A Web Tool for Assessing and Comparing Classifications and Taxonomies

R-018

Timothy G. Lilburn[1], Yuan Zhang[2] and George M. Garrity[2, 3]

[1]American Type Culture Collection, Manassas, VA 20110
[2]Dept. of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824
[3]Bergey's Manual Trust, Michigan State University, East Lansing, MI 48824

ABSTRACT

Recently, we developed an algorithm that builds self-organizing and self- correcting classifications. We have applied this algorithm to the problems arising from sequence annotation errors on prokaryotic classification. The comparison of the optimized classifications developed with our algorithm with other taxonomic proposals has allowed us to resolve outstanding problems in prokaryotic classification and taxonomy. To make such comparisons available to the research community, we have built a website that allows users to compare the current Bergey's Taxonomic Outline (DOI: 10.1007/bergeysoutline200310) with an optimized classification. The website serves as user interface to a dedicated analytic server, built using StatServer (Insightful). The application allows users to select the taxonomic group they are interested in, choose how they want the results to be organized (that is, at the species, genus or family level) and display the comparison. The organization of the compared classifications is visualized in the form of shaded evolutionary distance matrices. The colors of the matrix indicate the distances between the pairs of sequences in the matrix. The grouping of the colors in the matrix reflects the higher level groupings of the sequences (and, by extension, of the parent organisms). One matrix is arranged according to the hierarchy of the Outline and the other matrix is arranged according to the groupings generated by the classifier. Users can drill down in the display to see the comparisons at lower taxonomic levels or move up the hierarchy. The side-by-side comparison illuminates possible solutions to evident problems in the current classification. We illustrate how the taxonomy browser works by looking at the classsification and taxonomy of the *Archaea*.

INTRODUCTION

The Taxonomy Browser website gives users the opportunity to visualize the realtionships amongst the Prokaryotes in two relatively compact formats: principle components analysis (PCA) plots and heat maps. Users can browse the Taxonomic Atlas, which is made up of pre-computed plots and maps representing the phyla and lower taxa of the prokaryotes or they can use the Analytic Service to generate heat maps that illustrate the relationships amongst user-selected taxa. The relationships amongst the Prokaryotes are calculated using evolutionary distances among the small subunit (SSU) rRNA sequences. The sequences used have been carefully curated and can all be linked to a validly named prokaryote. This "vetted" set of sequences represents a valuable resource in itself.

THE DATA

Our analyses began with a set of roughly 14,000 SSU rRNA sequences gathered from GenBank, the Ribosomal Database Project, and the database provided with the ARB software suite. One of the major problems plaguing the use of 16S rDNA for deterministic purposes is the lack of a carefully vetted set of sequences, in which the taxonomic annotation was carefully reviewed and updated. We put a great deal of effort into linking sequence data with the organisms from which the sequences were obtained and into keeping pace with sequences from new species published in the International Journal of Systematic and Evolutionary Microbiology. The same degree of effort is put into keeping abreast of the opinions of the authors contributing to Bergey's Manual of Systematic Bacteriology. The result was a set of 6635 sequences (> 1399 nts, < 4% ambiguities) that could be identified as coming from type strains or from strains of validly named species. This first subset is identified as the "unresolved" set as there remained a number of taxonomic and nomenclatural errors within this data set. A second subset of 6377 sequences for which we could confirm identity and taxonomic placement is identified as the "resolved" set. There are still some likely placement errors in the resolved set that are indicative of misnamed species. These are predominantly within the phyla *Firmicutes* and *Actinobacteria*.

The evolutionary distances used to establish the relationships amongst the prokaryotes are based on an analysis of this SSU rRNA sequence data. Once gathered, the sequences are aligned in the ARB aligner, which allows users to take account of extensive secondary structure information about SSU rRNA when evaluating alignments. A mask is also applied to the alignments that restricts phylogenetic analysis to 1101 conserved positions in the Prokaryotic alignment.

Aligned, masked sequence sets are saved and distance matrices are generated for each set. Since the sequences represent the entire spectrum of prokaryotic diversity, a simple model of evolution, the Jukes-Cantor model, is used to correct the measured distances between the sequences. The distance matrices are concatenated and are the input data for the statistical treatment and exploratory data analyses that follow.

DESCRIPTION OF EDA PLOT TYPES

PCA plots - Dynamic Principal Components Analysis (PCA) plots highlight specific taxa in color, against a background of all taxa (global PCA plots) or subsets of taxa (phyla or class) based on predefined taxonomic criteria. The atlas presents a series of interactive two dimensional maps of the taxonomic space occupied by the Archaea and Bacteria and is based on the analyses that were done in writing the revision to the Road Map to the Manual of the . These maps provide a view of the temporospatial relationships among the major lineages of cultured and non-cultured species that are derived by performing a PCA on the covariance matrix computed from the evolutionary distances among the 16S rDNA genes. PCA finds uncorrelated composite measures in a multidimensional data set and allows expression of the original data in far fewer dimensions, without a significant loss of information. In the maps, each point represents the location of a given sequence within the derived coordinate system, in reference to all other samples within the data set.

For computational efficiency, we employ use an asymmetric matrix of pairwise evolutionary distances as input for our analysis. Each sequence (representing an individual species or strain) is treated as an independent variable and its evolutionary distance to 223 reference sequences (which we refer to as benchmarks) is computed. The benchmarks were selected based on their designation as type or reference strains for the 184 families in the described in the first revision of Bergey's Taxonomic Outline of the Procaryotes. Using this approach we are able to significantly reduce the dimensionality of the original data to yield 2-D and 3-D views that account for >85% of the total variance in the underlying data.

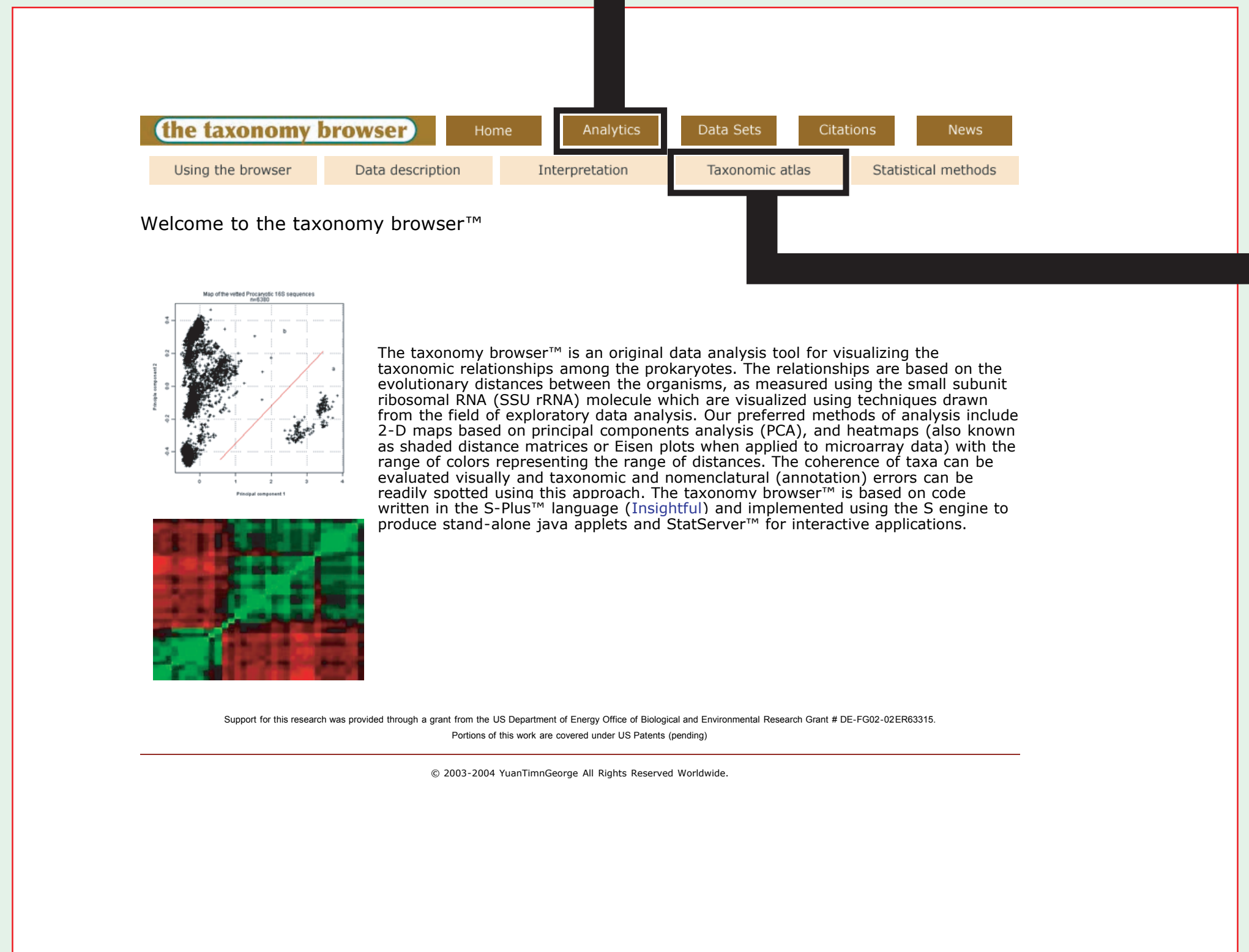
Placing the cursor over point(s) of interest will provide information about the identity of the point (current name, RDP-ID, higher taxon). For the purpose of comparison, we have included plots derived from two separate data sets. The unresolved plots contain a number of unresolved taxonomic and annotation errors, and many previously ~~unresolved~~ synonymies. The resolved plots were derived from a subset of those sequences in which the taxonomic and annotation errors were corrected and the synonymies were resolved to coincide with the taxon to which the source organism most likely belongs based on SSU rDNA sequence similarity.

Global PCA plots - The global PCA plots present a view of the two prokaryotic domains and are generated from the complete data set using all 223 benchmarks. To maintain a consistent perspective, taxa are selected at the phylum or class level and overlayed back onto the base plot. The identity of the highlighted points can be viewed by placing the cursor directly over those of interest. In all cases, global plots are for the first two principal components, which account for > 85% of the total variance within the data set.

Phylum- and class-level PCA plots - In some instances, adequate resolution of species is not possible within the global PCA plots. Reasons for this include overlap of point within a given 2-D space, point occlusion, or insufficient variability within the 16S rDNA to provide meaningful separation using the global coordinate system. In such cases, it is useful to recompute the principal components of a subset of sequences and benchmarks, selected based on taxonomic affilitation. Sepa ration and visualization of subgroups is enhanced by "rotating" the plots, which is accomplished by using various two-way combinations of the first, second and third principal components.

Heat maps - This method of visualization is simple; a distance matrix consists of a two dimensional table that contains pairwise evolutionary distances among a set of N taxa. The minimum and maximum values are then determined and numerical values are transformed into a number of discrete sub-ranges to which color/hue combinations are assigned.

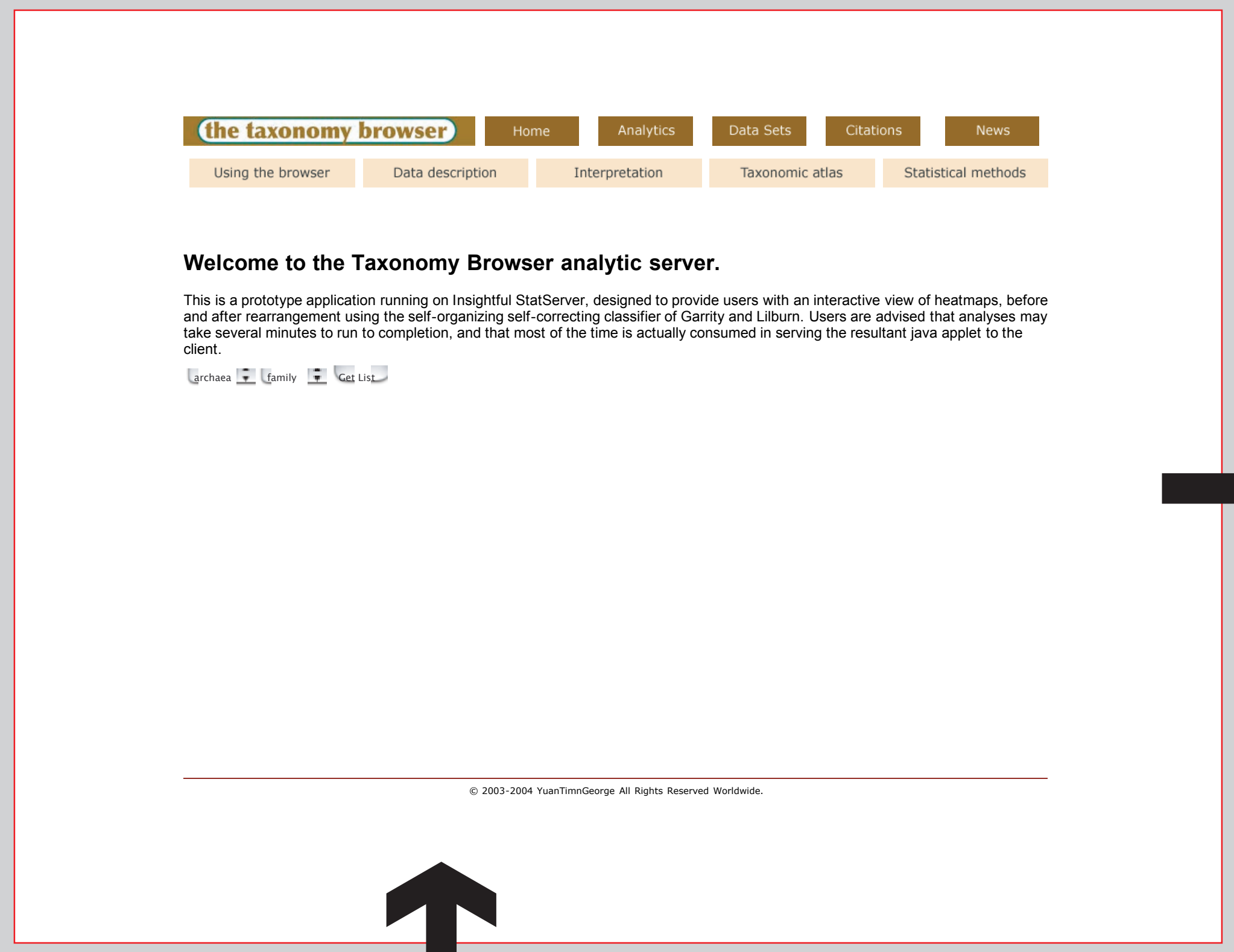
Dynamic heatmaps - To view interactive heatmaps (which are accessible from the taxonomic atlas and analytics pages), simply place the cursor over the area of interest. A legend for the pair of points, or a larger region (in the analytics) will appear in the upper right hand corner of the plotting region. S-Plus graphlets support zooming and allow visualization of regions of interest in greater detail as well.Non-optimized, dynamic heatmaps of subsets of the data used in PCA analysis were generated to help explain positioning of individual points in some plots. Note that the scale and coloration changes from on heatmap to the next.



Welcome to the taxonomy browser™

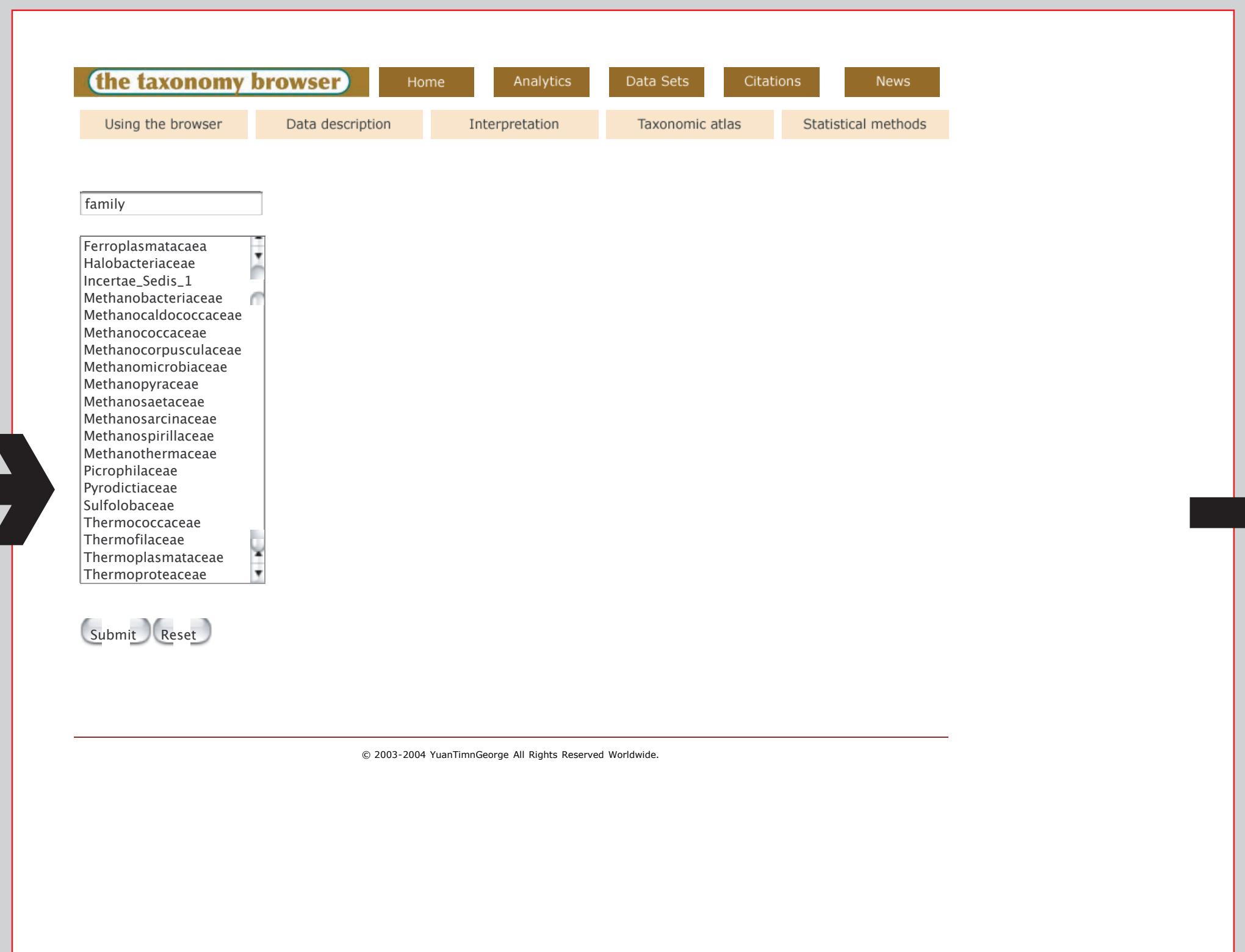
The taxonomy browser™ is an original data analysis tool for visualizing the taxonomic relationships among the prokaryotes. The relationships are based on the evolutionary distances between the organisms, as measured using the small subunit ribosomal RNA (SSU rRNA) molecule which are visualized using techniques drawn from the field of exploratory data analysis. Our preferred methods of analysis include 2-D maps based on principal components analysis (PCA), and heatmaps (also known as shaded distance matrices or Eisen plots when applied to microarray data) with the range of colors representing the range of distances. The coherence of taxa can be evaluated visually and taxonomic and nomenclatural (annotation) errors can be readily spotted using this approach. The taxonomy browser™ is based on code written in the S-Plus™ language (Insightful) and implemented using the S engine to produce stand-alone java applets and StatServer™ for interactive applications.

Support for this research was provided through a grant from the US Department of Energy Office of Biological and Environmental Research Grant # DE-FG02-02ER63315.
Portions of this work are covered under US Patents (pending)



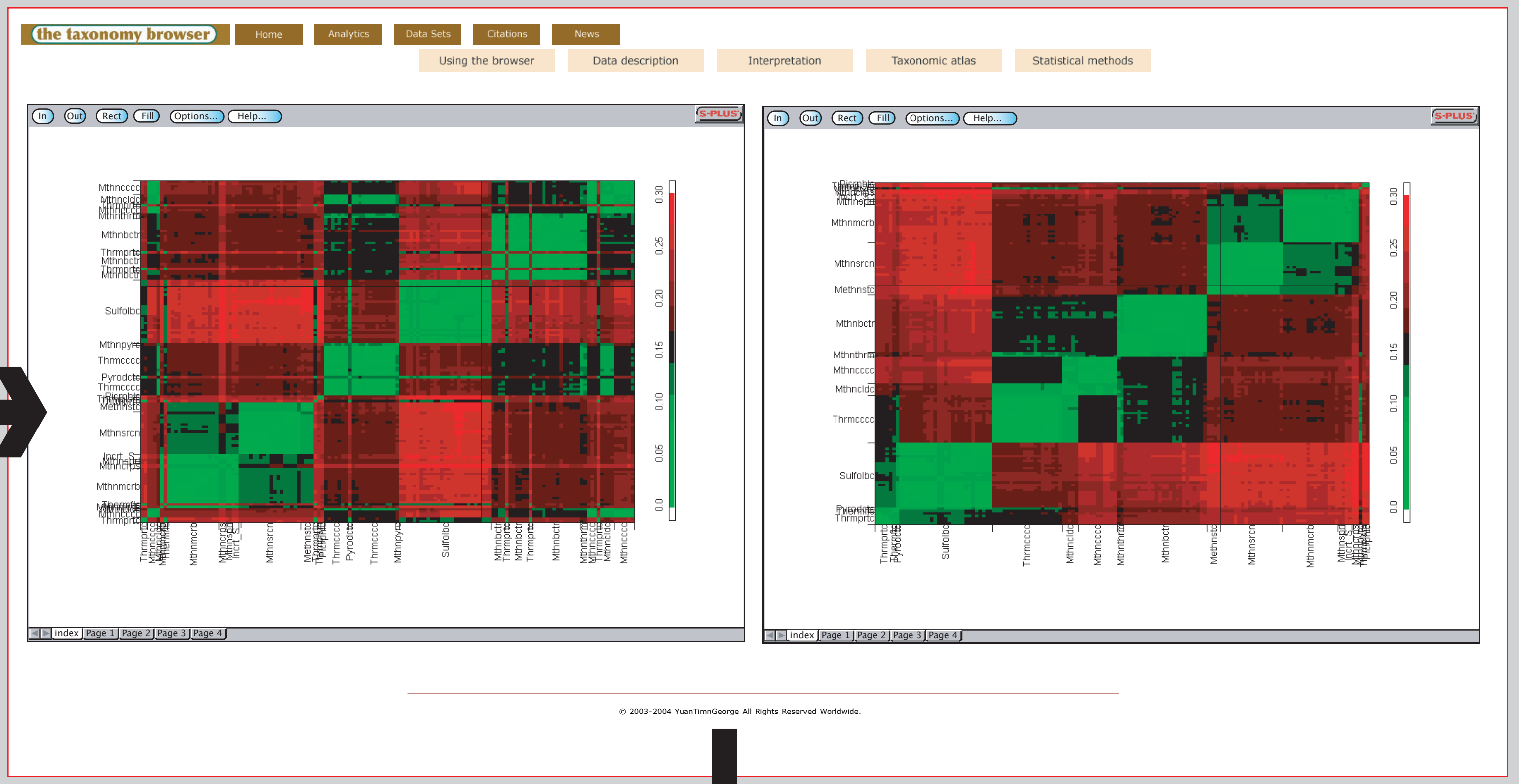
Welcome to the Taxonomy Browser analytic server.

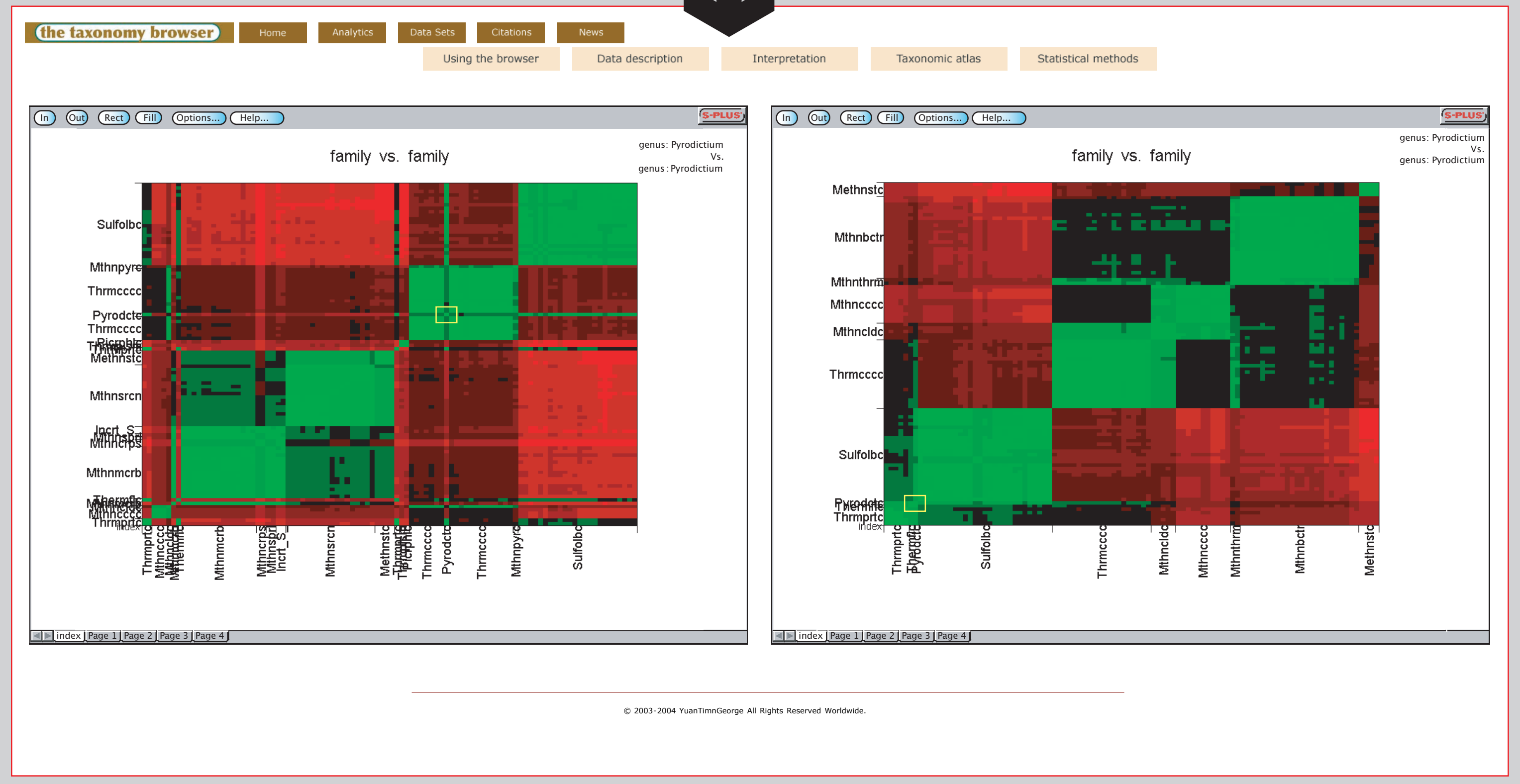
This is a prototype application running on Insightful StatServer, designed to provide users with an interactive view of heatmaps, before and after rearrangement using the self-organizing self-correcting classifier of Garrity and Lilburn. Users are advised that analyses may take several minutes to run to completion, and that most of the time is actually consumed in serving the resultant java applet to the client.

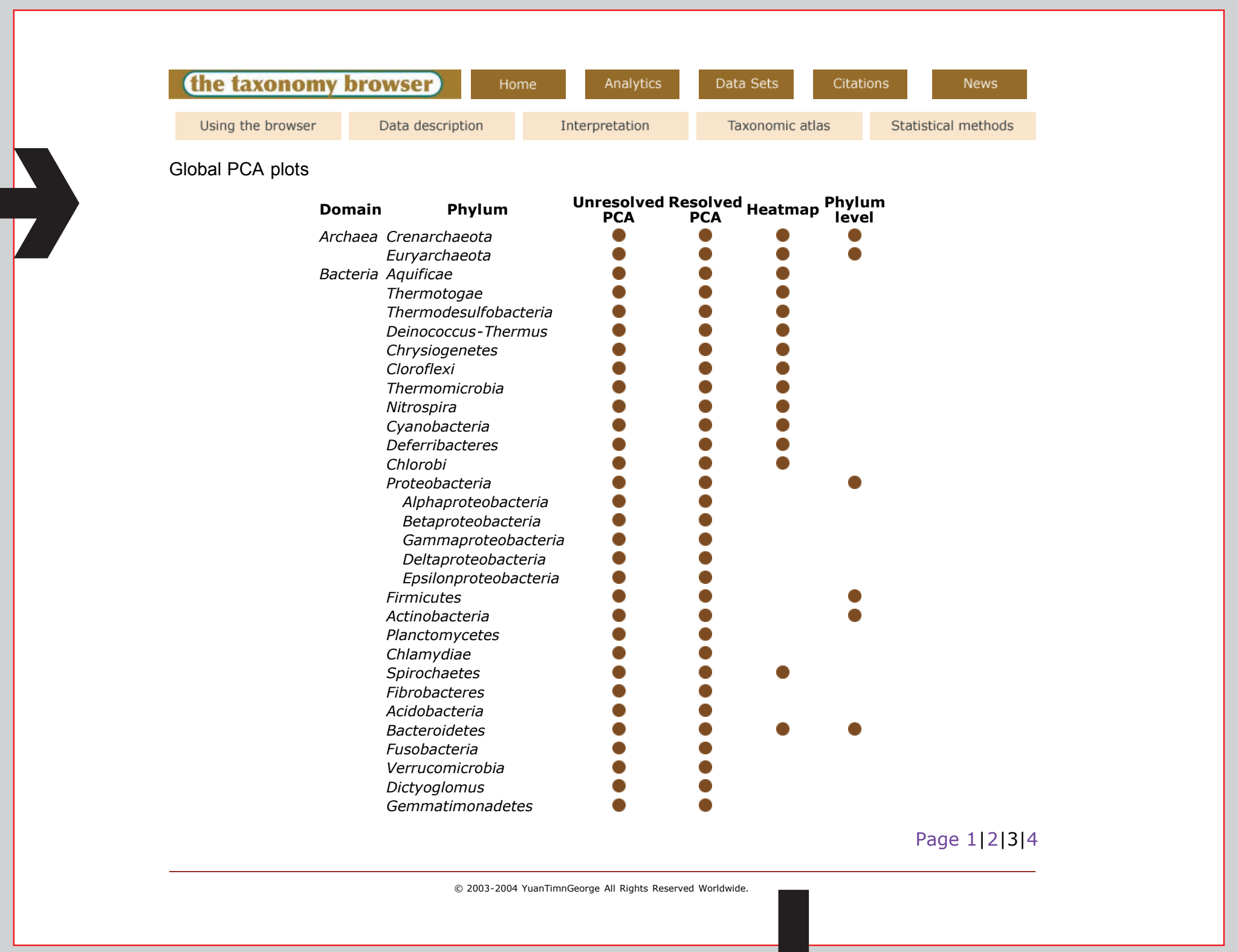


family

Ferroplasmatacaea
Halobacteriaceae
Incertae_Sedis_1
Methanobacteriaceae
Methanocaldococcaceae
Methanococcaceae
Methanocorpusculaceae
Methanomicrobiaceae
Methanopyraceae
Methanosaetaceae
Methanosarcinaceae
Methanospirillaceae
Methanothermaceae
Picrophilaceae
Pyrodictiaceae
Sulfolobaceae
Thermococcaceae
Thermofilaceae
Thermoplasmataceae
Thermoproteaceae



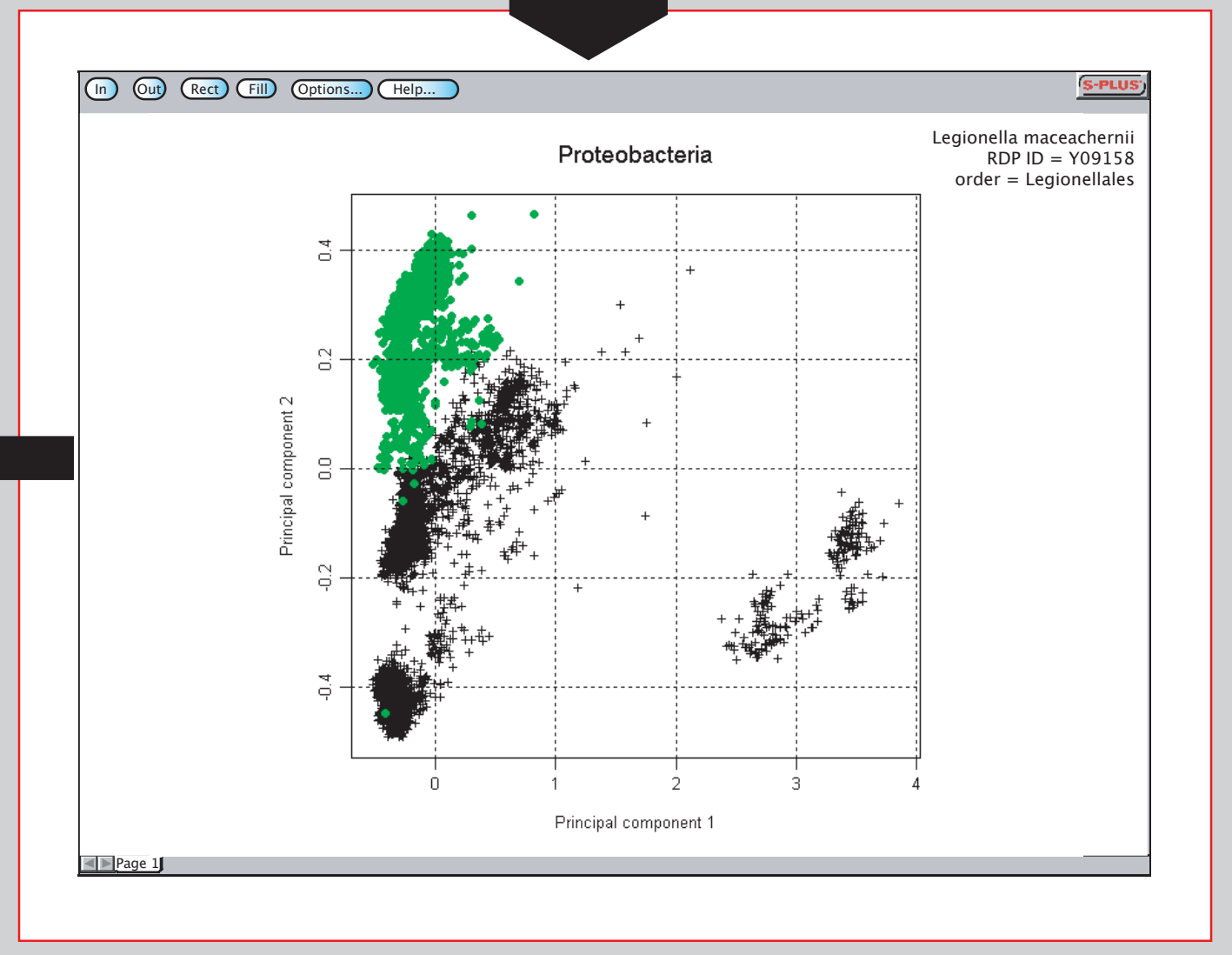


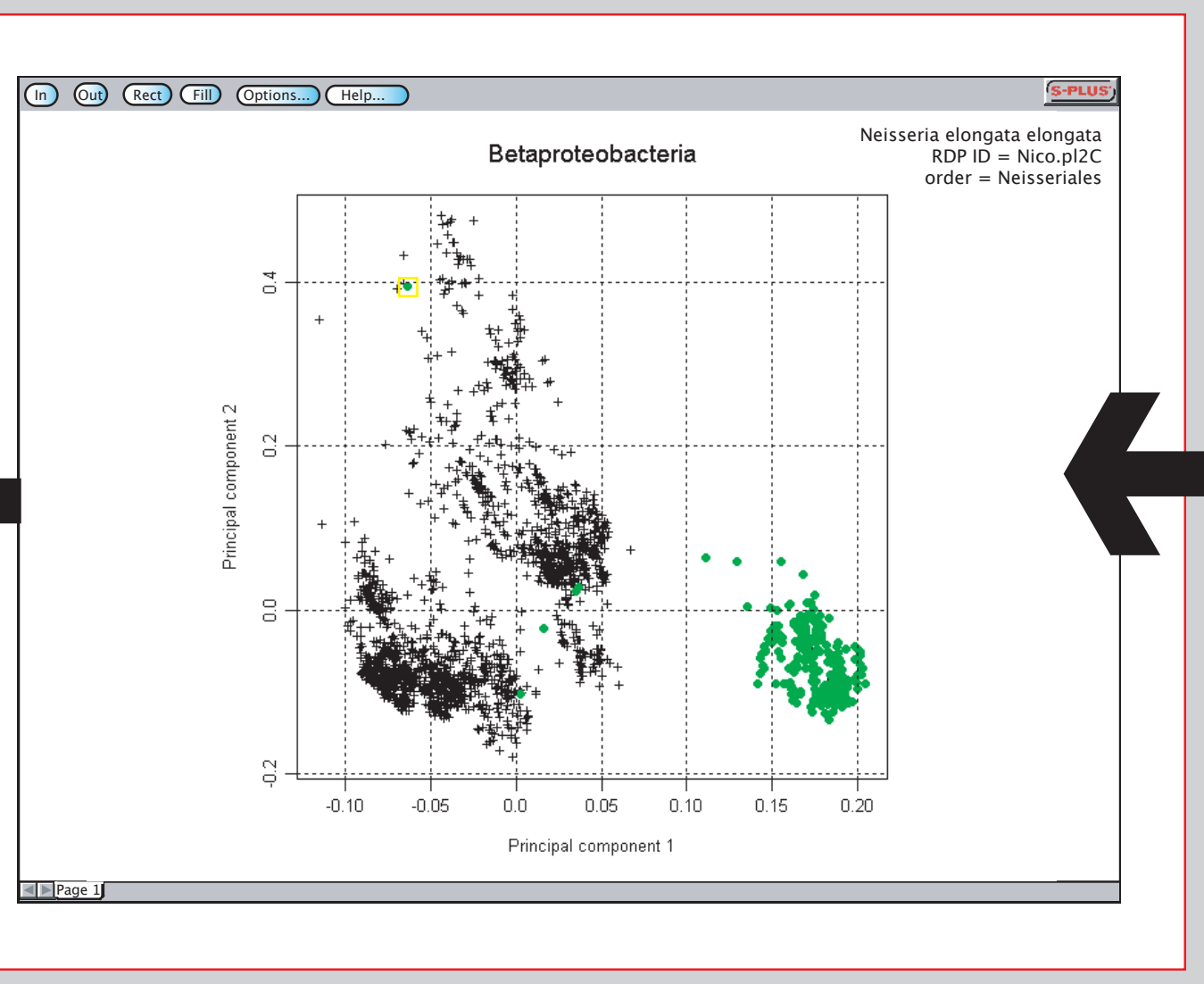


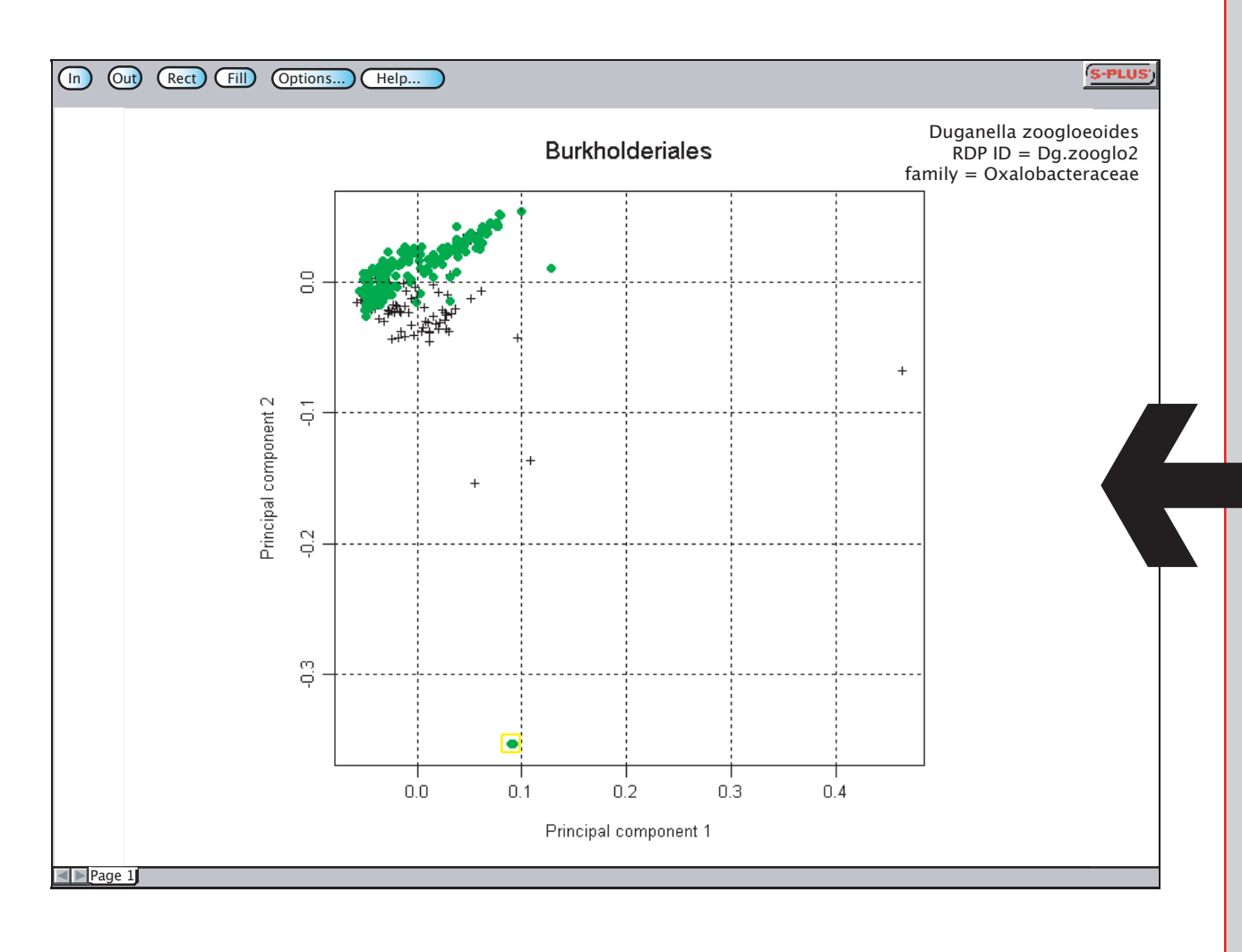
Global PCA plots

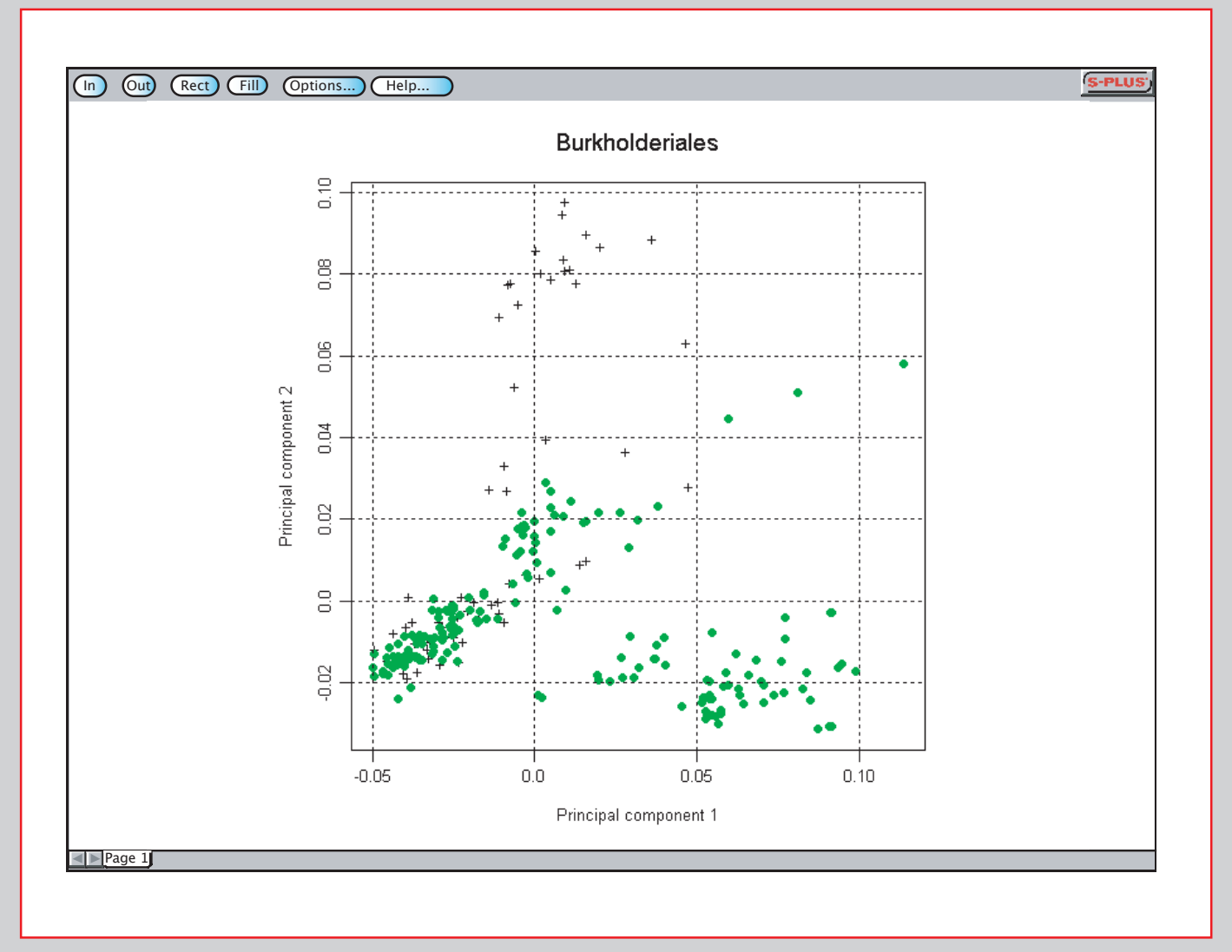
Domain	Phylum	Unresolved PCA	Resolved PCA	Heatmap	Phylum level
Archaea	*Crenarchaeota*	●	●	●	●
	Euryarchaeota	●	●	●	●
Bacteria	*Aquificae*	●	●	●	
	Thermotogae	●	●	●	
	Thermodesulfobacteria	●	●	●	
	Deinococcus-Thermus	●	●	●	
	Chrysiogenetes	●	●	●	
	Cloroflexi	●	●	●	
	Thermomicrobia	●	●	●	
	Nitrospira	●	●	●	
	Cyanobacteria	●	●	●	
	Deferribacteres	●	●	●	
	Chlorobi	●	●	●	
	Proteobacteria	●	●		●
	Alphaproteobacteria	●	●		
	Betaproteobacteria	●	●		
	Gammaproteobacteria	●	●		
	Deltaproteobacteria	●	●		
	Epsilonproteobacteria	●	●		
	Firmicutes	●	●		●
	Actinobacteria	●	●		●
	Planctomycetes	●	●		
	Chlamydiae	●	●		
	Spirochaetes	●	●	●	
	Fibrobacteres	●	●		
	Acidobacteria	●	●		
	Bacteroidetes	●	●	●	●
	Fusobacteria	●	●		
	Verrucomicrobia	●	●		
	Dictyoglomus	●	●		
	Gemmatimonadetes	●	●		

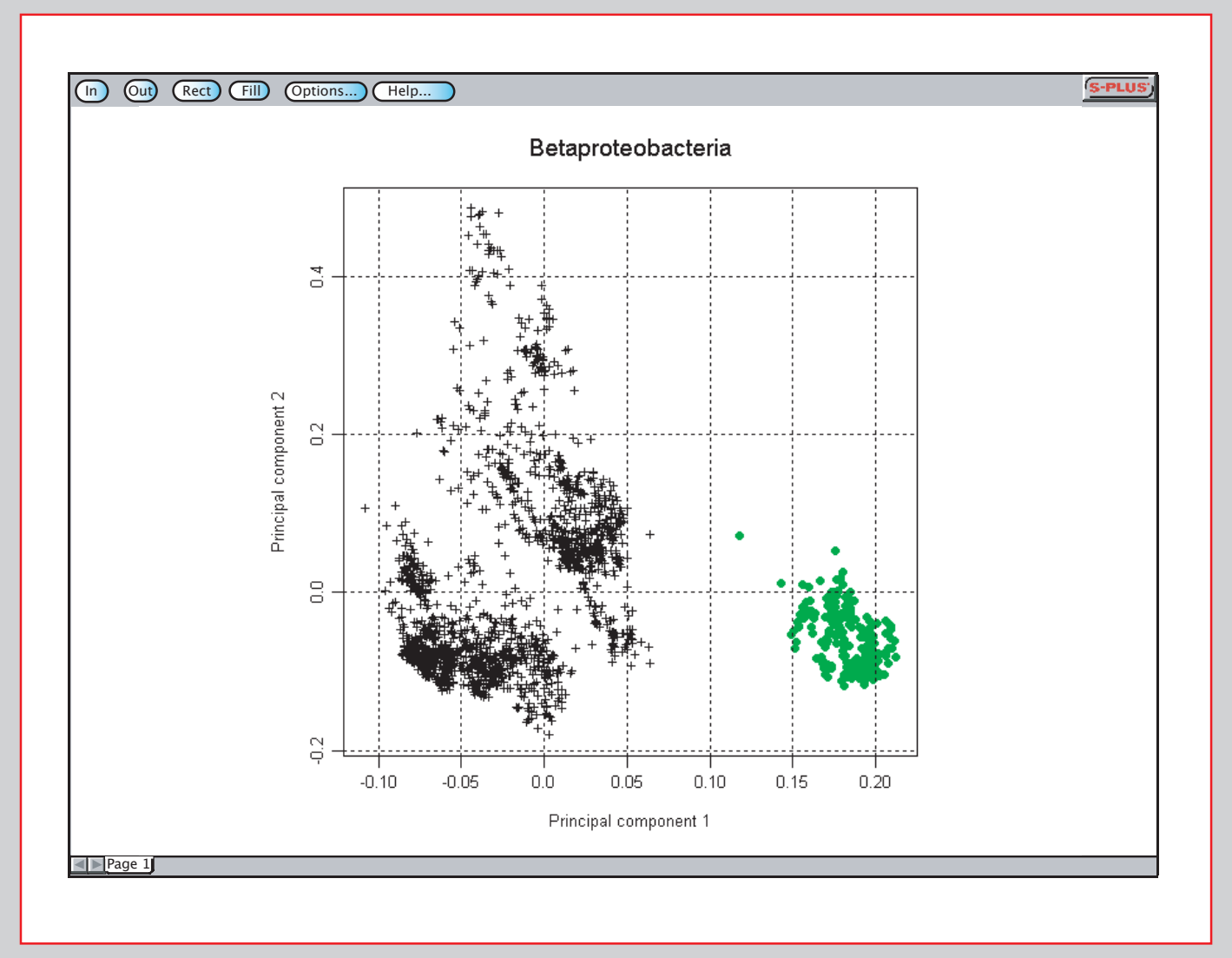
Page 1|2|3|4

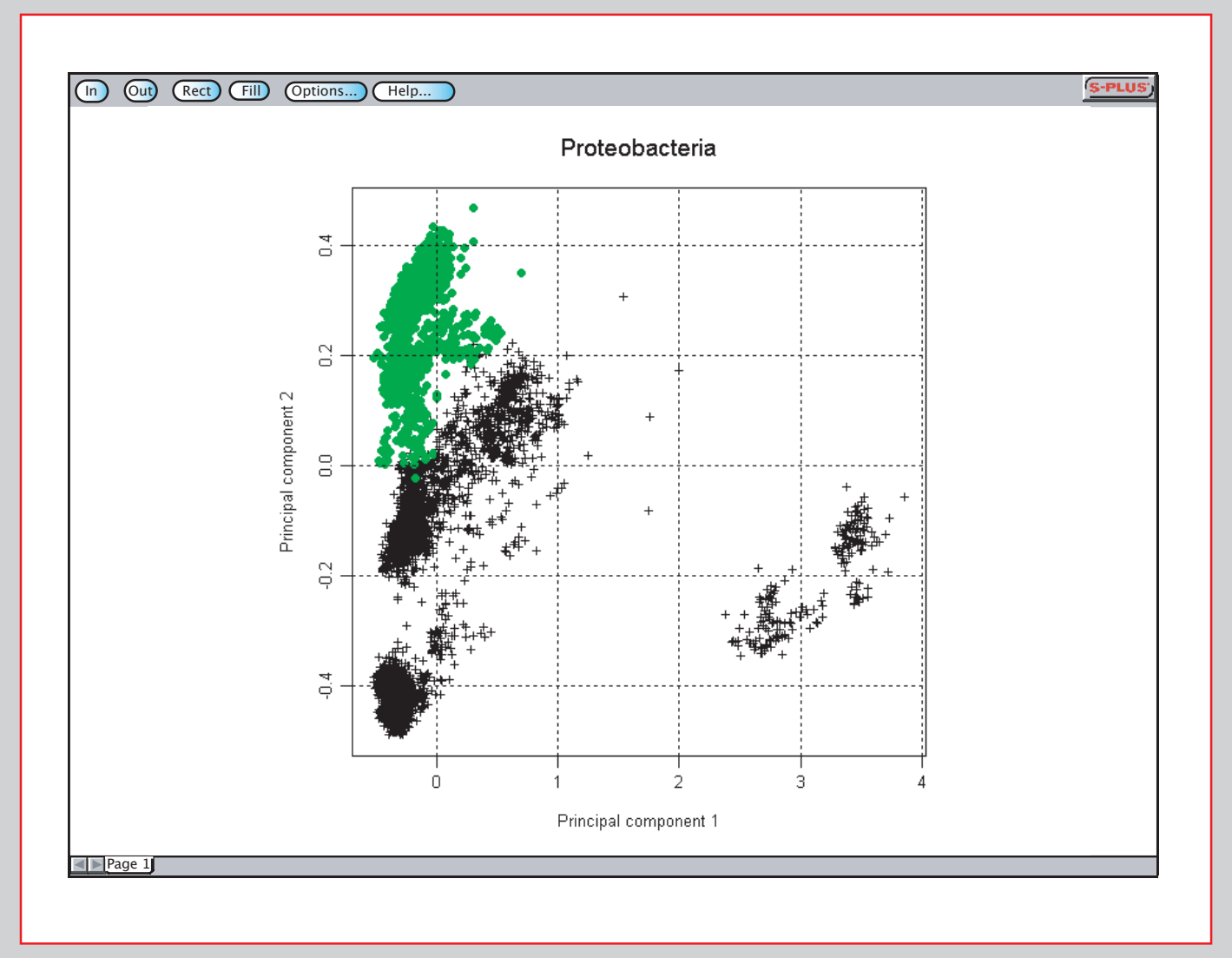




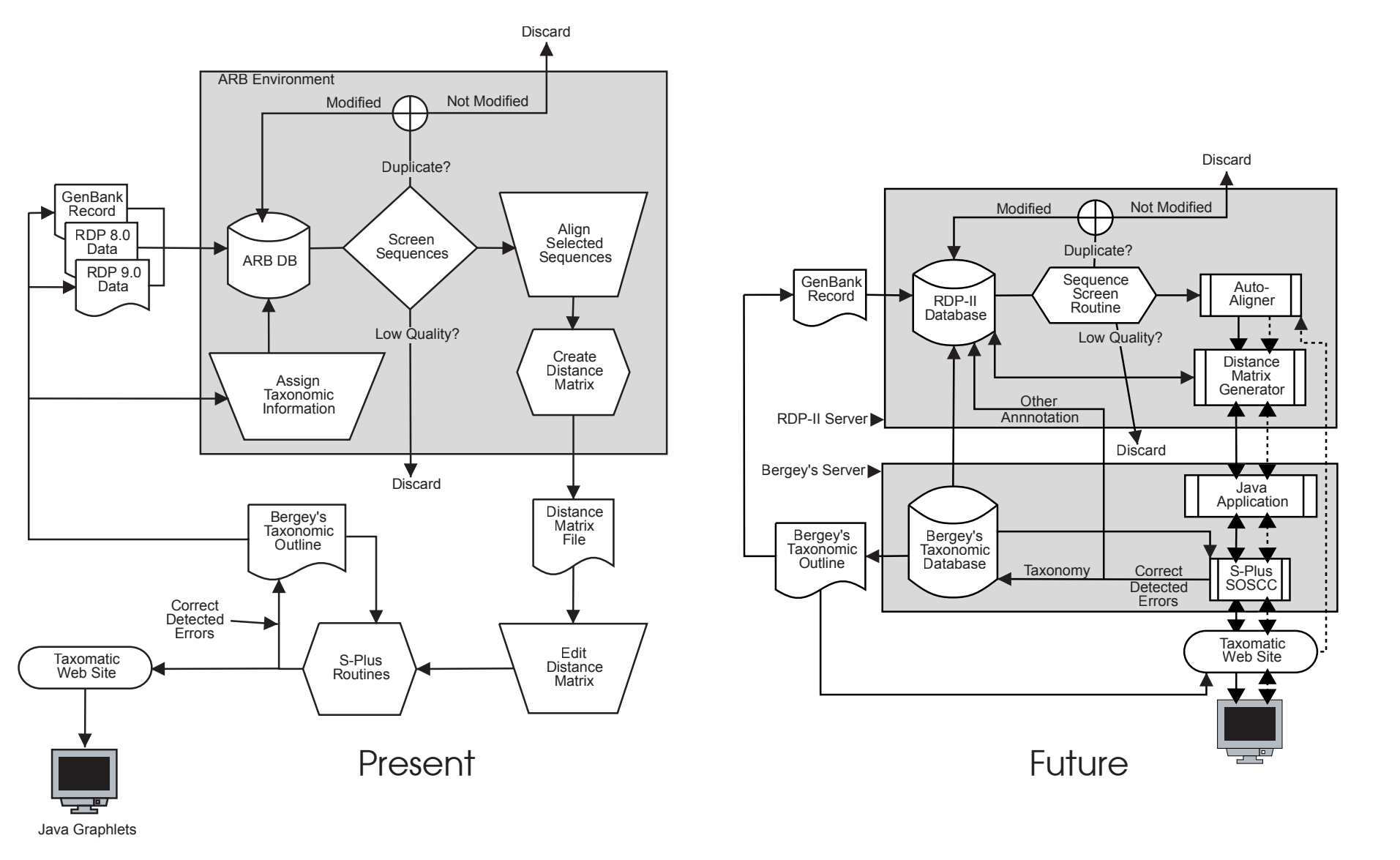




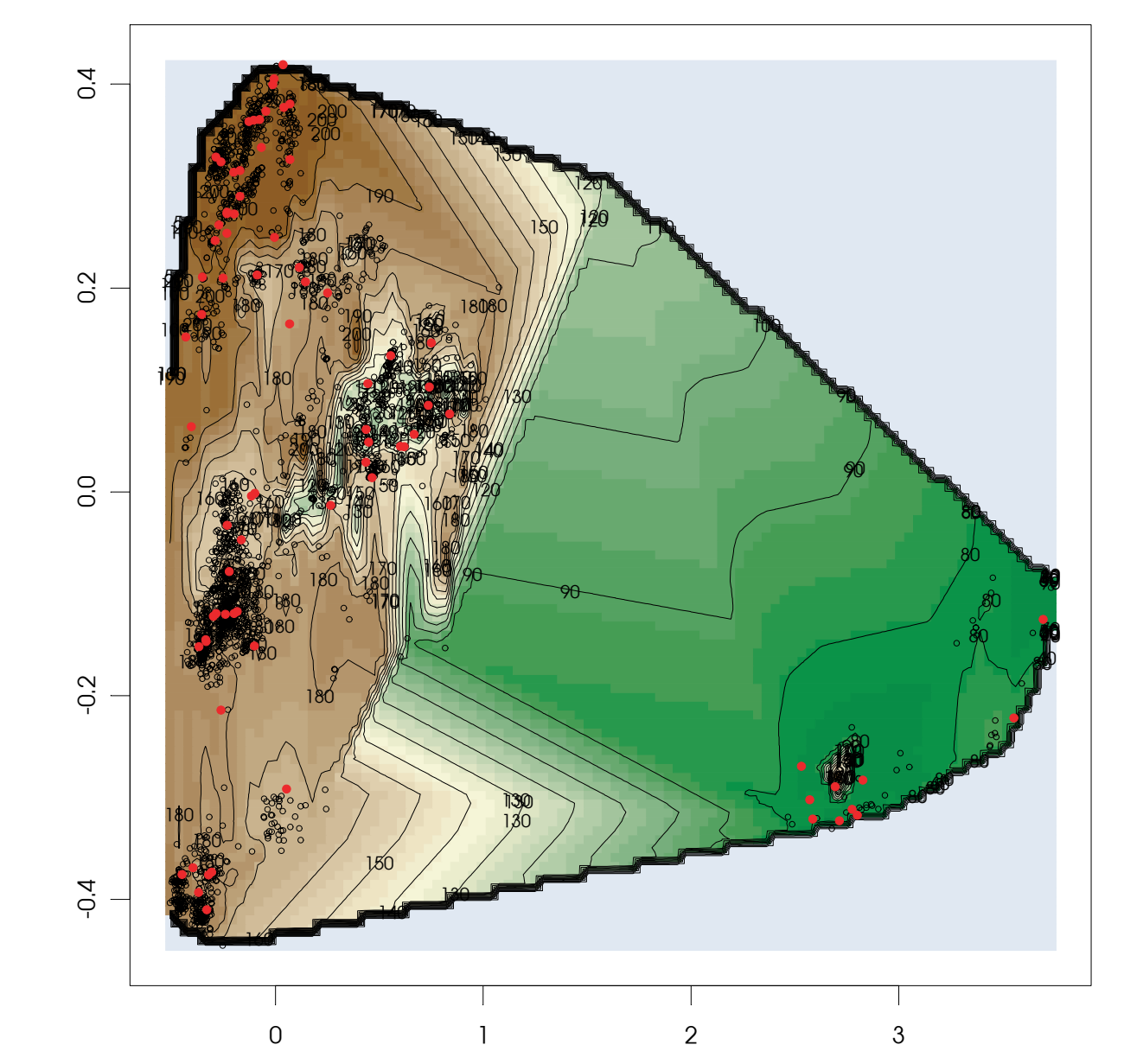




Work/Data Flow Supporting the Taxonomic Browser



Distribution of potential orthologs in prokaryotes



FUTURE DIRECTIONS

At this time much of the data manipulation and analysis involve manual steps. Some of the most time-consuming steps are enclosed by the trapezoids in the "Present" flow chart in the Figure shown above left. Probably the slowest step involves maintaining the alignment of SSU rRNA sequences. ~~The~~ Bergey's Manual Trust, the ATCC and the Ribosomal Database Project are collaborating on a plan to largely automate the maintenance of the data and analyses that underpin the Taxonomy Browser. An overview of the pipeline is shown in the "Future" flow chart in the Figure above left. The RDP-II's autoaligner will generate the alignments and distance matrices will be calculated from these alignmentss. The latter will be passed to a set of S-Plus functions that will generate either optimized or current taxonomies. Corrections in annotation and so on will be fed back to the RDP-II sequence database.

Another planned enhancement of the Taxonomy browser is the addition of facilities for handling user-submitted sequences. This would allow members of the research community ot classify and/or identify taxa based on their own sequence data. We also anticipate that this tool can be used to compare phylogenies constructed with data from other, non-SSU rRNAsequences or with any other data that allows a distance metric to be measured.

The Taxonomic Atlas can also be used to leverage annotation from the many prokaryotic genome sequencing projects. The Figure on the right above shows information about the distribution of 205 single copy homologs (drawn from Lerat et al. 2002) among 136 prokaryotic genomes, overlaid on a PCA plot. The ~~set~~ was selected from the *Gammaproteobacteria.* The red dots show the location of the sequenced organisms in the taxonomic space.The colors and contours represent the inferred number of homologs present in genomes in that area of the space.