**JMP Genomics Introductory Tutorial**

Developed by Kai Knutson ’11 and Milan Cvitkovic ‘13

JMP Genomics is a software package developed to analyze whole transcriptome data. It is built on the SAS statistical software package and is useful in identifying differences in the patterns of gene expression in different tissues. JMP Genomics will help you begin to ask questions at the scale of the whole genome. It is a very powerful and complex program. It is not an effective use of your time to try and understand what all the different parameters are in the program.

The goals of this tutorial are to give you a brief introduction to the software, to help you create useful visualizations of whole transcriptome data, and to get you thinking about asking genomics-scale questions.

## Part 1: Setting Up the Program and Importing/Preparing your Data

**1)** Start the computer and hold down the "Option" key.

**2)** Select "Windows XP" and click the "Return" key.

**3)** When the OS has loaded, create a folder on the desktop titled "FirstName\_LastName\_JMP\_Genomics\_Files" by right-clicking the desktop and selecting "New" > "Folder". Then create a second folder on the desktop titled “Folder of RAW files.’ In this folder download and save the JMP Genomics exercise file that can be found on the Explorer in the Gene Expression section. The link is called ‘Data File for JMP Genomics Gene Expression Exercise.’ The file name is ‘jmp\_exercise\_data.txt.’

**4)** Open JMP Genomics 4 by selecting Start >Programs > SAS > JMP > JMP Genomics 4. Close any windows that open automatically within the program.

**5)** From the menu "Genomics" select "Import" > "Text" > "Import Individual Text, CSV, or Excel Files".

**6)** ***a)*** Import the file titled ‘jmp\_exercise\_data.txt’ in your Folder of Raw Files.

***b)*** In "Row Number of Variable Names" enter "1".

***c)*** In "Data Start Row" enter "2".

***d)*** In "Output Folder" select the folder you created on the desktop.

***e)*** Do not change any of the other settings, click "Run".

**7)** When the process is complete an SAS Message will be displayed, click "Open".

**8)** Close the file and all other open windows within the program. From the menu "Genomics" "Data Set Utilities" > "Transform".

**9)** ***a)*** In "Input Data Set" select the file imported to the folder you created on the desktop with your name. This file should be named ‘jump\_exercise\_data.sas7bdat’.

***b)*** Select all available variables except "Transcript" and “NRn” by holding down the shift key and clicking each variable. Click the arrow button to select these variables to be transformed.

***c)*** In "Type of Transformation" select "log2".

***d)*** In "Shifting Factor" enter "1".

***e)*** In "Output Folder" select the folder you created on the desktop.

***f)*** Do not change any of the other settings, click "Run".**10)** From the menu "Genomics" select "Experimental Design File" > "Create a Design Data Set from an Existing Data Set".

**11)** ***a)*** In "Input Data Set" select your transformed data file (‘jump\_exercise\_data\_dtf.sas7bdat’)

***b)*** Select all "Available Variables" except "Transcript" and “NRn” by holding down the shift key and clicking each variable. Click the arrow button to select these variables for analysis.

***c)*** In "Output Folder" select the folder you created on the desktop.

***d)*** Do not change any of the other settings, click "Run".

**12)** When the process is complete an SAS Message will be displayed, click "Open".

**13)** Right-click the column titled "Label of Former..." and click "Delete Columns" (under the Columns option).

**14)** Right-click in the space next to the column titled "ColumnName" and click "New Column".

**15)** a) In "Column Name" enter "Tissue".

b) In "Data Type" select "Character".

c) Do not change any of the other settings, click "OK".

**16)** Enter the following data in the column "Tissue":

Nm, No, Np- “Nodules"

Rj, Rk, Rl- “Roots"

Sa, Sb, Sc- "Vegetative Shoots"

Sd, Se- "Pre-Floral Shoots"

Sf, Sg, Sh, Si- "Floral Shoots"

**17)** From the menu "File" select "Save As", in "Save as type" select "SAS Data Set (\*.sas7bdata,\*.sd2)", click "Save".

*You have now imported your transcriptome data into JMP and modified it to be useable by the program and understandable to you.*

## Part 2: Generating Visualizations

**18)** Close the file and all other windows open within the program. From the menu "Genomics" select "Workflows" > "Basic Expression Workflow”.

**19)** ***a)*** In "Study Name" enter "FirstNameLastName".

***b)*** In "Input Data Set" select your transformed data set (‘jump\_exercise\_data\_dtf.sas7bdat’)

***c)*** From "Available Variables" select "Transcript" and click the arrow next to "Label Variable". Select "Transcript" again and click the arrow next to "Variables to Keep in Output or By Which to Merge Annotation Data".

***d)*** In "Output Folder" select the folder you created on the desktop.

***e)*** Click the tab "Experimental Design".

***f)*** In "Experimental Design Data Set" select your experimental design file (‘jump\_exercise\_data\_dtf\_exp.sas7bdat’).

***g)*** From "Available Variables" select "Tissue" and click the arrow next to "Color Variables".

***h)*** Click the tab "QC and Normalization".

***i)*** Check the boxes next to "Distribution Analysis", "Correlation and Principal Components Analysis" and "Correlation and Grouped Scatterplots".

***j)*** Under "Normalization Method", select "None".

***k)*** Click the tab "ANOVA".

***l)*** From "Available Variables" select "Tissue" and click the arrow next to "Class Variables".

***m)*** In "Model these Fixed Effects" enter "Tissue".

***n)*** Click the tab "LSMeans".

***o)*** In "Estimate LSMeans for these Fixed Effects" enter "Tissue".

***p)*** Under "LSMeans Difference Set for Volcano Plots" enter "Simple Differences".

***q)*** Click the tab "Multiple Testing".

***r)*** Under "Multiple Testing Method" select "FDR".

***s)*** Under "Alpha" enter "0.05".

***t)*** We will not be using "Annotation" or "Tracks".

***u)*** Do not change any of the other settings, click "Run".

**20)** When the process is complete a dialogue titled “Journal” will be displayed. From this “Journal” window we will generate a number of useful visualizations of our data. Information about what each of the visualizations means can be found in the Data Analysis section at the end of this document. For now, click the “Results” button below “Process 1 – DataDistribution (NameOfStudy)”.**21)** In the “Overlayed Kernel Density Estimates” graph look for any outliers on the “Parallel Plot”. Outliers are curves which rise significantly above or below the others.

**22)** If you find any outliers, in “Filter Outliers” click “Create Subset Experimental Design Data Set, Excluding Selected Rows”.

**23)** Close all windows open within the program except “Journal”. Click the “Results” button below “Process 2 – DataCorrelation (NameOfStudy)”.

**24)** Arrange the “3D PCA Plot” and “Correlation Heat Map” side-by side. Rotate the “Scatterplot 3D” to view grouping of the data points, similarly colored data points should be grouped near each. Double-click on a point of interest to select it in both windows.

**25)** Close all windows open within the program except “Journal”. Click the “Results” button below “Process 3 – ArraryGroupCorrelation (NameOfStudy)”.

**26)** In “Correlation Scatterplots” click the red arrow next to “Multivariate” and select “Pairwise correlations”.

**27)** Close all windows open within the program except “Journal”. Click the “Results” button below “Process 4 – ANOVA (NameOfStudy)”.

**28)** In “Volcano Plots”, Use the cursor to draw a box around approximately ten of the points in the upper right corner on the plot of your choice. This allows you to choose the most highly expressed genes. You could use this approach to begin a candidate gene analsysis (one of the other strategies in the Explorer). More on how to use volcano plots is in the Data Analysis section below.

**29)** In “Action Buttons” click “Construct Oneway Plots”.

**30)** From the menu "Genomics" select "Genomics" > "Annotation Analysis" > "Venn Diagram”.

**31)** ***a)*** In “Available Variables” select no more than five of the “Sig Indexes” denoted by red

bars and click the arrow button next to “O-1 Variables” to select these.

***b)*** Select “Sum Sig Index for Diffs” and click the arrow button next to “Frequency variable to select this.

***c)*** Do not change any of the other settings, click "OK".

*You have nowused the Journal dialog to generate a number of visual representations of your data.*

## Part 3: Data Analysis

An explanation of the most important visualizations you made between steps 20 and 31 is found below. Talk with your lab partners about how you can use these visualizations to explore your findings and see what conclusions you can draw from JMP’s statistical analyses. Make sure to record your conclusions in your notebook. Useful hints for manipulating data in JMP are also found below.

|  |  |
| --- | --- |
| Visualization | Explanation |
| Overlayed Kernel Density Estimates, and Box Plots | The Overlayed Kernel Density Estimate shows an approximation of how many genes in the transcriptome had a particular amount of expression. The X axis represents expression, and the Y axis represents density of data points. Thus the higher the line rises above a given X value the more data points in the transcriptome had that level of expression. Each different colored lines represents the data from the different root, shoot, and nodule samples (this same color coding will be conserved throughout all the visualizations in JMP, which is really handy!). The Box Plots accomplish the same thing as the Kernel Density Estimates, but using a different visual.If you don’t understand what these graphs are showing you, don’t worry; they aren’t going to help you identify genes of interest anyway. All they do is let us eliminate outliers, which you did in step 21 above. If you’ve done that, close them. |
| 3D PCA Plot and Correlational Heat Map | These two visuals show relationships in expression between the 15 sample types in our transcriptome data. Each bubble represents the specific sample (e.g. Nm, Rj, etc., see the sample type PowerPoint for more information), and each color represents the overall type of sample (e.g. Root, Nodule, etc. same color scheme as above). The PCA plot shows in 3D what the Heat Map shows in 2D; it’s just two ways of showing the same thing. We won’t get into what the axes mean, but notice in the PCA that the bubbles of the same color are grouped together in space\*. Why does this grouping make sense? (think about what the colors and bubbles mean) Like the Kernel Density Estimates, these visuals are mostly for your own edification; they won’t help you find interesting genes. Once you’ve explored them to your satisfaction, close them and move on.  \*If they aren’t, or if you have only one bubble of a particular color, have someone check your data. You may have mislabeled something in step 16 above. |
| Volcano Plots | These are some of the most important visualizations you have of your data. Each plot compares expression of genes between two samples; there is one volcano plot for each pair of samples. Each dot represents one of the genes in the transcriptome. Difference in expression is on the X-axis and statistical significance is on the Y axis. Higher is more significant, and anything above the red line is considered statistically significant.The ordering of words in the Volcano Plot’s title is important in understanding that the plot shows. Imagine a gene that is upregulated in Roots compared to Nodules. If the Volcano Plot was titled “Roots – Nodules” this gene would have a positive X value, and if the title was “Nodules – Roots” the gene would have a negative X value, because the X value is the difference (positive or negative) in expression between the two samples. Keep this in mind when you are hunting for genes. |
| Venn Diagram | The Venn Diagrams are based off the Volcano Plot data, with the numbers in each region indicating the number of statistically significant up/downregulated genes. If, say, the intersection between the “Roots – Nodules” region and the “Roots – Pre-flowering” contained the number 42, you would know that there are 42 genes which are significantly upregulated in Roots vs. both Nodules and Pre-flowering. |

**Useful Tips for Using JMP:**

* One of the nice things about JMP is that if you highlight data in any one visualization, it is simultaneously highlighted in all the rest. For example, if you click on an intersection in a Venn Diagram, all the genes which are described by that intersection will be highlighted in the Volcano Plots, making it easy to identify them.
* Use the “Window List” on the left side of the screen to navigate more easily between windows.
* When you find interesting genes, make sure to look them up in the original data spreadsheet (jmp\_exercise\_data.txt, open it in Excel to make your life easier) to view the raw expression data. JMP looks at statistical significance, not actual numbers, so make sure you appraise your genes with the actual transcriptome information.
* When using the Volcano Plots, you can zoom by right-clicking the background and using the “size” option to increase the resolution. You can also change the color, size, and add labels to the data points you have selected from this right-click menu which will make them much easier to see.
* JMP is a big program with many features. Don’t be afraid to ask questions, and take some time to play around with the visualizations. You may well discover new tips and tricks for analyzing the data yourself, and if you do tell us so we can add them to this tutorial!

Helpful Links and Additional Resources:Wikipedia’s article on volcano plots - <http://en.wikipedia.org/wiki/Volcano_plot_(statistics)>

Schizophrenia transcriptome paper with good examples of JMP analysis and figures -<http://www.plosone.org/article/fetchArticle.action?annotationId=info%3Adoi%2F10.1371%2Fannotation%2F0fcac25f-2e63-4727-9759-635f5c212f6c&articleURI=info%3Adoi%2F10.1371%2Fjournal.pone.0003625> (Carleton’s 2008 Genetics class worked with the data in this paper before it was published).

A white paper on expression analysis with JMP Genomics is available on the *Chamaecrista* Explorer. You can find out a bit more about the different visual representations of the data in the paper.

You also might be interested in taking a quick look at a short article JMP Genomics developed based on the experiences of last years Genetics students: http://www.jmp.com/software/success/carleton\_college.shtml

Carleton has an educational license for faculty and students in the Biology Department. If you would like to download a copy onto your own computer (Windows OS only), please go to Collab/Departments/Biology to find the software depot that will allow you to download and install the hardware. Any student who is enrolled in a Biology course has permission to download the software.