 in 100 chance ($1E^{-2.0}$) that the base call is incorrect.
4. Chose the window size of the span to be considered collectively for quality.
5. Chose the percentage of bases which must pass the threshold for a region to be considered 'high quality'.
6. Provide your e-mail address for results delivery.

Note: This is a beta tool so please provide feedback.
Note: see Phred documentation at http://www.phrap.org/

Download

The download section contains links to database data such as greengenes.arb in addition to protocols, publications, presentations, supplemental material and taxonomic outlines.

Tip: The sequence database link contains the prokMSA in fasta and greengenes.arb formats.

More tools

This section contains other tools in development.

Note: Some of these tools are not fully tested, are under development and may only be available sporadically.

How to use Greengenes with ARB

Greengenes.arb

greengenes.arb is based on an early ARB release (6mrz97.arb) that has a fixed alignment length of 7682 characters. Records downloadable from greengenes.lbl.gov use this fixed alignment as does the NAST tool for aligning user-submitted sequences.

The database contains all of the greengenes-curated records that are >1250 bp and have been chimera-checked using bellerophon. This constituted 75208 records as of October 27, 2005. Records are initially added to the all_tree using the parsimony insertion tool. Putative chimeras with divergence ratios > 1.1 are not added to the all_tree, but remain in the database so that they are searchable within ARB. Bootstrapped neighbor joining trees are then inferred for each phylum-level lineage and taxonomic assignments are checked by hand, with the exceptions of the Proteobacteria, Bacteroidetes, Firmicutes and Actinobacteria. Therefore, we cannot vouch for the taxonomic integrity of these four phyla in the current release of this database.

Up to 50 representative sequences are selected from each phylum-level lineage to construct a rep tree. Representative sequences have been flagged in the remark field with the word “rep”.

ProkMSA ids are used as the unique identifiers in the database. When updating greengenes.arb from the website with public records, make sure to select the “use old names” when you import the sequences. When importing your own sequences aligned with the NAST aligner use the “create new names” option. Never use the “generate new names” option in the species menu as this will erase the prokMSA ids from the name field and create difficulties if you want to overwrite existing records.

Greengenes contacts:

LBNL	Gary L. Andersen Center for Environmental Biotechnology Lawrence Berkeley National Laboratory 1 Cyclotron Road, Mail Stop 70A-3317 Berkeley, CA 94720 USA Email: GLAndersen@lbl.gov Phone: (510) 495-2795 Fax: (510) 486-7152	Todd DeSantis Email: tdesantis@lbl.gov Phone: (510) 761-6720 Eoin Brodie Email: ELBrodie@lbl.gov Phone: (510) 486-6584 Ping Hu Email: PHu@lbl.gov Phone: (510) 486 5908	Richard Phan Email: RPhan@lbl.gov Phone: (510) 486-7280 Keith Keller Physical Biosciences Division Lawrence Berkeley National Laboratory Email: KKeller@lbl.gov
JGI	Phil Hugenholtz Microbial Ecology Program DOE Joint Genome Institute 2800 Mitchell Drive Bldg 400-404 Walnut Creek, CA 94598 USA Email: phugenholtz@lbl.gov Phone: (925) 296-5725		
Danish Genome Institute	Niels Larsen Danish Genome Institute Gustav Wieds vej 10 C DK-8000 Aarhus C Denmark Email: nel@birc.dk		
Baylor University	Mark Rojas Department of Bioinformatics Baylor University PO Box 97356, 1311 S. 5th St. Waco, TX 76798-7356 USA Email: Mark_Rojas@baylor.edu		
University of Queensland	Thomas Huber Departments of Biochemistry and Mathematics The University of Queensland Brisbane Qld 4072 Australia Email: huber@maths.uq.edu.au		
Chalmers University of Technology, Sweden.	Daniel Dalevi Department of Computer Science Chalmers University of Technology SE-412 96 Göteborg, Sweden Email: dalevi@cs.chalmers.se		