User Guide to Statistical Analyses

Most of the statistical analyses are conducted using SAS v.8.02 (SAS Institute, Inc. Cary NC 27513, http://www.sas.com/) programs. Analyses include:

(1) One way **ANOVAs** on the total number of clones found per treatment, or on fingerprint richness, diversity, or evenness in each treatment.
(2) **Stepwise Discriminant analysis**, to identify fingerprint clusters in which the distribution of clones per treatment varies.
(3) **Stepwise Regression/ Correlation analysis,** to identify fingerprint clusters in which the distribution of clones per treatment varies with some measurement that varies between treatments (for example, the ability of soils in different treatments to suppress a plant parasitic nematode).

(1) One-way **ANOVAs**: we use this test to see if there are differences between treatments. In order to compare treatments, we need multiple observations for each treatment. When you design your experiment, try to accommodate treatment replications, or you may not be able to compare parameter values for each treatment.

- Use SAS to perform this analysis. Below is a sample SAS program that you can rewrite to accommodate your needs.

```
options ls = 80 ps = 55 nocenter nodate;
/* Set up temporary SAS data set called onewaybacteria */
data onewaybacteria;
/* Two variables to be input, Treatment and bacteria count */
  input treatment $ baccount;
/* baccount is the total number of different bacterial clones found for each
individual replicate */
/* The "$" indicates that Treatment is a text variable*/
  title1 'Oneway ANOVA Example';
  title2 'Bacteria Count';
/* datalines Statement to indicate data is about to begin */
datalines;
101   644
101   502
101   495
zhou  610
zhou  525
zhou  560
mbbb  539
mbbb  513
mbbb  577
mbv   549
mbv   587
mbv   474
;
RUN;
proc glm;
/* This analysis has balanced data, but proc glm was used in case
there are unequal replicates on future data*/
      class treatment;
```

```
/*The MEANS statement with the TUKEY option generates the tukey pairwise
      comparisons for the cells. This test protects against inflation of
      the type I error rate due to multiple t-tests and uses a constant
      error term for the analysis.*/
      model baccount=treatment;
      means treatment  / tukey cldiff ;
run;
```

This one-way ANOVA example will test for differences in total number of bacteria clones identified between the four treatments. The Tukey means separation test is then applied to identify pairwise differences between the individual treatments. The output from this analysis is presented below.

```
Oneway ANOVA Example
Bacteria Count
The GLM Procedure
Dependent Variable: baccount

                              Sum of
Source                DF        Squares    Mean Square   F Value   Pr > F
Model                  3     1330.25000      443.41667      0.13   0.9370
Error                  8    26472.66667     3309.08333
Corrected Total       11    27802.91667

R-Square     Coeff Var      Root MSE    baccount Mean
0.047846      10.49879      57.52463         547.9167

Source                DF      Type I SS    Mean Square   F Value   Pr > F
treatment              3    1330.250000     443.416667      0.13   0.9370

Source                DF    Type III SS    Mean Square   F Value   Pr > F
treatment              3    1330.250000     443.416667      0.13   0.9370


The GLM Procedure
Tukey's Studentized Range (HSD) Test for baccount
NOTE: This test controls the Type I experimentwise error rate.

Alpha                                     0.05
Error Degrees of Freedom                     8
Error Mean Square                     3309.083
Critical Value of Studentized Range   4.52880
Minimum Significant Difference         150.41

Comparisons significant at the 0.05 level are indicated by ***.

                Difference       Simultaneous
 treatment         Between     95% Confidence
Comparison           Means          Limits

zhou - 101           18.00    -132.41  168.41
zhou - mbbb          22.00    -128.41  172.41
zhou - mbv           28.33    -122.08  178.74
101  - zhou         -18.00    -168.41  132.41
101  - mbbb           4.00    -146.41  154.41
```

```
101  - mbv          10.33    -140.08  160.74
mbbb - zhou        -22.00    -172.41  128.41
mbbb - 101          -4.00    -154.41  146.41
mbbb - mbv           6.33    -144.08  156.74
mbv  - zhou        -28.33    -178.74  122.08
mbv  - 101         -10.33    -160.74  140.08
mbv  - mbbb         -6.33    -156.74  144.08
```

In this example no significant differences were found between total number of bacteria clones between the four treatments (p=0.13) on the ANOVA and no significant differences were found on the pairwise comparisons. Typically the pairwise comparisons would not be used if the overall F-test was not significant.

(2) Stepwise Discriminant analysis: Given the large number of fingerprint groups in OFRG studies, it would be unfeasible to manually pick out groups, or clusters of groups, that demonstrate treatment differences. To help us locate differences between treatments, we use a stepwise discriminant analysis. We need to look at data from groups containing a sufficient number of clones for analysis. What the optimal number of clones would be for this analysis is unknown, but currently we analyze data from fingerprint groups containing n + 1 clones, where n = the number of treatments, or in the case of a large number of treatments, an arbitrarily chosen number such as 5 clones. Note that the greater the number of clones in a cluster, the easier it is to identify treatment differences, should they exist.

- First step: Open the file you made for Cluster analysis, above, or if you have not performed this analysis, obtain the data from groups containing 5 or more clones and transpose it, as described for generating Cluster analysis data files.
- The format of the file will be slightly different in that the treatment and replicate data are in separate columns and come at the beginning of the data:

Treat	Rep	Group1	Group2	Group3	Group4	Group5	Group6
One	a	9	4	0	0	0	0
One	b	9	3	1	0	0	0
One	c	9	2	3	1	2	0
Two	a	0	4	1	0	1	0
Two	b	0	1	3	2	0	0
Two	c	0	4	5	1	2	0
Three	a	2	2	4	1	0	0
Three	b	2	3	3	0	0	0
Three	c	2	3	0	0	0	0
Four	a	1	3	1	0	0	5
Four	b	1	2	0	0	0	5
Four	c	1	7	0	0	0	5

- Save the file under a new name.

An example of a stepwise discriminant SAS program is below.

- Note that you must input the names of each group (or, the Group Number for each fingerprint group) in both the input statement and the proc stepdisc var statement, and that each group name must be recognized by SAS as a text string and not a number.
- To insert data for datalines, you can cut (CTRL + C) and paste (CTRL + V) from Excel spreadsheets.

```
options ls = 80 ps = 55 nocenter nodate;
/* Set up temporary SAS data set called stepdisc */
data stepdisc;
/* variables to be input are Treatment Replicate, and number of clones per
treatment and replicate in each group */
  input Treatment $ Replicate $ Group1 Group2 Group3 Group4 Group5 Group6;
/* The "$" indicates that it is a text variable*/
  title1 'Stepwise Discriminant Analysis';
  title2 'Example Data';
/* datalines statement to indicate data is about to begin */
datalines;
One   a     9     4     0     0     0     0
One   b     9     3     1     0     0     0
One   c     9     2     3     1     2     0
Two   a     0     4     1     0     1     0
Two   b     0     1     3     2     0     0
Two   c     0     4     5     1     2     0
Three a     2     2     4     1     0     0
Three b     2     3     3     0     0     0
Three c     2     3     0     0     0     0
Four  a     1     3     1     0     0     5
Four  b     1     2     0     0     0     5
Four  c     1     7     0     0     0     5
;
RUN;
proc stepdisc data=stepdisc;
   class Treatment;
   var Group1 Group2 Group3 Group4 Group5 Group6;
run;
```

An example of the output is below.

```
The STEPDISC Procedure
Stepwise Selection: Step 1

          Statistics for Entry, DF = 3, 8

Variable    R-Square    F Value    Pr > F    Tolerance

Group1        1.0000      Infty    <.0001       1.0000
Group2        0.1169       0.35    0.7885       1.0000
Group3        0.3577       1.48    0.2905       1.0000
Group4        0.3220       1.27    0.3493       1.0000
Group5        0.3253       1.29    0.3437       1.0000
Group6        1.0000      Infty    <.0001       1.0000
```

- The program will automatically select groups it thinks should be excluded (or which demonstrate differences in clone distribution between treatments), but we recommend looking instead at the data before any groups are excluded (or, "Step 1" data), since sometimes the number of steps the program can take is less than the number of groups that should be excluded.
- In the above example, two groups have a significant p-value, Group1 and Group6. These groups probably represent rRNA gene sequences which have different distributions between treatments.

(3) Stepwise Regression analysis: In this case you are looking for fingerprint groups in which the abundance of clones between treatments varies with some measurement of ecosystem function that varies between treatments. An example of an ecosystem function would be the ability of a soil to suppress a plant parasitic nematode. The data used in this analysis is identical to that used in the Stepwise Discriminant analysis, except this time there is an additional data column for the measurement.

Treat	Rep	Group1	Group2	Group3	Group4	Group5	Group6	Measurement
One	A	9	4	0	0	0	0	6
One	B	9	3	1	0	0	0	6
One	C	9	2	3	1	2	0	6
Two	A	0	4	1	0	1	0	0
Two	B	0	1	3	2	0	0	0
Two	C	0	4	5	1	2	0	0
Three	A	2	2	4	1	0	0	3
Three	B	2	3	3	0	0	0	3
Three	C	2	3	0	0	0	0	3
Four	A	1	3	1	0	0	5	2
Four	B	1	2	0	0	0	5	2
Four	C	1	7	0	0	0	5	2

An example of a stepwise regression SAS program follows:

```
options ls = 80 ps = 55 nocenter nodate;
data stepwisemeasure;
/* Measurement represents the property to be measured, such as ability to
digest waste or suppress plant disease */
  input Treatment $ Replicate $ Group1 Group2 Group3
Group4 Group5 Group6 measurement;
/* The "$" indicates that it is a text variable*/
  title1 'Stepwise Regression';
  title2 'Example of Measurement Data';
/* datalines statement to indicate data is about to begin */
datalines;
One   a     9     4     0     0     0     0     6
One   b     9     3     1     0     0     0     6
One   c     9     2     3     1     2     0     6
Two   a     0     4     1     0     1     0     0
Two   b     0     1     3     2     0     0     0
Two   c     0     4     5     1     2     0     0
Three a     2     2     4     1     0     0     3
```

```
Three b     2     3     3     0     0     0     3
Three c     2     3     0     0     0     0     3
Four  a     1     3     1     0     0     5     2
Four  b     1     2     0     0     0     5     2
Four  c     1     7     0     0     0     5     2
;
RUN;
PROC REG DATA=stepwisemeasure;
MODEL measurement=Group1 Group2 Group3 Group4 Group5 Group6/
SELECTION=Backward;
TITLE1 'Stepwise Regression';
/* Add & substract one at a time & compare f0 */
run;
proc corr Data=stepwisemeasure;
with Group1 Group2 Group3 Group4 Group5 Group6;
var measurement;
run;
end;
```

When the above program is run, the output looks like this:

```
Stepwise Regression                                                    13

The REG Procedure
Model: MODEL1
Dependent Variable: measurement

                          Analysis of Variance

                                    Sum of           Mean
Source                      DF       Squares         Square    F Value    Pr > F

Model                        6       53.68920        8.94820      17.47    0.0033
Error                        5        2.56080        0.51216
Corrected Total             11       56.25000


Root MSE              0.71565    R-Square     0.9545
Dependent Mean        2.75000    Adj R-Sq     0.8998
Coeff Var            26.02378


                       Parameter Estimates

                     Parameter       Standard
Variable       DF     Estimate          Error    t Value    Pr > |t|

Intercept       1      1.11042        0.85450       1.30      0.2505
Group1          1      0.61054        0.06991       8.73      0.0003
Group2          1     -0.08145        0.17714      -0.46      0.6650
Group3          1      0.26296        0.21564       1.22      0.2771
Group4          1     -0.72013        0.50774      -1.42      0.2153
Group5          1     -0.53650        0.34683      -1.55      0.1826
Group6          1      0.10344        0.12499       0.83      0.4456
```

```
Pearson Correlation Coefficients, N = 12
       Prob > |r| under H0: Rho=0

            measurement

Group1          0.94714
                 <.0001

Group2         -0.06580
                 0.8390

Group3         -0.24492
                 0.4430

Group4         -0.28563
                 0.3682

Group5         -0.08872
                 0.7839

Group6         -0.20000
                 0.5331
```

- The program will automatically select groups it thinks should be excluded (or which demonstrate correlations between the number of clones per treatment and "measurement"), but we recommend looking instead at the data before any groups are excluded (or, "Step 1" data), since sometimes the number of steps the program can take is less than the number of groups that should be excluded.
- It may also be useful to skip the stepwise regression altogether and just focus on the correlation coefficients, which will give you the same information. Note that the r^2 information also tells you whether the correlation is positive or negative.
- In the above example, one group has a significant p-value at Step 1 and for its correlation between clone number and "measurement", Group1. This group probably represents rRNA gene sequences that change in abundance as the level of "measurement" changes.

The CORR Procedure

Pearson Correlation Coefficients, N = 12
Prob > |r| under HO: Rho=O

	measurement
Group6	-0.20000
	0.5331