# Aligning & analyzing TRFLP data

Based on Rees et al., 2004

Rees, G. N., Baldwin, D. S., Watson, G. O., Perryman, S., and D. L. Nielson, 2004. Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics. *Antonie van Leeuwenhoek*, 86:339-347.

1) **Standardize** each peak height as a % of the total peak height for that sample. Reassign all peaks that are less than 1% to 0, and recalculate the % of the remaining peaks for that sample.

   **a)** Import data tables that you exported from Genescan into Excel. Create a workbook for each project and a worksheet page for each sample.

   **b)** For each sample, calculate the total of all the peak heights (usually column D).

   **c)** Next calculate each individual peak height in that sample a percentage of the total peak heights for that sample (new column,"%peak height").
   - (peak1_height/total_peak_heights)*100,
   - (peak2_height/total_peak_heights)*100, etc.

   **d)** Use a simple If-Then statement to carry over the peak heights and areas that are $\geq$1%. Use one column for peak heights ("new peak height"), and another for peak areas ("new peak area") (even though we will only use peak height data). Type your actual cell numbers into the formulas below in their respective columns.
   - =IF(%peakheight<1,0,peakheight)
   - =IF(%peakheight<1,0,peakarea)

   **e)** If you had any peaks that were below the 1% threshold in the previous step, they would have resulted in zeros, which means we need to re-adjust our total peak height. So sum up your "new peak height" column and also your "new peak area" column.

   **f)** Next calculate your "adjusted % peak height", and repeat for your "adjusted %peak area".
   - (peak1_newheight/total_peak_newheight)*100
   - (peak1_newarea/total_peak_newarea)*100

2) **Align** data using a crosstab macro ("treeflap"). This macro rounds each peak size to the nearest integer (or value that you select) and reorganizes the data in crosstab form so it is ready to import into analysis software. You can select to use peak heights or peak areas. It was written by Dr. C. Walsh and is available from: http://www.wsc.monash.edu.au/~cwalsh/treeflap.xls

**a)** Create a new workbook and set up the following column titles: ID, size (bp), adjusted %peak height, adjusted %peak area.

**b)** Copy & paste in the data from each sample that you wish to analyze together in the appropriate rows.

**c)** Insert a new column to the very left of your data. Start with the number 1 at your first sample row and drag all the way down your sheet until you are 1 cell just below your last sample row. If it doesn't do it automatically, select "fill series" so that you have a sequential number series starting at 1 at the very top.

**d)** Open the treeflap.xls file and if prompted, click to enable macros.

**e)** Press Ctrl–Shift-T at the same time, and select Peak Heights and round to the nearest 1 for now. Your data should automatically align itself into a crosstab matrix. One sheet will have your aligned peak height %'s, another will have the number of peaks aligned in each cell (this shows you if you have multiple peaks in the same size window).

**f)** Select the worksheet that has your aligned peak heights and copy and paste special (values only) to a new sheet in the same workbook.

**g)** Insert 3 rows at the top of the sheet. At the bottom of the first column with numbers calculate the sum of all numbers in that size. Then copy the formula to the bottom of all columns with the data.

- =SUM(first_number_at_top:last_number_at_bottom)

**h)** Now go through and delete any columns that have a sum of zero. We need to remove these peaks so that they won't interfere with our analysis.

**i)** In the very first cell at the top of your sheet, enter the total # of samples (=rows). Just below it enter the total # of peaks (=columns of peak sizes).

3) **Analyze** using non-metric MDS in Primer v6. The standardized & aligned peak heights are either transformed into presence/absence, or left unaltered with the heights intact. A Bray-Curtis similarity matrix is then performed, followed by the MDS analysis using 100 random restarts.

   **a)** Open Primer v6 software & click on the new data icon (or go to File, New). Select "Sample data" and press Enter.

   **b)** Enter your title, select "Abundance", and samples as "Rows". Enter your number of peaks in the "Number of columns" section and the number of samples in the "Number of rows" section. Press Enter.

   **c)** Right click on the table and select "Labels" then "Samples". Copy and paste your column of row labels into this section.

   **d)** Back in your aligned spreadsheet copy the row with all of your peak sizes and then right click on a cell below your table. Choose "Paste special" then "Transpose". This should have created a single column listing all of your sizes.

   **e)** Back in Primer right click on the table again and select "Labels" then "Variables". Copy and paste your transposed list of sizes into this section.

   **f)** Go to the top menu and select "Analyze", then "Pre-treatment", then "Transform (overall)". Select "square root" then click Ok.

   **g)** From the top menu select "Analyze", then "Resemblance". The window that pops up should already have "Samples" selected and "Bray-Curtis similarity". Click Ok.

   **h)** Select "Analyze", then "MDS". Set the number of restarts to 100, and keep the minimum stress at 0.01. Click Ok.

   **i)** Your MDS plot will now appear. If you would like to see the actual sample points, right click on the plot and select "Data labels & symbols". Under the Symbols heading, click on the small box next to "Plot". You can also change the color, size and symbol here. If you want to change the font size or type you can do so under the Labels heading.

   **j)** You can rotate the plot any way you wish. Play around with including and removing samples to see how the relationships change with those remaining. Check your final stress value listed on the plot and compare with the values listed on page 5-6 in the supplementary book that we ordered along with the software.

DIVERSE
Overall transform
    Square Root
    Presence / Absence
SIMPER
Resemblance (= Similarity Matrix)
CLUSTER
    SIMPROF test + create factors

quick, mainly-automated method that would allow us to easily add new samples without having to spend a lot of time re-normalizing or re-aligning everything each time.

**Other things to consider:**

- Rees is now using a variable percentage threshold method to normalize data –

  Osborne, C.A., Rees, G. N., Bernstein, Y., and P. H. Janssen, 2006. New threshold and confidence estimates for terminal restriction fragment length polymorphism analysis of complex bacterial communities. *Applied and Environmental Microbiology*, 72(2):1270-1278.

- Hewson & Fuhrman use a shifting window bin type method –

  Hewson, I., and J. A. Fuhrman, 2006. Improved strategy for comparing microbial assemblage fingerprints. *Microbial Ecology*, 51:147-153.

- Craig Nelson has created the "Genescan Aligner Workbook" for Excel. It is another semi-automated process except that this one allows more human interface. You set manual cutoff points, but supposedly it will provide some visual aids and calculations to help you with your peak lumping. It might be easier than trying to align everything by yourself if you are aligning by hand.

  Noah sent me the Excel workbook and instructions, so let me know if would like to try it out and I will send it to you.