Instructions for Analyzing μArray Data

From the computer without SAM installed
Read Howto Install SAM on XP.doc for instructions on installing SAM.

From the computer with SAM installed
Launch Microsoft Excel
Select File, Open (Ctrl+O) and choose the file for analysis (Read Howto Grid μArray Images.doc and Howto Isolate One Median.doc for instructions on converting images to numeric values and preparing the numeric values for analysis respectively).

At this point, the excel sheet contains one median value for each unique antigen. Now we want to process these medians for statistical analysis. The standard procedure for processing μArray data is to set all low intensity values to some nominal value. In our case, we use the nominal value of 10. Any value less than 10 is assigned a value of 10. All other values keep their original value. The next step is to divide each value by some constant. In our case, we use the value of 300. This value scales the data for optimal visual output (see image below).

The numbers across the top represent the value of the divisor. The numbers along the right column represent the value of the raw data before any processing (the contrast setting is at 5.5).

The final step for processing the data is to take the log base 2 of each value. The log scale is useful for bringing out differences in reactivity in both a visually appealing way and an analysis friendly manner. Below is a summary of the algorithm used process the data.

- if "raw value" < 10 assign it to 10
- else "raw value" = "raw value"
- divide all each value by 300
- take the log2 of the final answer

I consider the data to be processed after all the bulleted items are complete.

From processed to formatted
The instructions that follow guide one to do a two-class, unpaired data analysis (see Explanation of Two-Class, Unpaired Data Analysis that appears below for a description of what that means). A multiclass analysis is identical to a two-class, unpaired data analysis except there are more than two groups and the Data in Log Scale? option is not available.

---

1 The program we use for μArray analysis is called Significance Analysis of Microarrays (SAM)
2 Antigen refers to proteins and peptides.
3 I will only describe the two-class, unpaired and the multiclass analyses in this document because they are the only ones I have used in analyzing my data. Please see the SAM manual (sam.pdf) or the examples for a description of additional analyses SAM is capable of doing.
Insert a row (shown in blue) beneath the slide numbers. This row will contain the group labels for the analysis.

Compute the standard deviation of each antigen by selecting an entire row of data and calculating the standard deviation. It’s probably best for the output to appear on the far right of the data (shown in plum).

Sort the standard deviation column in ascending order to bring the antigens with the smallest variation to the top of the list.

Eliminate any antigen (entire row) whose standard deviation is zero.

Add a group number (blue cells) beneath each sample.

Now the sheet is formatted for conducting a SAM analysis.

From formatted to analyzed

- Highlight the following cells only – unique id (rose), antigen names (pale blue), data (tan), and group numbers (blue).
- Select the SAM button
  - Select an analysis from the Choose Response Type
  - If conducting a two-class, unpaired data analysis check Logged (base 2), otherwise leave alone
  - Select the OK button

SAM will create two new worksheets SAM Plot & SAM Output. The SAM Plot worksheet appears with the SAM Plot Controller dialog box. One can adjust the number of significant genes that are included or excluded in the output by putting a number in the Fold Change box or adjusting the value of the Delta Value. The fold change selects only the significant genes with a fold change greater than the value entered. The delta value adjusts the q-value threshold. A higher delta value means the output reflects antigens with lower q-values. 

## Diagram for Reference

The sheet containing processed data (shown in tan) appears in a format where the slide numbers and sample names appear across the first row (shown in lavender) and the antigens appear along the first column (shown in pale blue).

Make sure the text in the upper left corner (shown in rose) is anything besides “name” (e.g. unique id)

1 Excel will rarely output an exact zero because numerical computations involve truncating values. Therefore, a value like X.XXXXXE-07 is zero.

2 The q-value represents the chance that the antigen is really a false positive. It is the lowest false discovery rate where the antigen is considered significant. As a rule of thumb, however, one can think of the q-value as being similar to a p value. Therefore, for maximum confidence in how significantly different the groups are, pick antigens with q-values < 5%.
Explanation of Two-Class, Unpaired Data Analysis

All two-class analyses follow the form of a question, “Is there a difference in reactivity between (1) and (2)?” SAM outputs many parameters that help us decide which differences in antibody-antigen reactivity are statistically significant between the two groups. I’m briefly reviewing the parameters since they are relevant in choosing antigens for further inquiry. More information can be found in the SAM documentation.

A typical two-class output looks like the following. Note that gene means antigen in our case. The developers created SAM for gene microarray analysis instead of protein microarray analysis. I will use genes because they appear below, but know that I really mean antigen.

## Two-Class, Unpaired SAM Output

### 8 Positive Significant Genes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene Name</th>
<th>Gene ID</th>
<th>Score (d)</th>
<th>Number (r)</th>
<th>Difference (s + s0)</th>
<th>Fold Change</th>
<th>q-Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>221 222 PIP 25-39</td>
<td>-3.769060761683962</td>
<td>5.14276719616969</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>2</td>
<td>443 444 PIP 40.89-141</td>
<td>-1.008600731620542</td>
<td>5.43604683736879</td>
<td>3.31474007521092</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>3</td>
<td>125 126 PIP 130-161</td>
<td>0.676940565682032</td>
<td>5.14276719616969</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>4</td>
<td>35 36 PIP 26-39</td>
<td>-2.30223838176301</td>
<td>2.9543071256114</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>5</td>
<td>243 244 Pip All</td>
<td>2.152381478704812</td>
<td>5.1711198562052</td>
<td>0.198182002</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>6</td>
<td>13 14 PIP 1-11</td>
<td>4.0491318923920</td>
<td>3.16468717 2.53416815</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>262 Pip RNA 1</td>
<td>1.1624810188120</td>
<td>3.16468717 2.53416815</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>143 144</td>
<td>2.152381478704812</td>
<td>5.1711198562052</td>
<td>0.198182002</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
</tbody>
</table>

### 193 Negative Significant Genes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene Name</th>
<th>Gene ID</th>
<th>Score (d)</th>
<th>Number (r)</th>
<th>Difference (s + s0)</th>
<th>Fold Change</th>
<th>q-Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>261 262 Pip RNA Polymers</td>
<td>-2.093092415164262</td>
<td>-2.9543071256114</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>2</td>
<td>223 224 Pip 19-39</td>
<td>-1.082027460412302</td>
<td>-2.9543071256114</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>3</td>
<td>131 132 Pip 1-11</td>
<td>-1.229273238176301</td>
<td>-3.3619516314201</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>4</td>
<td>92 Pip ABC 16</td>
<td>121-143</td>
<td>-5.297578392406201</td>
<td>-2.9543071256114</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
</tr>
<tr>
<td>5</td>
<td>12 PIP 1-15 70-12 90</td>
<td>2.152381478704812</td>
<td>-4.718953080329</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
<tr>
<td>6</td>
<td>13 PIP 1-15 70-12 90</td>
<td>2.152381478704812</td>
<td>-4.718953080329</td>
<td>0.46675635107581</td>
<td>0.6500196314201</td>
<td>1.2954314201</td>
<td>0.000196314201</td>
</tr>
</tbody>
</table>

The number of positive significant genes refers to the genes that are positively correlated while the negative significant genes are the negative correlations. A positive correlation means that reactivity of group 1 is higher than the reactivity of group 2 and the opposite is true for a negative correlation.

- The row refers to where the gene is located on the excel spreadsheet.
- The Gene Name is the Unique Id of the antigen.
- The Gene ID is a hyperlink to a gene database (we don’t use this feature).
- The score (d) represents the value of the T-statistic. A higher score means a larger difference between the two groups.
- The numerator (r) represents the difference between the means of the two groups. A larger absolute value of the numerator says that difference between the means of the two groups is greater.
- The denominator (s + s0) represents the denominator of the T-statistic (we don’t really care about this value).
- The fold change is the ratio of the averages of the two groups.
- The q-value (%) was explained above.

### Explanation of Multiclass Analysis

A multiclass analysis outputs the same information as the two-class except for the fold-change since a fold-change cannot be computed in a multiclass analysis.