Aug 15, 2007

K9

A T-RFLP and Community Data Analysis Toolkit

Welcome to K9! K9 is a modular application whose primary application is for working with analyzing T-RFLP data. That said I've recently greatly expanded it's ability to work with microbial community data as it is analyzed at the Microbiome Core at the University of Michigan. As more functions are needed I anticipate adding them but that is a project for the future.

First and foremost K9 will take the peak file that is generated by either GeneScan (a commercial program from Applied Biosystems Inc. (ABI) that is no longer supported but people still use) or Peak Scanner (a newer and free program also by ABI), bin the data and perform a Bray-Curtis clustering analysis of said data to generate a Dendrogram. Additionally, K9 leverages a number of functions from R for carrying out or display various elements of microbial community analysis. Sounds pretty straight-forward right? Well, if I've done my job right it certainly should seem that way

What you will need:

A Macintosh computer running OS 10.4.x or greater (at the time of writing the highest version is 10.4.9), and the latest version of R installed (http://www.r-project.org/ (at the time of writing the latest OS X version is 2.5.1)). For functions other than the T-RFLP data analysis you will also need to install the following R packages (selecting install dependencies): vegan, rgl, and gplots. Finally, to use DOTUR you will need to have the DOTUR executable in you /usr/bin directory. You will, of course, also need the peak files generated by GeneScan or Peak Scanner in the proper format (see setting up GeneScan or Peak Scanner T-RFLP peak file analysis)

A little history:

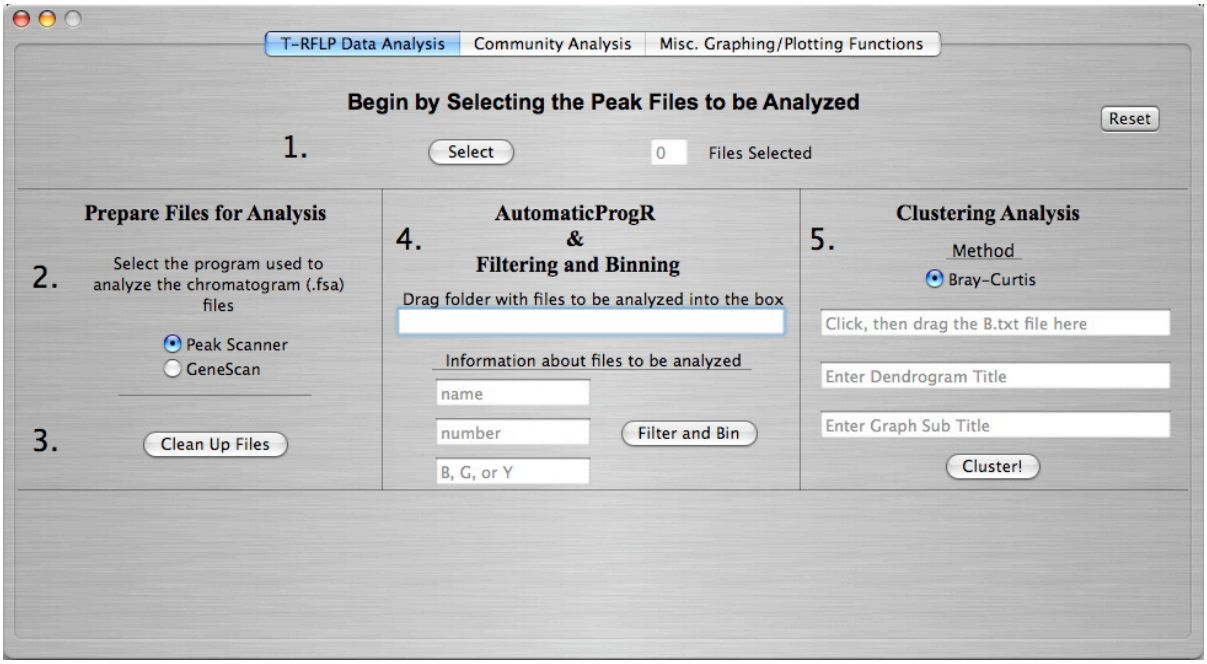
When our lab first started doing this kind of analysis several things became quickly clear: 1) Not all ABI sequencers work with GeneScan (Peak Scanner has solved this problem), 2) the output files from GeneScan (or Peak Scanner) re-

quire specific formatted before they can be processed by this nifty little Filtering and Binning R program, 3) the Perl script that the gets this format ready (AutomaticProgR.pl) was not complete and required a good amount of editing of the output files in a text editor before they could be made useful, 4) before one could format the files with the Perl script, one needed to clean out all sorts of characters that cause Perl to get confused and 5) most people don't feel comfortable working in a command-line environment deciphering cryptic error messages. All-in-all, this process needed to be made more user friendly. So I wrote K9.

How to use K9:

When you launch K9 you will notice the window pictured below. The different aspects of what this program displayed when the requisite tab is clicked. The explanation that follows of how to use K9 will also be subdivided based on the tab that is chosen.

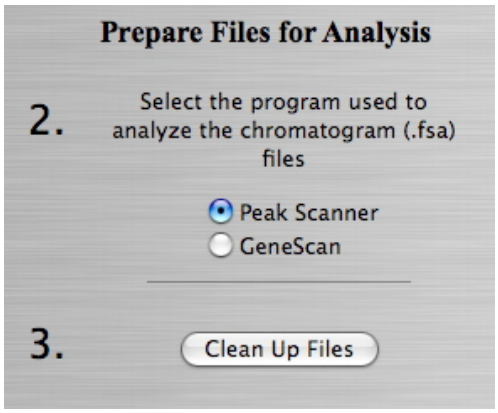
T-RFLP Data Analysis:



The steps going from peak file to a Bray-Curtis distance matrix are nicely numbered from 1 to 5.

To begin:

1) Press the "**Select**" button in the top middle of the application window and select the files that you want to analyze (Note: you can select multiple files).



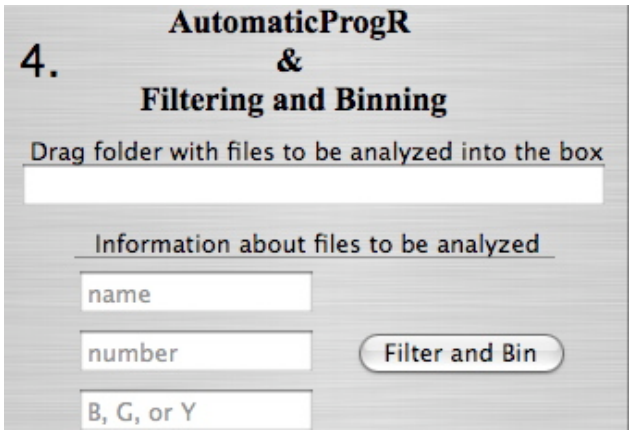
The "**Files Selected**" indicator will tell you how many files you have selected.

Note: The way that K9 cleans up the files is "destructive" in that it will change the original peak files. It is recommended that you carry out the analysis on copies of these files.

2) Select the radio button that corresponds to the program that was used to generate the peak files.

Next, click on the button "**Clean Up Files**". A progress indicator will become active and stop when the cleaning is complete.

Now your files are ready to be formatted and the peak set in the appropriate bins

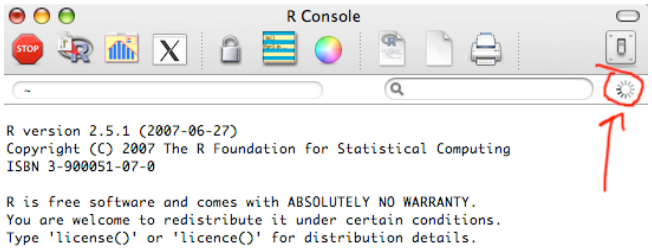


3) Drag the folder that contains the files that you just cleaned up to the center text field box (alternatively you can type in the path name (unix / de-limited not apple : delimited (e.g. /Users/yourusername/Desktop/T-RFLP_Data)).

Note: **If you have a space in the path name K9 will not work!** That means that "/Users/yourusername/Desktop/T-RFLP_Data" is ok but "/Users/ yourusername/Desktop/T-RFLP Data" is not. Keep track of this as it will save you headaches down the road.

Next enter the information for the files you want processed. Put the name of the files in the "**name**" box, how many there are in the "**number**" box and the color of fluorophore that you want analyzed (usually B) in "**B, G, or Y**" box. A quick note on files names: No spaces in the name and you need to enter it up to where the numbers start. So for example if you had 8 files called M1, M2, M3....M8, you would need to enter M into the name box. On the other hand you had HA1, HA2, HA3...HA8 you would enter HA into the name box. Again, the number field is the number of files. You must make sure that you match the letter that you enter into the "**B, G, or Y**" field with the letter that is in your peak file (the peak file that you output from GeneScan or Peak Scanner must have only one peak color per file (so all blue peaks, all green, etc. but not blue and green in the same file).

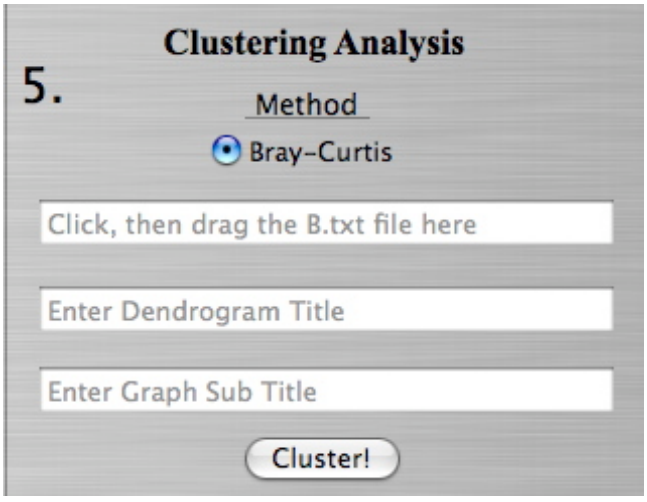
Click on the "**Filter and Bin**" button. A progress indicator will start and the R program will launch. This step may take awhile (might be instant, but also might take ~15 min.) To know R is still thinking check the spinning progress indicator in the R window (shown to the left). You should be returned to the application when the R-program concludes (if not just click on the K9 window).



If you look at the box "**Click, then drag the B.txt file here**" it should now have a path name in it (beginning with /Users/...).

4) Lastly, K9 lets you do a furtherest neighbor cluster analysis of the Bray-Curtis distance of your data.

Currently, the Bray-Curtis analysis is all that is offered; however, should others be made available in the future you would select them by clicking on the



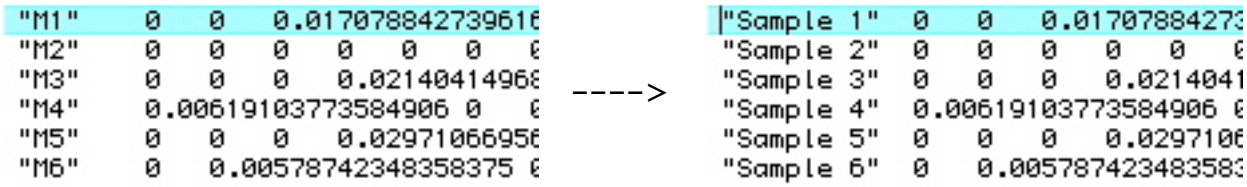
appropriate radio button.

Next fill in the title that you want for your dendrogram (this goes at the top) and the sub-title (this goes at the bottom).

Click the button "**Cluster**!" and again R will spring to life and produce a dendrogram of the data. If you want to save that graph, click on the graph, go to the R menu, select Save As and save it where you want to save it. This file will be a PDF file that is readily converted to other formats using the application Preview (Export the file as format you want).

Now you may be wondering why you would want to analyze the same B.txt file more than once. I can think of a several reasons but the one that stands out this most is if you like the graph but you want better labels for the groups.

To do this open the B.txt file in a text editor (I use TextWrangler (great program free from BareBones Software)) and change the title in the first column to what you want it to read on the dendrogram (see below)



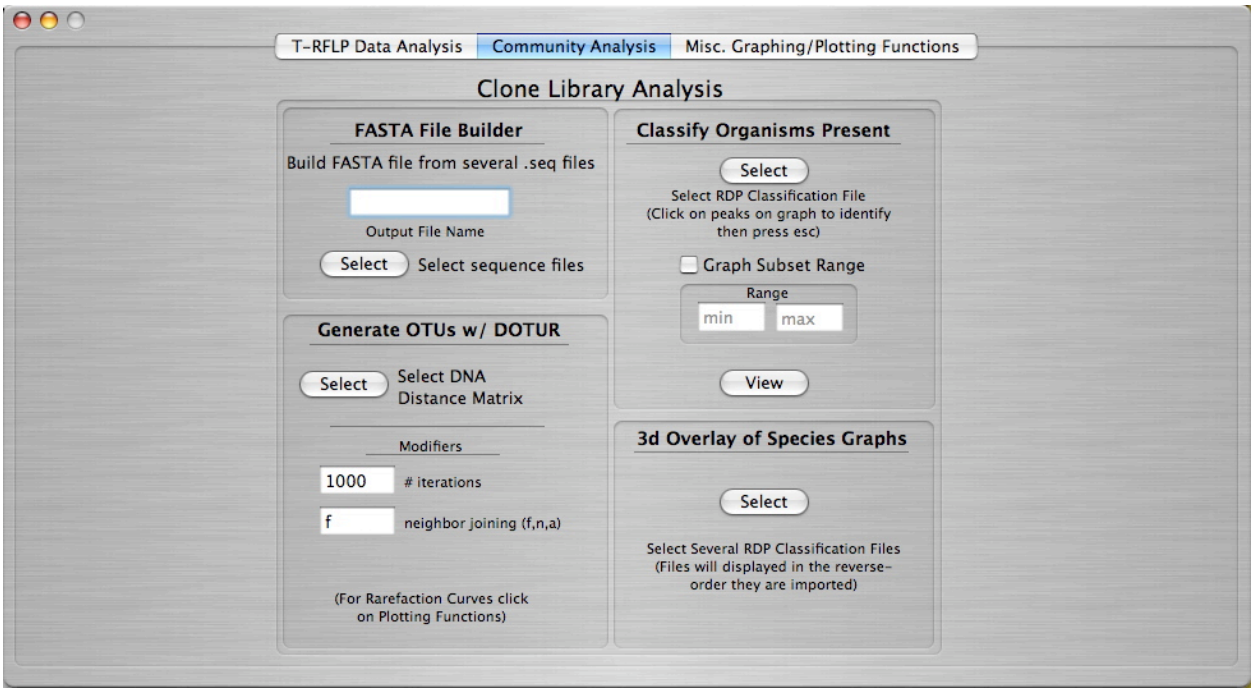
And you are done!

Now I mentioned earlier that this program is modular, this is because if you have the right file you can jump in at different parts of the analysis (e.g. your files are clean, start at filtering and binning; you just want to redraw a dendrogram, start at Clustering Analysis.

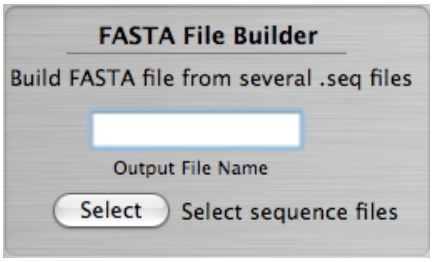
The very last thing to know is the "Reset" button up in the upper right hand corner of the application window. If you click this button the program will be restored to it's defaults, temp files will be cleaned up and R will be quit. Basically it's a general house keeping function.

Community Analysis:

When the Community Analysis tab is clicked on the window should appear as shown below.



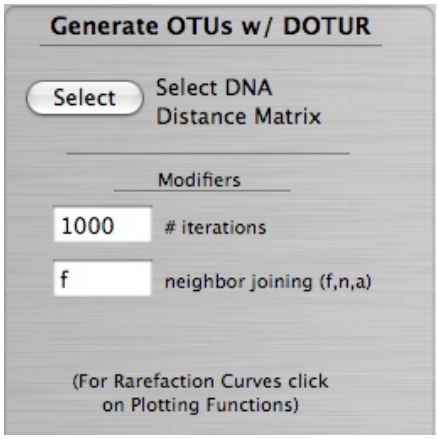
Contained herein are 4 functions that we have been using to facilitate analysis of clone libraries. The first of these is the FASTA File Builder.



Typically, when one gets sequence data back from a core it is simply the sequence data without any other identifiers. What FASTA File Builder does is to take multiple sequence files and build them into one FASTA formated file. To use this, type in the title for the file you want to output and select all of the sequence files that you want to format. The .fasta file will be placed in the same directory as the sequence files that you select. Not too complicated, but it saves a lot of time cutting and pasting.

Next there is DOTUR. This is a graphical front-end for the DOTUR program created by P.D. Schloss at U. Mass Amherst (detailed in the paper Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness, Applied and Environmental Microbiology, March 2005, p. 1501-1506, Vol. 71, No. 3). For this part of the program to work the user (or whomever set this program up) needs to have done a couple things in advance.

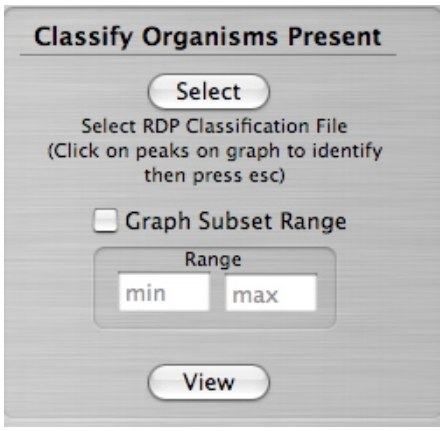
First they need to have the DOTUR executable installed into the /usr/bin directory (to access it open Terminal in the Utilities folder and type open /usr/bin then drag the DOTUR executable into that window (authenticate as needed).



To run DOTUR, you need a DNA distance file (DNAdist or PHYLIP formatted distance file). These can be generated using ClustalX (an OS X implementation of ClustalW), or what we have been doing, a distance file obtained from the Ribosomal Database Project (http://rdp.cme.msu.edu/). Once you have your distance file set the number of iterations that you want and the desired method of neighbor joining (furthest, nearest, or average) then click on Select to select the distance file and begin analyzing. By default DOTUR does a rarefaction analysis on everything. After clicking select the Terminal application will be launched and DOTUR will begin running. When it is finished (and it can take awhile) the Terminal window will close and return you to K9.

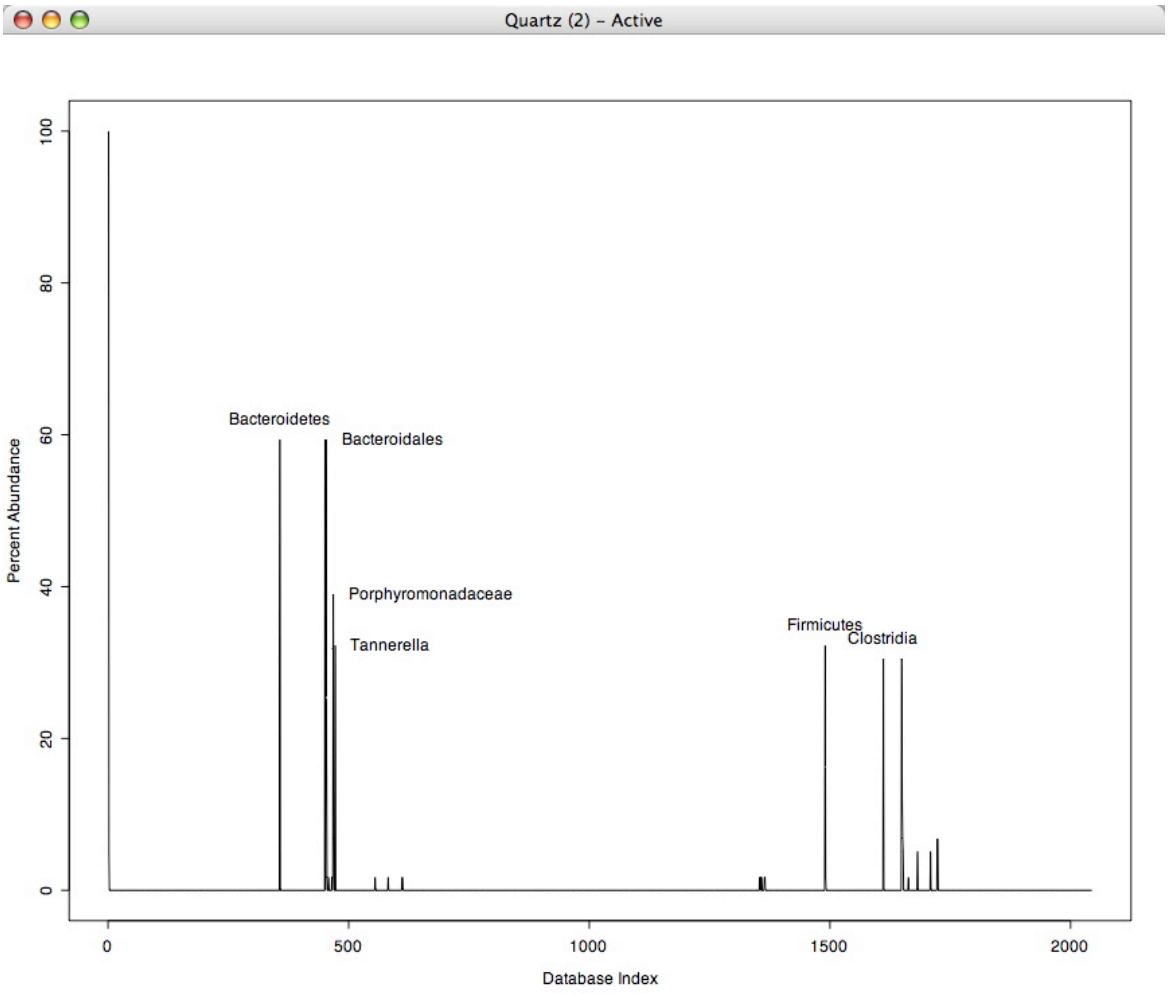
The next function allows the classification of the organisms present in a clone library. To do this you need the RDP Classification File (again downloaded from http://rdp.cme.msu.edu/).

To get this file log in to RDP, click on the Group Name of the data set (used to generate the distance file), then click on the button labeled Select Aligned Only, then View Classification and finally the button labeled download as text file.



Now that you have the file, click on the Select button as shown below and select the Classification File. This selects the file for further analysis, but doesn't do anything else just yet. To see the species present click on the View button. Here the % abundance of all the species present are graphed vs. the RDP database position. To identify which species are present click on the tops of the peaks and when you are finished press esc. If things worked correctly you should have the labels pop up on the graph (see below). Sometimes there are so many peaks in a region that it makes separating the labels difficult. To visualize

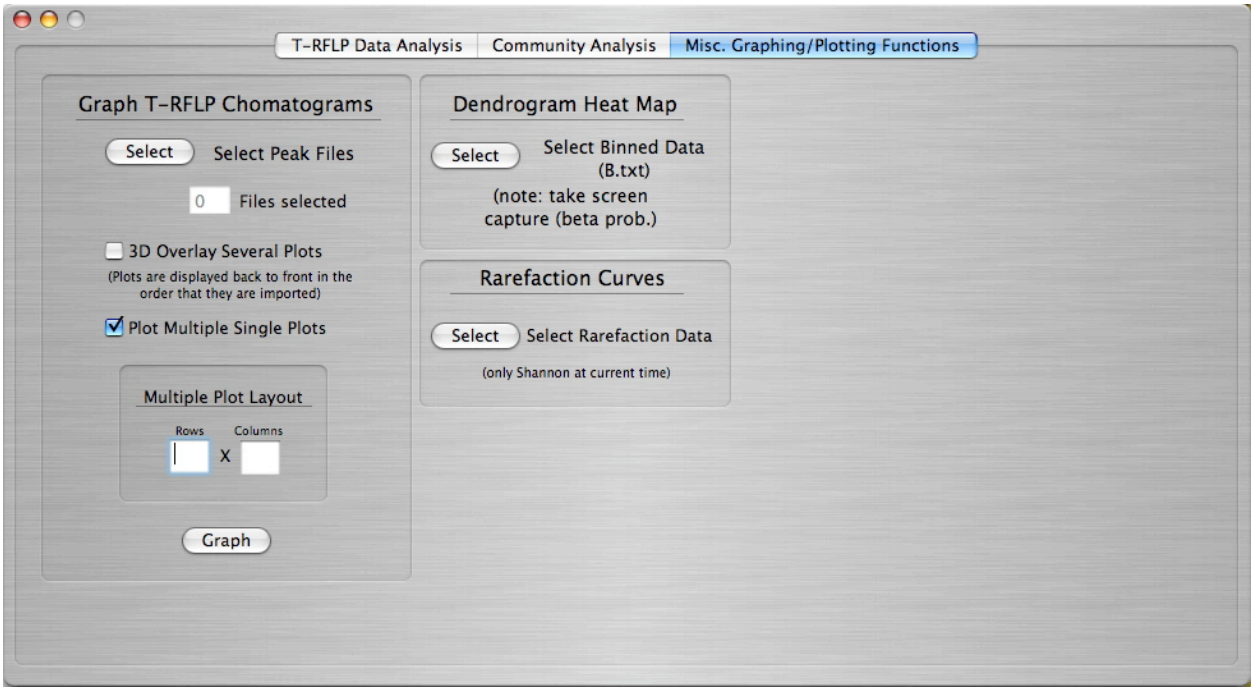
this click the Graph Subset Range button then type in the subset range you want to analyze. Clicking to identify the peaks will still be the same.



The next function is a 3D-overlay of several species abundance graphs if you would like to visualize them in this way. To use, click Select and select several RDP Classification FIles that you would like to visualize. You can't identify peaks from this view, but I suppose it does have it's uses.

Misc. Graphing/Plotting Functions:

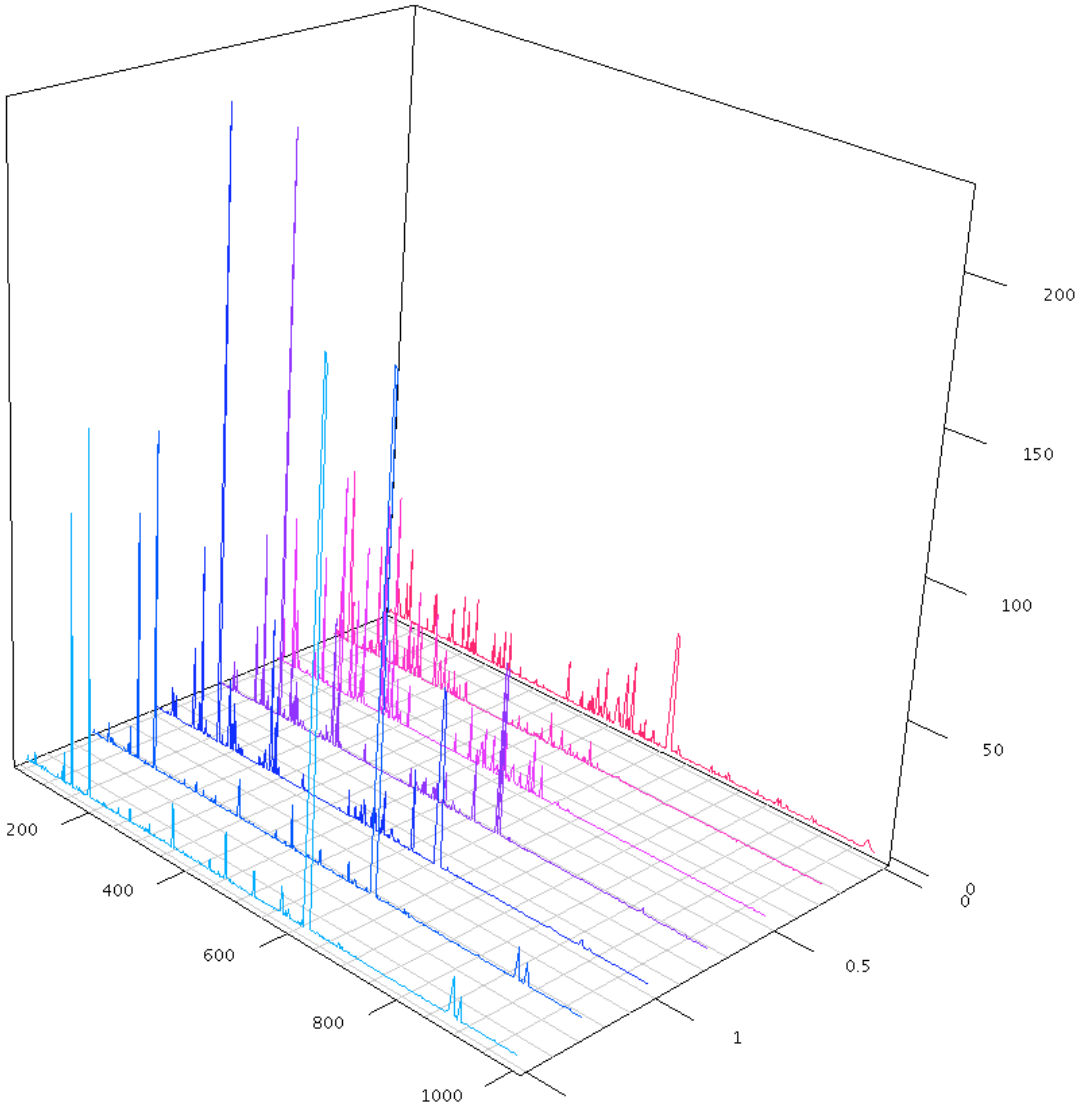
Upon selecting the Misc. Graphing/Plotting Functions tab you will see the window that is shown below. These functions will facilitate visualization of the data; however, it should be noted that they will not produce "final" figures (proper labels, axis descriptions, etc. still need to be added in postwork).



The first set of functions pertain to graphing T-RFLP chromatograms. This is largely because I've found that peak scanner can't do nearly as good a job of this as one would like.

First select the peak files that you would like to graph (you can select multiple files). The Files Selected counter will tell you how many files you have selected. There are currently two ways that you can view this data. The default is to Plot Multiple Single Plots. To view, set the layout of the graphs (I'd advise against putting too many on one page) so that as many files as you have selected can fit and click view. If you select a layout that can't fit the total number of graphs that you have the figure will start over with the remainder. The plots are displayed in the order that they were listed when you selected them (there is a number above each graph denoting the order but not the file name). To save this figure click on the graph then go to the file menu and select Save As and give it a name.

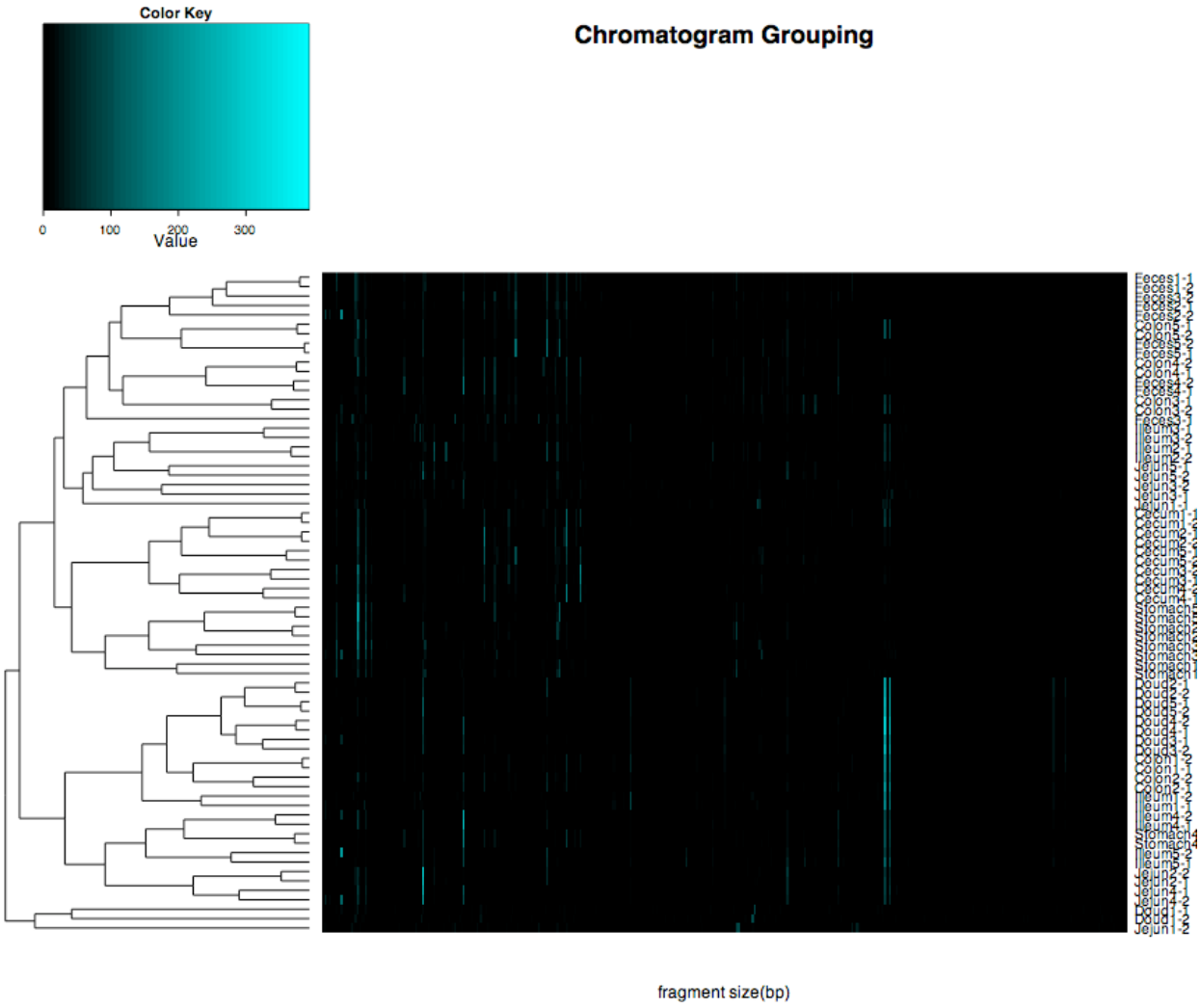
The other option is to select the 3D-Overlay button and then click view. When you do this you will be treated a 3D-rotatable figure containing the chromatograms (the weighted average of the signal) for comparison. Make the little window bigger and move the figure around until you find a view that you like, then take a screen capture of that view (apple+shift+4 then click and drag the cross-hairs over the image) as shown below.



The chromatograms are loaded onto the graph from back to front (so in the figure above the red is the first, then magenta, etc.).

The next function is the Dendrogram Heat Map. A heat map displays the "intensity" of data as bright color. For the purposes of this program this function is a nice way to visualize all of the T-RFLP data at once. To use it click on the select button and select a B.txt file. Previously, we had used this file to gener-

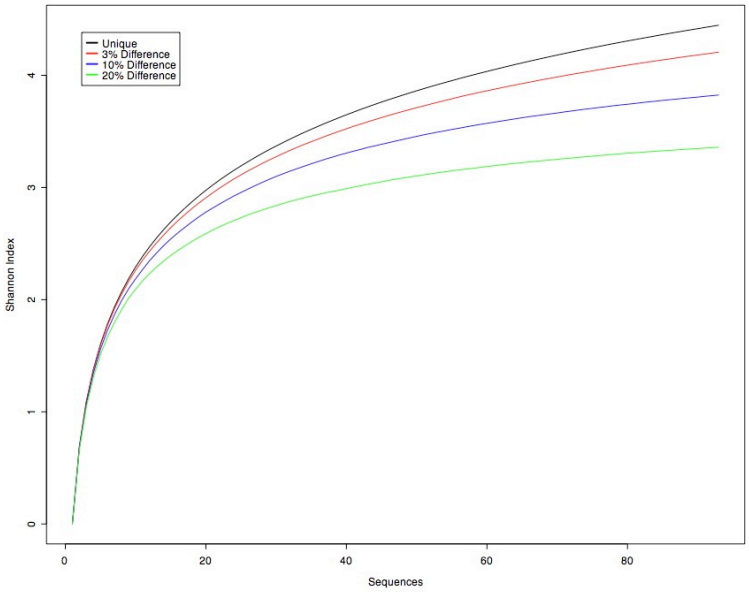
ate a dendrogram. The B.txt file contains the binned peaks from the T-RFLP analysis and as such contains the root mean squared averages of all of the peaks in each peak file. Thus, with a simple transformation (multiply by 1000 so that the values don't look silly), all of the "peaks" from all the sample together can be visualized as intensity in a heat plot as shown below.



On the left again is a dendrogram showing how the sample can be grouped. The body of the plot shows the background values as black peaks in blue. The degree of blue is based on the how large the value of the peak is. At the bottom of the plot the label "fragment size (bp)" means that as you move from left to right on the plot the fragment length increases. Finally on the right are the names of the samples. Of all the visualization functions I consider this one to be the most underdeveloped, but it does have it's utility. To save this figure you need to do a screen capture (apple+shift+4 and drag the cross-hairs over the figure). Unlike the 3D functions, you can save this like other figures by clicking on the graph and selecting Save As; however, due to a problem with the

pdf output you will see hundreds of thin white lines which separate the various color “boxes” in the figure (looks terrible, I really don’t recommend getting the figure in this way).

The final visualization tool is for the visualization of Rarefaction curves (in particular the Shannon-Weaver rarefaction curve created by DOTUR). The Shannon-Weaver index is really a measure of entropy or the total “information” in the system; however, it is also a useful tool for studying diversity. To use click on the Select button in the Rarefaction Curves box and select the DOTUR output file that contains .fn.r.shannon and you will be presented with a graph as seen below.



A high Shannon Index value indicates an increase in the amount of information in the system (thus an increase in the amount of diversity). The first curve is the amount of information when each of sequences are treated as unique, the second is where they are grouped into those that differ by 3%, the third curve by 10% and the fourth by 20%. One can imagine that by the time one is grouping sequences by those that differ by up to 20%, one should have many fewer groups and a much more “organized” system. Organized = decreased entropy = lower Shannon Index value. In the graph shown here there is an unusually high amount of diversity (which highlights the problem of just using ClustalW for the generation of the distance matrix* when comparing 16S-RNA sequences - RDP also factors in secondary structure which creates a much more reasonable alignment). Other .r files may be graphed using this function; however, but the axis labels will have to be changed in Photoshop.

*although I’m sure it would work better with a large database of aligned 16S sequences.

Acknowledgments:

While I have put together this program, I can't claim to have created most of the content. It is best to think of this program as a graphical interface for a number of powerful, but "ugly" (i.e. command-line) utilities. So to give proper credit, the T-RFLP R Filtering and Binning functions and the perl script that prepares the data for these functions was originally created by Ziad Abdo at the University of Idaho (described in the paper ***Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes*** (http://www.ibest.uidaho.edu/tools/trflp_stats/documentation/T-RFLP-Published.pdf)). I have made some modifications to these but if the original functions are desired they can be downloaded at http://www.ibest.uidaho.edu/tools/trflp_stats/index.php. The Vegan package for R was created by Jari Oksanen, Roeland Kindt, Pierre Legendre, Bob O'Hara (http://cc.oulu.fi/~jarioksa/). Gplots was created by Gregory R. Warnes (http://cran.r-project.org/src/contrib/Descriptions/gplots.html). Rgl was created by Daniel Adler, Duncan Murdoch (http://rgl.neoscientists.org). Finally I'd like to that those that created R. It is an amazing piece of software and that it is free makes it all the more incredible. I've left the R output from the various function available when one is running K9 so that if one would like to try to tweak outputs or indeed learn how R works (at least within this context), they can (something that I highly recommend).

Side Notes:

I actually don't know what this will do with colors other than blue (B) because I have not had any data to analyze and test it out. I've done my best to ensure that there should be no color specific conflicts, but there is a limit to which I can test.

Troubleshooting:

As this is the first release of this program I don't know of major problems yet. That said, I know they exist. Here are some basic things which can cause problems, but have an easy fix.

Problem	Solution
Filtering and binning step does not produce an "output" file.	1) Make sure there are no spaces in the file path. 2) Look to see if the files are still being processed (this can take a while). 3) Make sure that the peak files are in the format 6 column tab-delimited format needed for analysis (see section on Peak Scan and GeneScan file format) 4) the input files might still have the non-sense characters that GeneScan and Peak Scanner leave. Try running the previous step again. 5) Make sure that you selected the right program for cleaning the files.
Filtering and binning does not produce a B.txt file	1) Make sure there are no spaces in the file path. 2) Make sure that there is only one class of color information in the input files (there will be a letter in the first column of each line and that should only be the color you want to analyze).
I just got some really goofy output that doesn't make sense.	1) Click on the Reset button to quit R and reset K9, then try again. 2) if it still looks goofy, it may just be your data.

Problem	Solution
DOTUR is not working	1) remember that in order for this to work the DOTUR executable needs to be in the /usr/bin directory.
Nothing is working	1) Odds-on-are that the K9 application resides in a path that has a space in it. I recommend putting it on your desktop or Applications folder or in a folder where words are only separated by a "_". 2) make sure you have all the necessary R-packages installed.
operation xxx just isn't working	Check the log on the R screen to see if there are error messages there. They are usually pretty clear.