**Workflows in JMP-G** software version?

Raw Data File

* Open an extracted chip file in Excel
* Delete all “header information rows” except for row that starts with “Flags” “Normalize” “Autoflag” “Block” “Column” “Row” etc.
* Select All (ctrl A) – Sort by Flag (ascending), then F635median-B532
* Change the DATA for any “Bad” Flags (-50, -100) to a value of 1

*Note: This removes the false data from analysis.*

Select All (ctrl A) 🡪 Sort by BLOCK, then ROW, then COLUMN (all ascending)

* You can also change the negative values in the raw data column (F635median-B532) by 0. There is a step further on that does that but you may occasionally have small data sets that may lose rows (data) in that step. To change the negative values you can use Excel or JMP.
  + In Excel, Select All (ctrl A), Filter and go to the data column, hit the right mouse button and hit CustomAutoFilter. Select “is less than” and type in 0. The screen will show only the rows containing negative numbers. Change by 0.
  + In JMP, hit Cols (columns) 🡪 Recode 🡪 Sample 🡪 Change in place 🡪 Change manually. Do to each sample. Sort and check. Since you have to do it sample per sample, this is only recommendable with few samples.
* Select the column for “Block” and click “Sort & Filter” button on Toolbar
  + Filter “Block” tab to only contain data for block 1
* Select columns **Flags, Column, Row, Name**, **ID,** **F532 Median - B532** and **REFNumber** from each block
* Copy and Paste into a new Excel spreadsheet labeled:

“Raw Data File\_<experiment name>\_<date>”

* Rename F532 Median - B532 column (the raw data) for each block as:

**chip#\_hyb date\_block#\_sample\_variable1\_variable2\_variable3…**

(this will match your EDDS file information)

Keep in mind JMP doesn’t like:

variables that start with numbers

spaces

dots

flashdrives

long names, avoid redundance

*Note: must include* **chip** *or* **array** *as a variable!*

* Repeat for each ***block*** on the chip.
* Repeat for every ***chip*** in the experiment.
* Final Excel file should have **N** columns corresponding to each sample hybridized in the experiment and 15,744 rows of data
* Open Excel file in JMP Genomics and save as BOTH:
  + JMP Data Table (\*.jmp)
  + SAS Data Set (\*.sas7bdat)

You can work with your entire dataset including 15744 rows (features dataset) or with the probes dataset. In order to obtain the probes data set you have to:

* Delete the empty features, named as NA. Select All (ctrl A), Filter by name. Type in “NA”. Select all the rows containing NA and delete. This will take a moment and will reduce the dataset to 12890 rows. You can do this in Excel and save as an excel file.
* Open the file in JMP Genomics. Hit Table 🡪 Summary 🡪 Select your samples of interest and hit “Mean”. Group by ID. With the current microarray you will obtain 4178 rows. If the RefNumber column is missing after this step you can copy paste it from the “clean annotation file”.
* Keep in mind each file should match its own annotation file.
* Save as JMP Data Table (\*.jmp) and SAS Data Set (\*.sas7bdat)

Experimental Design Dataset File Creation (EDDS)

*Note: this generates the file with the experimental variables to test; do not begin any variable with a number, i.e. instead of 60912 for a date use d60912 or jun912 etc.*

* Genomics – Import – Experimental Design File – Create a Design Data Set from Existing Data Set
  + - Input SAS Data Set: open raw data file created above (.sas7bdat format)
    - Variables containing primary data: select all columns with data & sample heading

*Note: Do not select RefNumber.*

MUST INCLUDE: *chip* or *array* as a variable. Sometimes you will have to add a *fake array* just to have the complete word “array” as a variable, you can add it at the end of each name. The fake array has to be unique per sample. JMP believes it’s an array but you know it is not.

* + - Choose an output folder (this generates many SAS files so you should create an EDDS folder)
    - Options tab:

SAS code for EDDS to create variables (copy and paste into JMP-G field; change variable names to *your* experimental variables):

%str(

Line = scan(columnname,1,"\_");

Sex = scan(columnname,2,"\_");

Age = scan(columnname,3,"\_");

Channel = scan(columnname,4,"\_");

)

*Note: Keep number of characters in variable names to a minimum. Upper/lower case does not matter.*

* Run

Annotation File

*Note: this generates the file with all of your gene/probe information.*

* Create from .gal file associated with microarray data extraction
  + - Open .gal file in Excel
    - Add additional columns to file describing genes/probes, species, accession #, etc.
    - Make sure to have Column, Row, RefNumber, Name and ID columns from original .gal file
    - Save as Feature Annotation\_<experiment name\_date>

Once you have the annotation file most times you will only:

* + - Open with JMP Genomics
    - Save as BOTH:
      * JMP Data Table (\*.jmp)
      * SAS Data Set (\*.sas7bdat)
    - There is an updated annotation file in progress.

Sample Distribution (EDDS file) - *optional*

Shows the distribution of samples in experiment (i.e. balanced design or not)

* Open EDDS File in JMP
* Analyze → Distribution
* Select variables (not ColumnName) for “Y, Columns” field
* Run

Raw Data Distribution Analysis

*Note: Shows if there are any outlier arrays and how your raw data looks before transformation/normalization*

* Genomics → Workflow → Advanced → Expression QC Workflow or Basic Expression Workflow
  + General Tab
    - Input SAS Data Set: open raw data file \*.sas7bdat
    - From now on every time you open a file in JMP, it will pop up a window including all the files you have opened in the past. The files will be organized by name or date created or modified. Keep track of the files you need to avoid confusions.
  + Experimental Design Tab
    - Label Variable: ColumnName
    - Color: choose a treatment variable or sample label of interest, i.e. Temp, Date, etc.
    - Variance Component: Chip Array Date Site Temp (experimental & technical variables)
  + Replace Low/High Tab
    - Do not check – leave all blank
  + Delete Rows Tab
    - Do not select – leave all blank
  + Groups Tab
    - leave blank
  + QC & Normalization Tab
    - **√** Distribution Analysis
    - **√** Correlation & PCA
    - PVCA:
      * Cumm. Portion of Variation: 0.9 (90%)
      * Max Number of Principal Comp to Model: 0
      * Number of the First Principal Comp to Model: 1
      * Norm Method: None
      * Run QC Analysis: Before Norm
  + Run
* Results: Distribution
* JOURNAL → Close All Other Windows button
* Results: Data Correlation

*Note: High residual variation = unexplained variance (synonymous with error)*

* + Simplify Principle Components
    - Variance Components Charts
      * Red Triangle: Var Prop by Prin Comp → red triangle → Script → Relaunch Analysis
      * Remove all but 1st 3 or 5 principle components
      * Connect Points: Red Triangle → Y-options → Connect Points
    - JOURNAL → Close All Other Windows

*Note: If JMP Genomics crashes or you need to come back to an analysis later, open the .jsl script file automatically generated by JMP.*

Dataset File Conversions

*Note: cannot log transform negative #s or ‘0’, so need lowest # in dataset to = 1 [log2(1)=0]*

* Missing Value Imputation (to replace negative or missing numbers in dataset with ‘0’)
  + Genomics → Expression → Quality Control → Missing Value Imputation (mvi)
    - Load Raw Data File & EDDS (\*.sas7bdat)
    - Choose and Output Folder
    - Variables to Include in the Imputation Process: blank
    - Options:
      * Select “Constant”
      * Constant to Apply: 0 or 1
      * Set Negative Values to Missing: check
      * Delete Rows with at Least this Percentage Missing: set to 100%
      * Renaming output data set is unnecessary unless you want to.
  + Run
* Log2 Transform Data
  + Genomics → SAS Dataset Utilities → Columns → Transform
    - Load MVI File (\*.sas7bdat) (This is what you just generated in the previous step.)
    - Variables to be Transformed: everything you need. Do not include RefNumber.
    - List-Style Specification of Variables: blank
    - Type of Transformation: Log2
    - Shifting Factor: 1
    - Do not need to load EDDS file or use Options Tab
      * Choose Output Folder (It is recommended to make a new folder each time output is generated so each analysis doesn’t get confused).
  + Run
* Loess Normalization
  + Genomics → Expression → Normalization → Loess Model
    - Load Log2 Transformed Data File and EDDS File (\*.sas7bdat)
    - Smoothing Parameter: do not change
    - Number of LOESS Iterations: 1
    - Choose Output Folder
  + All Other Tabs: leave blank
  + Run

Other options are IQR, median, mean, Std. You should run them and find the one the describes best your data although Loess is the more common one.

File Names:

\_mvi - missing value imputation

\_dtf - log2 transformed

\_lnm - loess normalized (this will be your working data file)

\_amr - significantly different genes/probes from ANOVA

\_edds or \_exp = experimental design dataset (edds)

Quality Control (QC)  
*Note: Best to start by looking at main components (Date, Site) separately; add interactions (Date\*Site) later*

* Genomics → Workflow → Expression QC Workflow
  + General Tab
    - Study Name: Name file to indicate QC Only Loess Data
    - Input SAS Data Set: open loess data file (lnm) \*.sas7bdat
    - By Variable or Intensity columns to filter: blank
    - Label Variable: blank
    - Make sure to choose output folder – make a new folder

*Note: JMP is “buggy” and when you make a new folder, after you type in the name, click on another folder, then click back on the new folder. Look down at the bottom left corner to make sure your new folder’s name appears. Then select ok.*

* + Experimental Design Tab
    - Label Variable: ColumnName
    - Color: choose a treatment variable or sample label of interest, i.e. Temp, Date, etc.
    - Variance Component: Chip Array Date Site Temp (experimental & technical variables) Copy and paste these into the Variance Component field.
  + Replace Low/High Tab
    - Do not check – leave all blank
  + Delete Rows Tab
    - Do not select – leave all blank
  + Groups Tab
    - leave blank
  + QC & Normalization Tab
    - **√** Distribution Analysis
    - **√** Correlation & PCA
    - PVCA:
      * Cumm. Portion of Variation: 0.9 (90%)
      * Max Number of Principal Comp to Model: 0
      * Number of the First Principal Comp to Model: 1
      * Norm Method: None
      * Run QC Analysis: Before Norm
  + Run
* Results: Distribution
* JOURNAL → Close All Other Windows
* Results: Data Correlation

*Note: High residual variation = unexplained variance (synonymous with error)*

* + Simplify Principle Components
    - Variance Components Charts
      * Red Triangle: Var Prop by Prin Comp → Script → Relaunch Analysis
      * Remove all but 1st, 4 or 5 principle components
      * Connect Points: Red Triangle → Y-options → Connect Points
    - JOURNAL → Close All Other Windows
* Re-Run - - - To Test Interactions (optional)
  + Open Workflow Dialogue
    - Highlight Data Correlation in Workflow Run → Edit
    - Change Variance Components: Date Site Date\*Site Colony… etc.
  + Run

This will allow you to look at the Kernel Density and Box Plots.

ANOVA (F-test)

*Note: ANOVA has 2 parts*

1. *F-test (ANOVA part) – shows which genes are different across an effect (i.e. Date), no direct comparison between Date 1 and Date2*
2. *Clustering – not as informative to cluster all genes, all samples; better to do F-test first to find subset of interesting (or informative) genes to cluster*

* Basic Expression Workflow
  + Genomics → Workflow → Expression Statistics Workflow

**-or-**

Windows → Close All → Save None

File → Load Genomic Settings → Output Folder → Load: Basic Expression Workflow (BEW)

* + Variable to merge Annotation: RefNnumber (Agilent) or Spot (Ecogenomics/Combimatrix)
  + Input SAS Data Set: open Loess Transformed Data File (lnm) \*.sas7bdat
  + Experimental Design Tab
    - Load EDDS file \*.sas7bdat
    - Label Variable: ColumnName
    - Color: choose a treatment variable or sample label of interest, i.e. Temp, Date, etc.
    - Filter to Include: blank
  + Model Tab –or- ANOVA Tab
    - Class Variables: ALL variables to be tested, i.e. Date Site Date\*Site

*Note: can look at variables separately first to see general differences; to look at cross reactions you can type:* Date Site Date\*Site***-or-***Date|Site

* + - Fixed Effects: variables to test (MUST be in Class Variables field)

i.e. Date Site Date\*Site ***-or-*** Date|Site

* + - Random Effects: technical replicates (MUST be in Class Variables field)

i.e. Chip Array Hyb Date… etc (especially important if you have a large amount of samples)

* + - Estimate Statements: blank
  + LSMeans Tab
    - Estimate LSMeans for Fixed Effects: Date Site Date\*Site (MUST be same as Fixed Effects)
    - LSMeans Differences for Volcano Plots: None or Simple Differences
    - If Simple Differences, you can check Cluster Significant LSMean Profiles
  + Multiple Testing Tab

*Note: If doing differences, “simple differences” should be used; p-value adjustment is on F-test. ‘Difference Chooser’ – allows you to select which differences or comparisons you want to make and in what direction. In Difference Chooser’, if you hit “Reverse” without selecting ‘Include’ it will automatically select ‘Include’ too. It creates a folder automatically.*

* + - Multiple Testing Method: FDR or Bonferroni

-OR-

* + - Choose “blank” and set –log10(p-value) Cutoff to desired level (usually 2 – 3)

-log10(p-value) = the significance value, i.e. if p = 0.001 then –log10(0.005) = 2.3 (the higher the number the lower the p value)

* + Clustering Tab (optional)
    - Hierarchical Cluster: check if doing “Simple Differences” (do not change other parameters). Sometimes it will not show it. To obtain it follow the steps in page 10.
  + PCA Tab (optional)
    - Principal Component: check if doing “Simple Differences” (do not change other parameters)
  + Annotation
    - Load Annotation File \*.sas7bdat
    - Annotation Merge Variables: RefNumber (Agilent), Spot (Ecogenomics)
    - *Note: This step is extremely important as it merges the Annotation file with your Data file.*Annotation Label Variable: Gene Name or Gene Identifier or Probe
    - GenBank Accession Variable: Accession Number or Probe
    - Gene Symbol: name, description, function, etc.
    - Gene Description: name, description, function, etc.

*Note: other than “Annotation Merge Variable” these fields are just identifiers for output data to identify which genes/probes are significantly different. Can fill out as desired.*

* + Run
* Analysis of Results
  + Workflow Journal: Select Results (opens new results window)
  + ANOVA Results Window
    - Tabs → Results: View Data
    - Table → Tabulate → Scroll to bottom → highlight ALL “Sig Index for Diff”
    - Drag to ‘Drop Zone for Rows’ → Add Columns by Categories
      * Shows how many genes are sig (1) vs. non-sig (0) in each effect (uses -log(p))  
        Example:

|  |  |  |
| --- | --- | --- |
|  | **0** | **1** |
| Date | 315 | 427 |
| Site | 350 | 392 |
| Date\*Site | 417 | 325 |

* + Venn Diagram – *how common are sig genes* ***between*** *effects (how much overlap)*
    - Make sure ANOVA \_amr file is open and selected
    - Genomics → General Utilities → Venn Diagram (single table)
    - Select ≤ 5 “Sig Index for Diff” (*or Sig Index PrF\_*) and include in “0-1 Variables” field. Hit ok.
  + Refine Analysis to Look at Individual Genes Across All Samples
    - Select portion of Venn Diagram of interest (i.e. overlapping genes between 2+ comparisons)

*Note: this highlights corresponding rows of significant genes on the \_amr table*

* + - ANOVA Results Window → Action Buttons → Launch Plot Intensities
      * Intensity Columns to Plot: blank
      * Label Variables: blank
      * Groups Tab: Make sure EDDS file is loaded (Variables Defining Groups - blank)
      * Annotation: not necessary, *can remove but do not have to*
  + Run
  + Tool: animates intensity plots
    - Rows → Data Filter → Spot (Ecogenomics/Combimatrix) or RefNumber or ID (Agilent)

*Note: make sure “spot” is in table & is ‘Nominal’ <not Continuous>; can then indicate “\_subt\_pi”*

* + - Highlight a spot → Select “Show”
    - Red Triangle → Animation → Play

Post-Hoc Testing – to determine in which effects genes are significant

* Tukey-Kramer
  + Select subset of significant genes (or all genes) in \_amr file
    - Tables → Stack
    - Stacked Columns: Highlight all LsMean
    - Non-stacked columns: check “select”
    - Select a few descriptors, i.e. RefNumber, ID… Hit ok. Save this as a separate file!
    - Help → Statistics Index → Statistics → Tukey-Kramer HSD → Launch
    - Tukey-Kramer HSD (Oneway - Distribution by Group) → Launch
      * Y, Response → Select Data
      * X, Grouping → ID, Name, Sample Name (labe), etc. Pick one.
      * By → Functional Group, Species, Probe Group, etc. You hit ok and it looks like nothing happened. Be patient. Save the results since JMP doesn’t save them automatically.
    - Red Triangle → Compare Means → All Pairs, Tukey HSD

Hierarchical Clustering – *2 ways to do it*

*Recommended:*

* Method 1: JMP – Select subset of interesting genes from \_amr
  + Example 1:
    - Rows → Data Filter → Highlight “Sig Index for Diff” for ONE comparison → Add
    - Highlight 0 in “0 ≤ Sig Index…” field and change to 1
    - Select **:** button at bottom of window to add additional ‘or’ comparisons → change 0 to 1
    - Repeat for all comparisons to cluster
    - Table → Subset
  + Example 2:
    - Highlight significant genes or any genes of interest (i.e. see Venn Diagram Example)
    - Table → Subset
  + Cluster with JMP
    - Analyze → Multivariate Methods → Cluster
    - Y, Columns: LSMeans for effect (i.e. Date) – *Note: Keep standardize data checked.*
    - Red Triangle → Select 2-way clustering

Not recommended:

* Method 2: JMP-G – Hierarchical Clustering (more detailed cluster analysis)

*Note: Memory intensive! May cause screen to freeze.*

* + Highlight significant genes in \_amr file for effect of interest (see method 1). You can also use the \_sig file
  + Action Button (associated with ANOVA)
    - Open subset in Wide Format
      * Select “Spot” as common prefix for wide column names (= sp\_)
    - JMP-G → Pattern Discovery → Hierarchical Clustering – *Note: can run PCA, etc. as well*
      * Open Wide Format file just created
      * List style specification of rows to be clustered = sp\_: (or just select “spot”)
      * Compare Variables: Date Site Colony… etc
      * Label Variables: ColumnName
      * Options:
        + √ standardize variables
        + X center rows (leave unchecked)
        + √ create heat map for compare var…
      * “Fast Ward” or “Avg” are usually used
  + Easy Tool: to visualize genes in heat map associated with effect
    - Analyze → Distribution → Select effect (Date, Site, etc.)

*Note: this highlights genes in heat map associated with the selected effect*

K-Means Clustering

* JMP-G → Pattern Discovery → K-Means Clusters
  + Select table of significant genes produced by ANOVA -or- subset of it  
    *Note: Input file must be in “rectangular format and contain the variables (columns) whose observations (rows) are to be clustered”.*
  + Label Variable → genes, probes, spots… (rows)
  + Variables to Cluster → Sample or Effect
  + Select → Automated Radius K Means
    - Correlation Radius for Clustering: high value = ↑ number clusters; lower value = ↓ number clusters   
      *Note: not statistical, but validates (via correlation) why there are n=X clusters*
    - Run
    - Tool: Color rows by variable (i.e. Functional Group)
      * Rows → “Color or Mark by Column” → select variable & color grouping

Multidimensional Scaling – Distance Matrix  
*Note: Produces a Distance Matrix and runs a SAS Clustering on the matrix which gives distances between clusters. Similar to PCA, shows different dimension & how much dimensionality is taken care of in each dimension*

* Windows → Close All
* Genomics → Pattern Discovery → Distance Matrix
  + Input Data Set → use Wide Format data table to cluster effects (not genes); i.e. see Hierarchical Clustering, Method 2 (above); example = wide\_ftest\_date
  + Variables = “spot” (sp\_1… sp\_x )
    - ID Variable = ColumnName
    - Copy Variable = Effect (i.e. Date Site … etc.)
  + Analysis
    - Level of Measurement = Interval
    - Distance Matrix = Euclid
    - Standardization Method = STD
  + Clustering → Select “Perform Clustering on Distance Matrix”
  + JMP → Analyze → Distribution
    - Select an effect (i.e. Date) to see where clusters are occurring in matrix
  + Table → Subset – to look at what samples are in a cluster
* To Reduce Dimensionality Further
  + “Launch AP Using Output Data Set” (in windows left screen) = Multidimensional Scaling
    - Color Variable: select an effect (i.e. Date)
    - Label Variable: ColumnName
  + Run
  + Results: Overlay Plot
    - Select one of the dimension points in the plot, usually want one close to or at the inflection point; *i.e. in Miami data, the 2nd dimension is already up to 96% of dimensionality*
    - Display Cluster Heat Map

Re-Running Analyses of Specific Effects

* Windows → Close All
* File → Load Genomic Settings → Select Basic Expression Workflow for ANOVA (see above)
  + Change study name to reflect current ‘effect’ comparison (i.e. Date Only)
  + LSMeans → Simple Differences → Difference Chooser
    - Select which differences to compare

*Example: compare those with greatest effect (i.e. J05 – O05, F05 – J06)*

* + - Save selected differences to output data set
  + Cluster Sig LSMeans Profiles
* Run
* Tables → Tabulate → “Sig Index for Diff” → Drop Zone for Rows → Add Columns by Categories (0 = not sig; 1 = sig)
* Venn Diagram (see above)
  + Genomics → Annotation Analysis → Venn Diagram → Select “Sig Index for Diff” of interest