

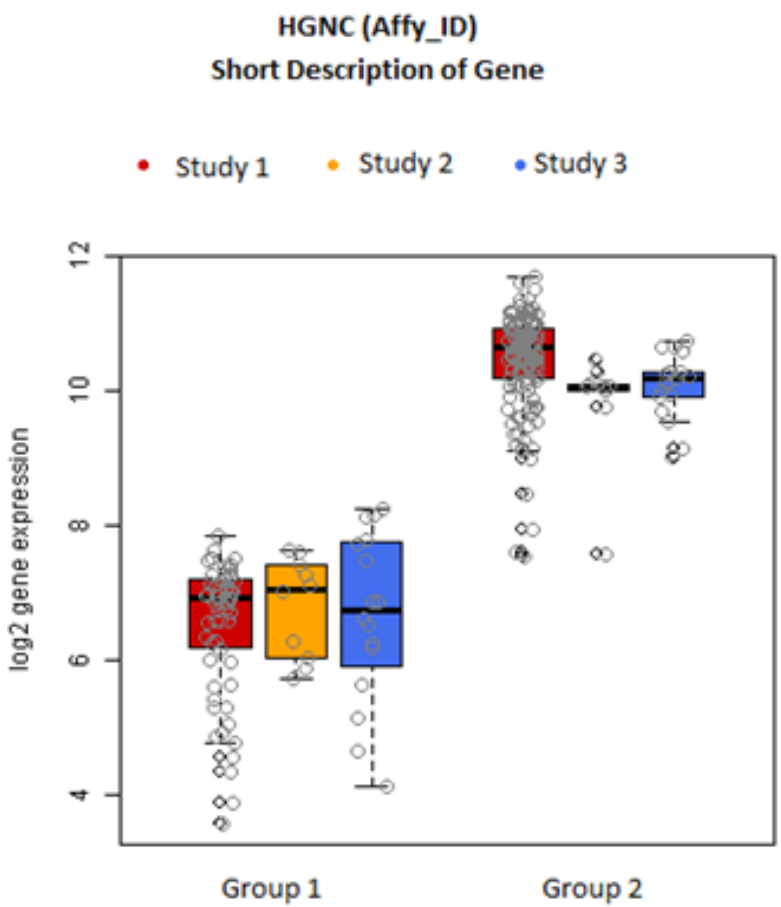
Science For A Better Life

Grouped Jittered Box Plots

in SAS 9.2 and SAS 9.3

Volker Harm and Catalina Mejía Herrera, PhUSE Conference 2012, Budapest

Presenting Results of Gene Expression Analysis



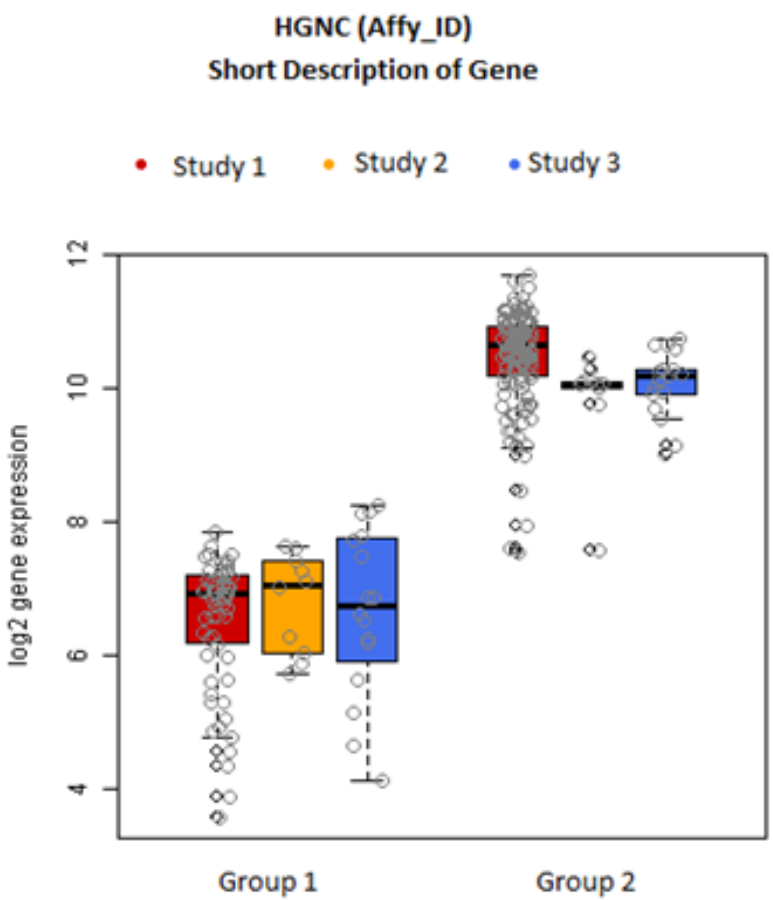
Typical task

- Present results of two factors for different studies by box plots
- Guide the audience in interpreting the boxplots by overlaying the individual data
  - spreaded for a certain random amount within the area of the box plot to reduce overplotting
- Make the graphs comparable by a standardized familiar appearance

Standard tools used

- Bioconductor software that uses the R statistical programming language
- R functions boxplot () and stripchart () to create the graphs

The Data



Measurements

Variable name	Description	Type
Affy_ID	Affymetrix gene identification	Character
GroupNo	Group number	Numeric
GroupName	Group name	Character
StudyNo	Study number	Numeric
StudyName	Study name	Character
Value	Log2 gene expression	Numeric

Jittering

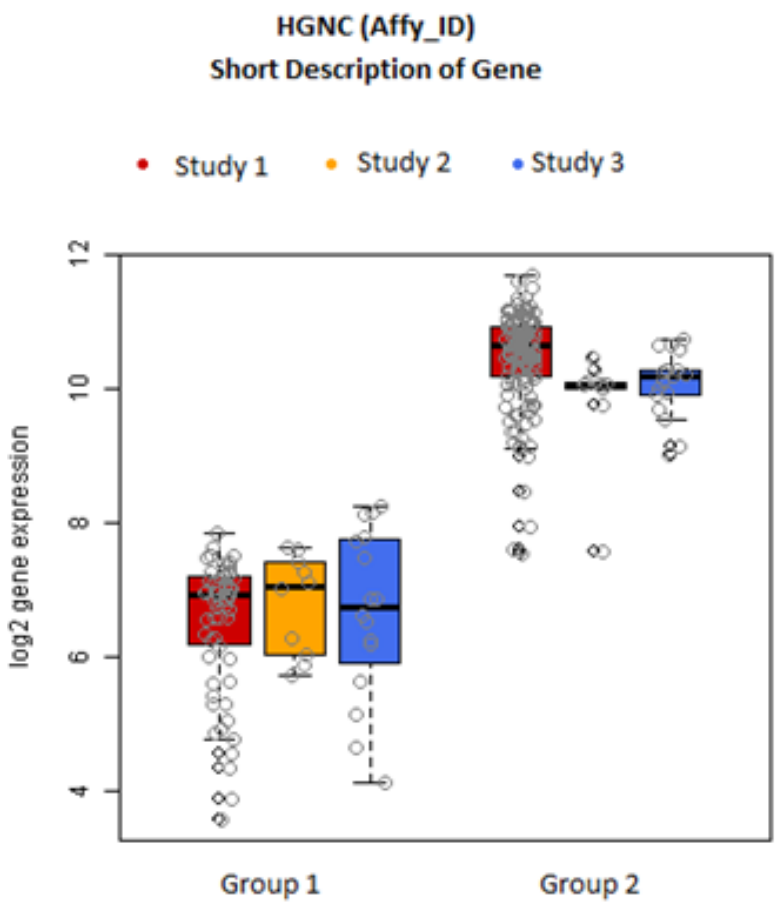
StudyNoJittered = StudyNo + 0.3*(uniform (1) – 0.5

Gene Information

Variable name	Description	Type
Affy_ID	Affymetrix gene identification	Character
HGNC	Unique gene name	Character
Short_Description	Short description of gene	Character

Example data sets will be given in the paper

The Task: Develop a SAS Macro



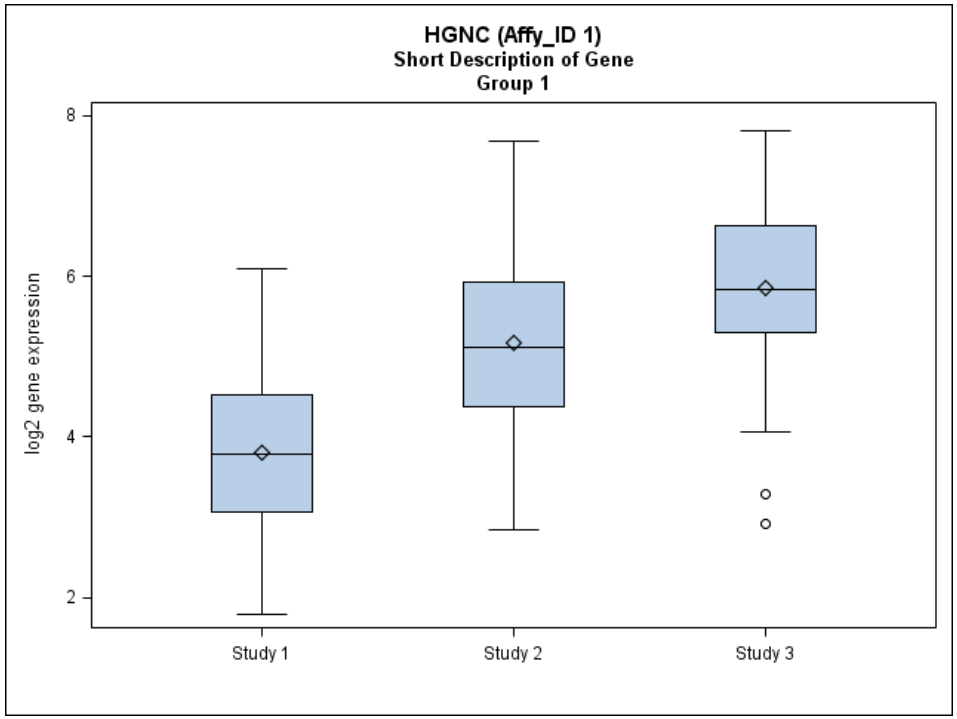
Given

- the set of data for each gene in the measurements table and
- the metadata in the gene information table

For each gene create a graph displaying

- the log2 gene expressions
  - for each group
    - for each study
      - by a boxplot (schematic, without mean) with the individual measurements overlaid jittered within the boxplot area
      - indicated by a color
- the study names
  - in a legend above the graph
  - indicated by different colors
- the group names below the x-axis.
- the gene information in the titles

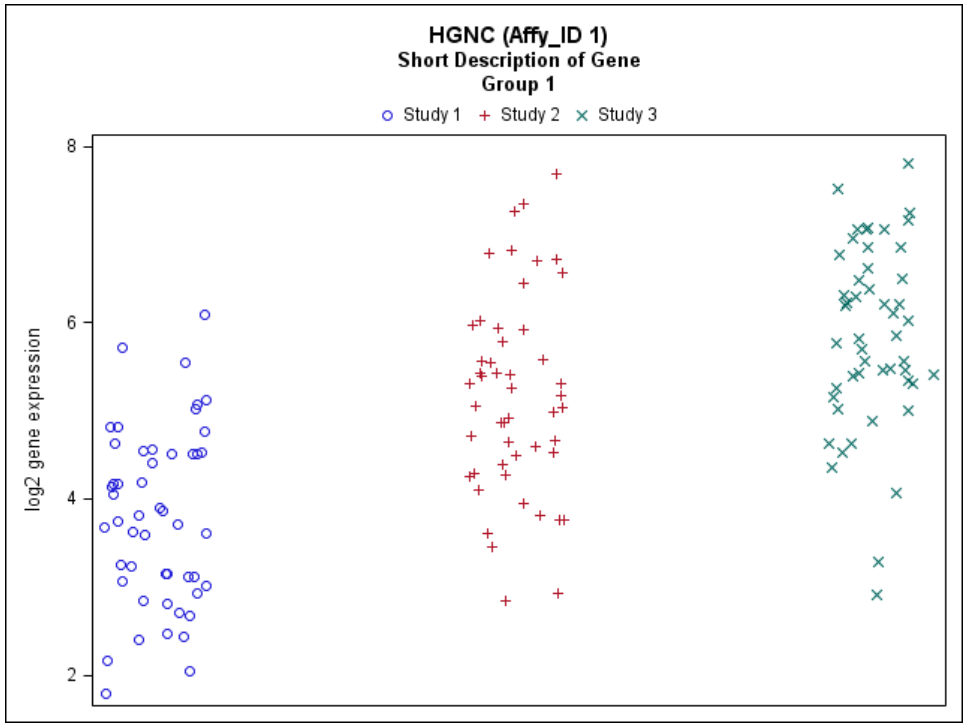
SAS 9.2: Boxplot with proc sgplot




```
%macro ExpressionBoxPlotGroup1;
  proc sgplot data = Expressions;
    xaxis label = " ";
    vbox Value/ category = StudyName;
    where GroupName = 'Group 1';
  run;
%mend;
```

- The vbox statement creates the boxplot for each study in the group.
- The x-axis label is suppressed with xaxis statement.
- Titles are set outside the macro.
- Example for one group

SAS 9.2: Jittered Expressions with proc sgplot




```
%macro ExpressionScatterPlotGroup1;
  proc sgplot data = Expressions;
    xaxis display = none;
    scatter x = StudyNo + 0.3*(uniform (1) - 0.5)  y = Value/
       group = StudyName;
    keylegend/ position = top noborder;
    where GroupName = 'Group 1';
  run;
%mend;
```

- The scatter statement draws the jittered measurements for each study in the group.
- The x-axis is suppressed with xaxis statement.
- The keylegend statement allows a colored legend above the plot area.
- Titles are set outside the macro.
- Example for one group

Overlaying the Plots Does Not Work

Box plots cannot be combined with any other plot types.

```
%macro OverlayPlots;
 proc sgplot data = GraphData;
  xaxis label = " ";
  vbox Value/ category = StudyName;
  scatter x = StudyNoJittered y = Value/ group = StudyName;
 keylegend/ position = top noborder;
 where GroupName = 'Group 1';
run;
%mend;
```

```
ERROR: Attempting to overlay incompatible plot or chart types
NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           1.31 seconds
      cpu time            0.65 seconds
```

This is true for proc sgplot and proc sgpanel in SAS 9.2 and SAS 9.3.

Furthermore: Box plots cannot be combined with any other plot types.

How to get on in SAS 9.2

IF THERE IS ONLY ONE LEVEL OF GROUPING

- **Graph Template Language (GTL)** (approach also discussed in Pratt. 2012)
  - Box and scatter plots can be overlaid, but have different types of axes (discrete and linear).
  - There are two independent x-axes, which each of the two plots can be assigned to.
  - These axes are built independently from each other and one has to care to get them synchronized. This can be done choosing clever formats.

EXTENDING IT TO TWO LEVELS OF GROUPING

- **Layout datapanel/layout datalattice**
  - The boxplot statement is not available in these layouts, because it is a *computed* plot.
  - The boxplot statement could be replaced by using the boxplotparm statement (This creates a *parametrized* plot that can be used in layout datapanel or layout datalattice.
  - Box plots can be produced, but there is no second x-axis available in these layouts to get the different axes adjusted.
- **Layout lattice** (see Holland. 2010. for an introduction to layout lattice)
  - Layout lattice puts single plots in a lattice.
  - Preprocessing the data into a wide format acceptable results are possible.
  - Needs a lot of programming to embed it in a general macro..

The vector Statement Solution: Building our own Box Plots (Harris. 2010)

The vector Statement

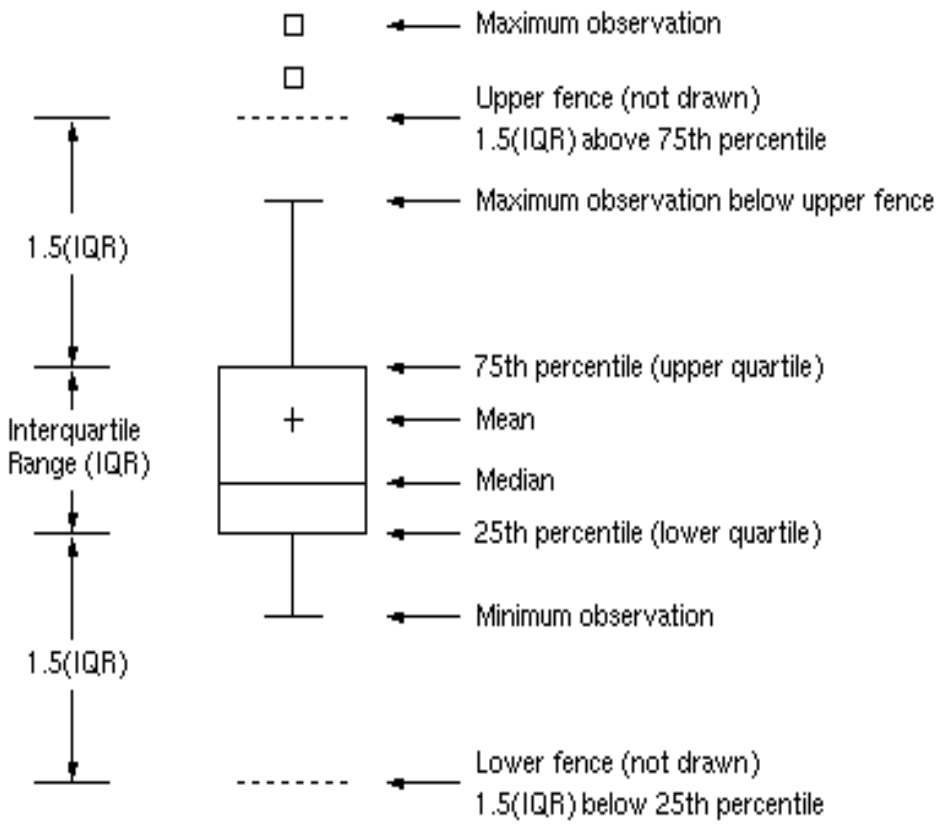
- Available for sgplot and sgpanel with SAS 9.2 Phase 2 and later
- Draws arrows from a point of origin to each data point
- Creates a basic plot, which can be overlaid by a scatter plot
- Has a linear axis
- Has an option noarrowheads

Getting the Statistics for the Box Plots

```
proc means data = Expressions noprint;
   by GroupName StudyNo;
   var Value;
   output out = BoxPlotStats mean = mean median = median q1 = q1 q3 = q3 qrange = qrange min = min max = max;
run;
```

- Merge the statistics to the expressions data set to get a graph data set

The Anatomy of a Schematic Box Plot

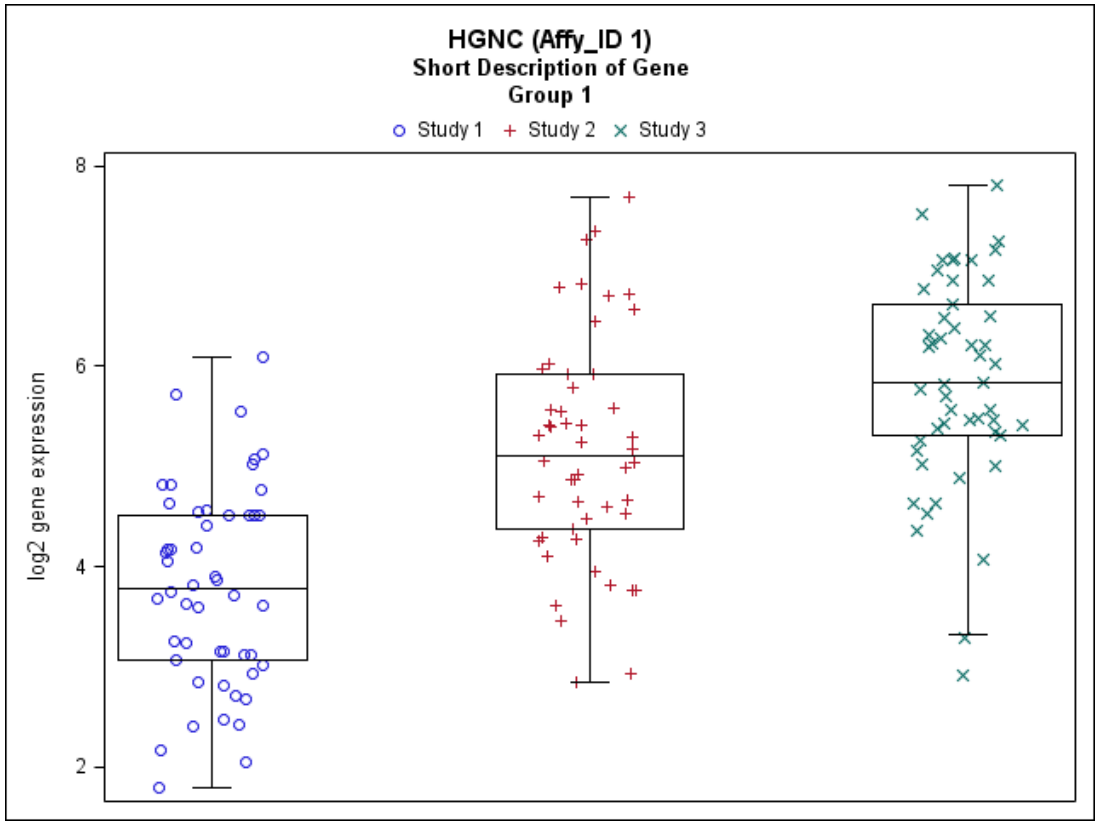



```
* calculate box coordinates;
 data GraphData;
   set GraphData;
   XBoxLeft = StudyNo - 0.25;
   XBoxRight = StudyNo + 0.25;
   XWhiskerLeft = StudyNo - 0.05;
   XWhiskerRight = StudyNo + 0.05;
   YWhiskerTop = (1.5*qrange) + q3;
   YWhiskerBottom = q1 - (1.5*qrange);
   if max le YWhiskerTop then
     YWhiskerTop = max;
   if min ge YWhiskerBottom then
     YWhiskerBottom = min;
run;
```

Drawing the Box Plots

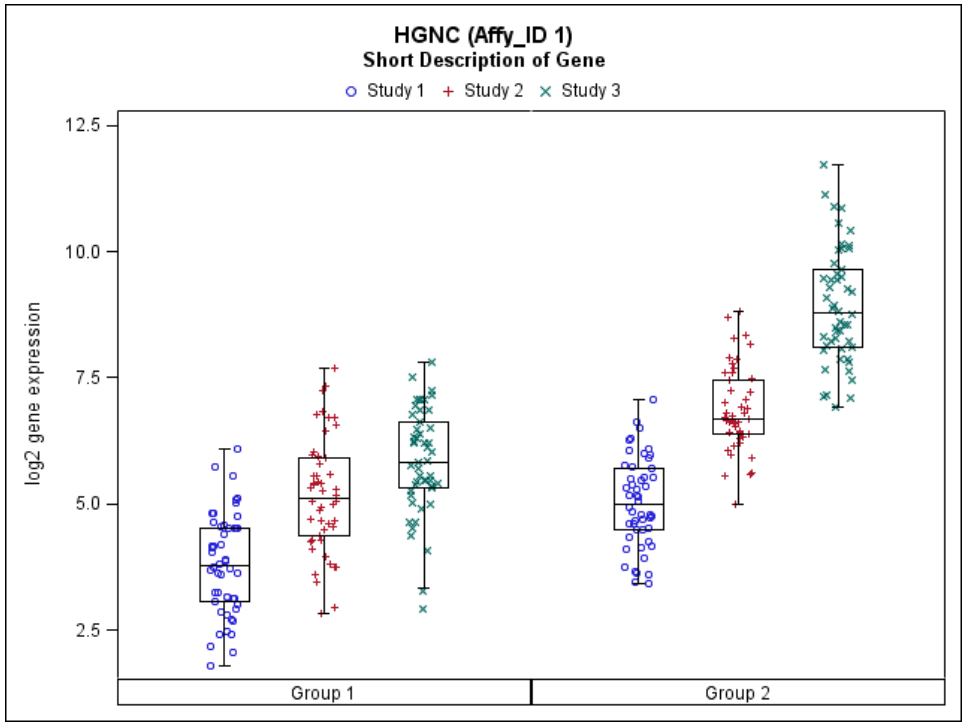
```
%macro DrawBoxPlots;
   * draw median line;
   vector x = XBoxRight y = median / noarrowheads xorigin = XBoxLeft yorigin = median;

   * draw quantile box;
   vector x = XBoxRight y = q1/ noarrowheads xorigin = XBoxLeft yorigin = q1;
   vector x = XBoxRight y = q3/ noarrowheads xorigin = XBoxLeft yorigin = q3;
   vector x = XBoxLeft y = q3/ noarrowheads xorigin = XBoxLeft yorigin = q1;
   vector x = XBoxRight y = q3/ noarrowheads xorigin = XBoxRight yorigin = q1;

   * draw whiskers;
   vector x = StudyNo y = YWhiskerTop/ noarrowheads xorigin = StudyNo yorigin = q3;
   vector x = StudyNo y = YWhiskerBottom/ noarrowheads xorigin = StudyNo yorigin = q1;
   vector x = XWhiskerRight y = YWhiskerTop/ noarrowheads xorigin = XWhiskerLeft yorigin = YWhiskerTop;
   vector x = XWhiskerRight y = YWhiskerBottom/ noarrowheads xorigin = XWhiskerLeft yorigin = YWhiskerBottom;
%mend;
```

Vector Solution for One Group



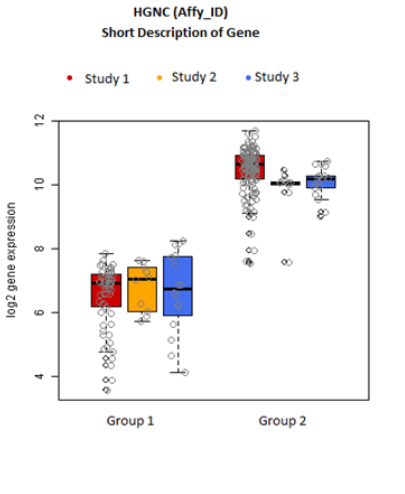

```
proc sgplot data = GraphData;
   keylegend/ position = top noborder;
   xaxis display = none;

   * draw jittered values;
   scatter x = StudyNoJittered y = Value/
       group = StudyName;
   keylegend/ position = top noborder;

   %DrawBoxPlots;
run;
```

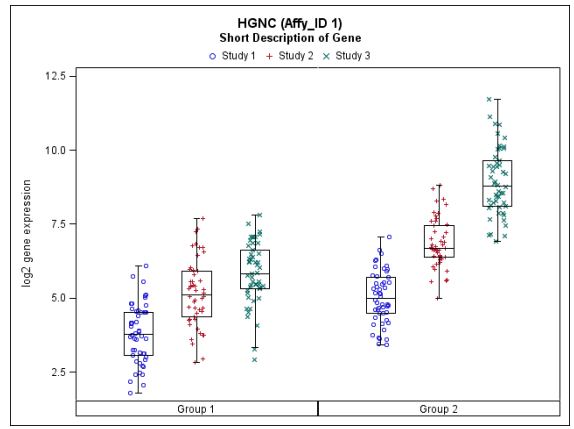
Vector Solution with proc sgpanel




```
proc sgpanel data = GraphData;
   * define the structure of the panel;
   panelby GroupName/
      novarname layout = columnlattice
      colheaderpos = bottom noborder;

   * leave some room between the columns of the panel;
   colaxis display = none offsetmax = 0.2 offsetmin = 0.2;

   * draw jittered values;
   scatter x = StudyNoJittered y = Value /
      group = StudyName;
   keylegend/ position = top noborder;

   %DrawBoxPlots;
run;
```

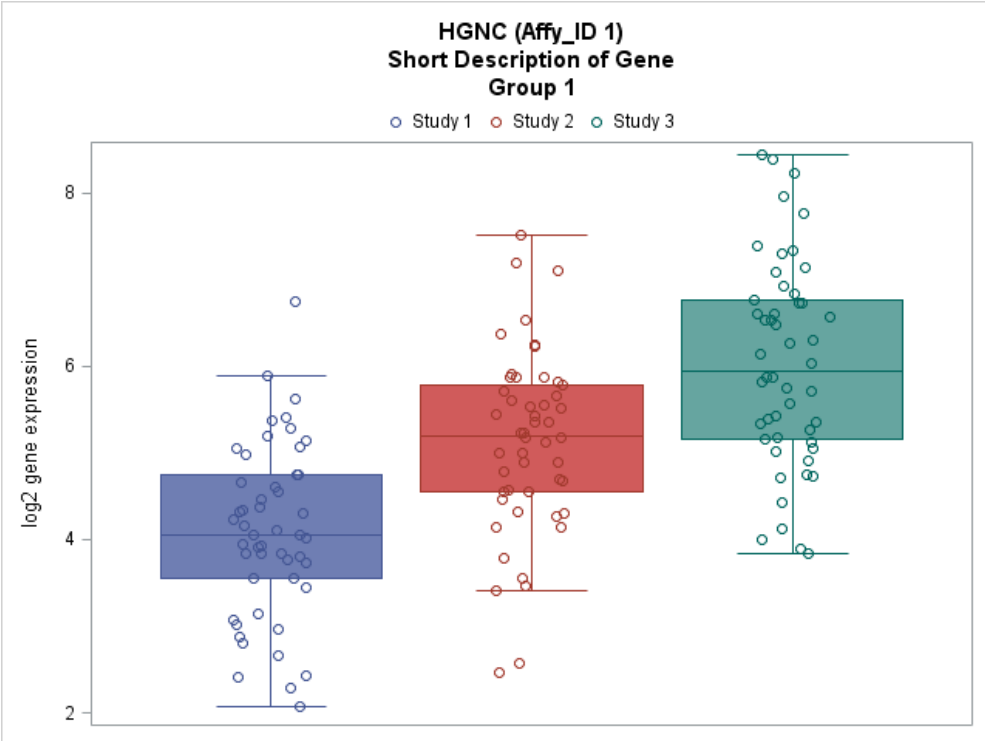
What Do We Have Achieved?





+ Box plots with jittered data overlaid
+ Study indication by colors and legend
+ Grouping possible
+ Correct handling of titles and axis labels
+ Easy handling and programming
- Data are colored, not the boxes
- Group indication cannot be changed flexibly in proc sgpanel
- No SAS functionality used for producing the box plots

What´s New in SAS 9.3?



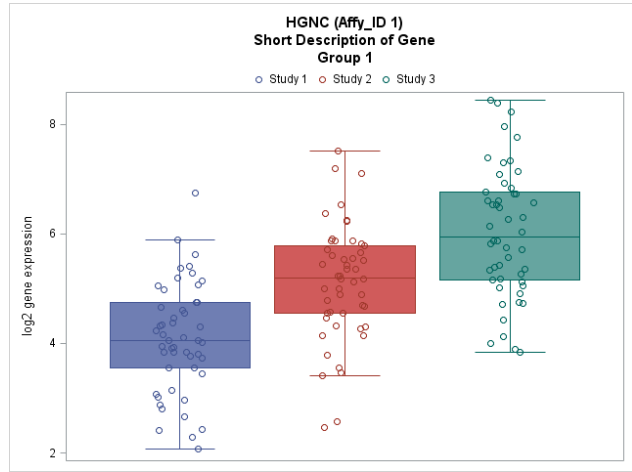
- Boxplots support an independent, numeric axis.
- There is a new group option for box and scatter plots.

That means:

- In the Graph Template Language (GTL) box and scatter plots can easily be overlaid.
- The grouping feature allows to get filled boxes in an easy way.

proc template Code for One Group

```
proc template;
    define statgraph Overlaid;
      begingraph /;
        EntryTitle "HGNC (Affy_ID 1)";
        EntryTitle "Short Description of Gene";
        EntryTitle "Group 1";
        layout overlay/ xaxisopts = (display = none type = linear);
          BoxPlot x = StudyNo y = Value/ Group = StudyNo display = (caps median fill) LegendLabel = "log2 gene
            expression";
          ScatterPlot x = StudyNoJittered y = Value/ Group = StudyName name = "SCATTER";
          DiscreteLegend "SCATTER"/ Valign = top Border = false;
        endlayout;
      endgraph;
    end;
run;
```



Grouped Jittered Box Plots in SAS 9.3

In the GTL layout datalattice the graphs for one group can be combined, but in layout datalattice only the boxplotparm statement is available.

For boxplotparm the statistics must be calculated before creating the graph.

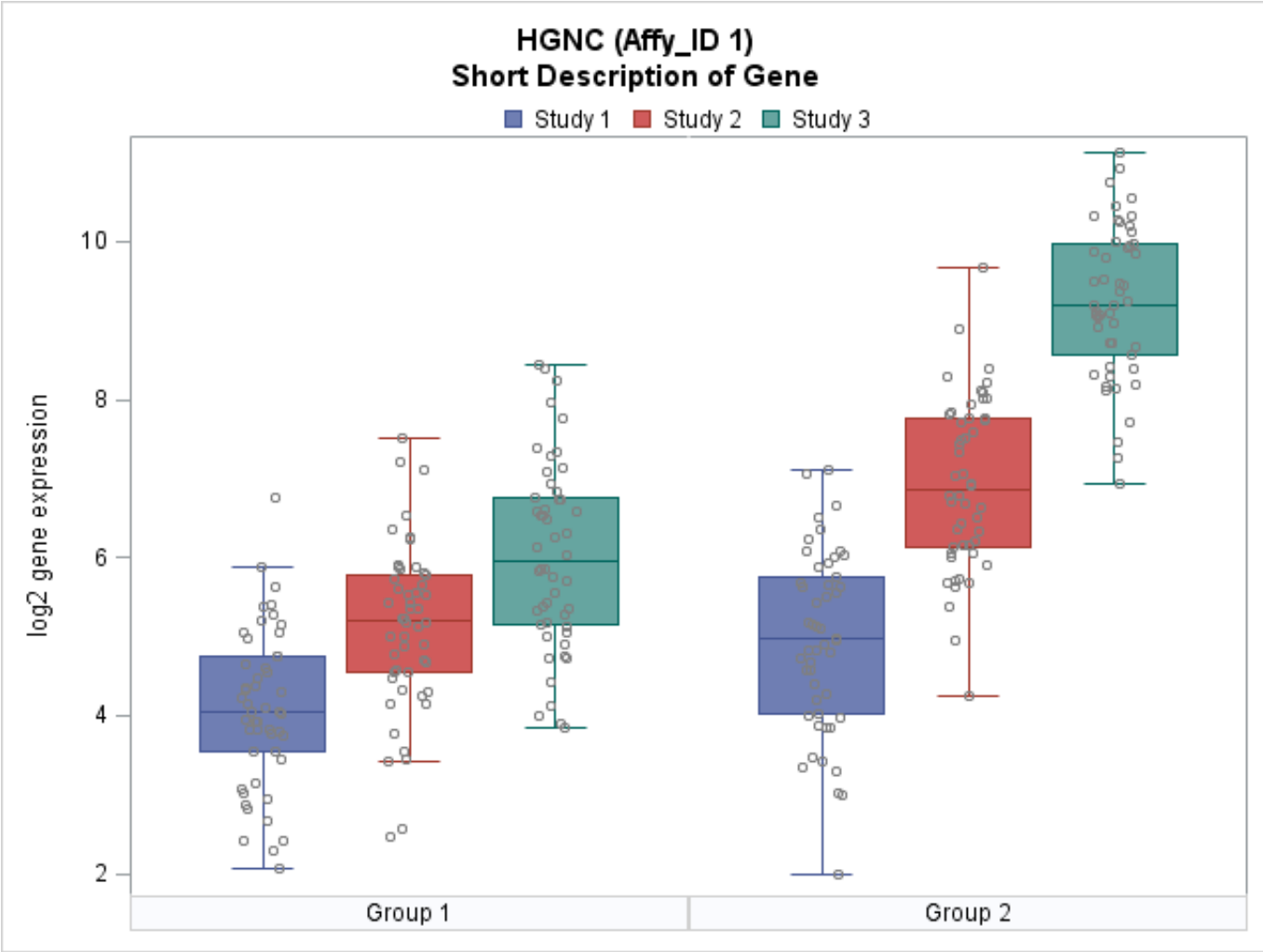
SAS provides a macro %boxcompute for this.

The results of this macro have to be added to the graph data set.

The final input data set has the following structure:

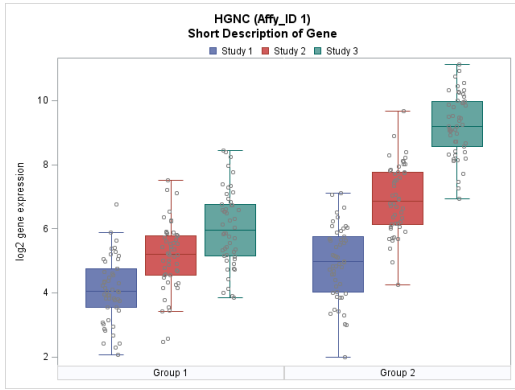
Variable name	Description	Type
STAT	Box plot statistic	Numeric
BoxValue	Value of box plot statistic	Numeric
Affy_ID	Affymetrix gene identification	Character
GroupNo	Group number	Numeric
GroupName	Group name	Character
StudyNo	Study number	Numeric
StudyName	Study name	Character
StudyNoJittered	Jittered study number	Numeric
Value	Log2 gene expression	Numeric

Solution with layout datalattice

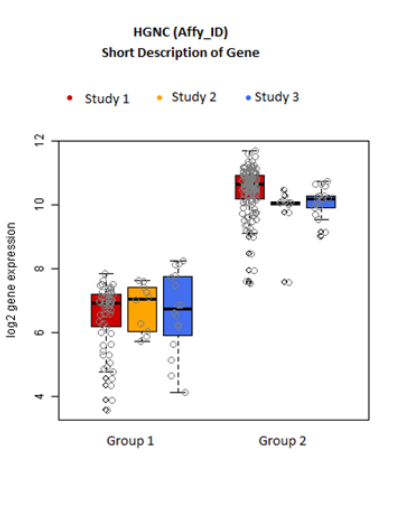


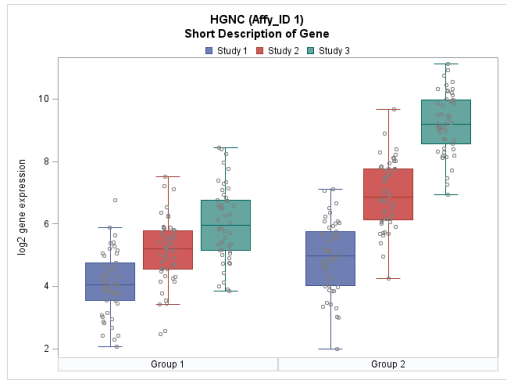
The Template

```
proc template;
   define statgraph JitteredGroupedBoxPlots;
      begingraph;
         entrytitle  "&HGNC. (%trim (&Affy_ID.))";
         entrytitle "&Short_Description.";
         layout datalattice columnvar = GroupName/ columnheaders = bottom border = false
            columnaxisopts = (display = none offsetmax = 0.2 offsetmin = 0.2) headerlabeldisplay = value;
            layout prototype/ walldisplay = none;
               boxplotparm x = StudyNo Y = BoxValue stat = STAT/ display = (caps fill median) group = StudyName name = 'Box';
               scatterplot x = StudyNoJittered y = Value/ primary = true markerattrs = (color = gray);
            endlayout;
            sidebar/ align=top;
               discretelegend "Box" /border = false;
            endsidebar;
         endlayout;
      endgraph;
   end;
run;
```



What Do We Have Achieved?





+ Box plots with jittered data overlaid
+ Study indication by colors and legend
+ Grouping possible
+ Correct handling of titles and axis labels
+ Easy handling and programming
+ Boxes are colored
- SAS functionality used for producing the box plots
- Group indication in datalattice style

Take Home Message

It is not just a snip to generate Jittered Grouped Box Plots.

There are quite flexible good looking solutions possible in SAS 9.2 and 9.3.

It was a lot of fun exploring all this: I like ODS Graphics.



Science For A Better Life

Thank you!