



JMP Genomics

Version 4.1

Release Notes

"Creativity involves breaking out of established patterns in order to look at things in a different way." Edward de Bono



JMP. A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513
www.jmp.com

Release Notes for JMP Genomics 4.1

Copyright ©2009, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

JMP[®], SAS[®] and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

This document describes changes and enhancements from JMP Genomics 4.0 to JMP Genomics 4.1. New and improved features in JMP Genomics Analytical Processes (APs) are described in the following sections. Changes to specific analytical processes are organized according to the JMP Genomics main menu.

General Features

SAS and JMP

- The SAS and JMP Components underlying JMP Genomics 4.1 have been updated to SAS 9.2M2 and JMP 8.0.2.

Client-Server Functionality *New!*

- The APs in JMP Genomics 4.1 can be run in an experimental client-server mode by connecting to an existing SAS metadata server. While connected to the metadata server, the APs in a local desktop installation of JMP Genomics use the server to run SAS, which can be particularly helpful in processing large jobs.
- Please note that a server installation used in this way must have a metadata server running, and must have installed, at minimum, SAS Integration Technologies as well as the SAS components found within desktop JMP Genomics (Base, Stat, Graph, Genetics, IML and Access to PC File Formats). No JMP Genomics-specific installation is required on the server.

SAS Message Windows

- Many of the JMP Genomics APs surface the names and paths to their output data sets in a SAS Message window. The identity and function of each output data set is now identified in the SAS Message window.

Tip of the Day *New!*

- A randomly-selected JMP Genomics *Tip of the Day* is surfaced when JMP Genomics is opened.
- Tips may be accessed within JMP by selecting **Help > JMPG Tip of the Day**.

Accessing the JMP Genomics User Guide

- Individual manuals in the JMP Genomics User Guide may be accessed by selecting **Genomics > Documentation and Help > User Guide > Name of Specific Manual**.
- You may access individual manuals directly by clicking the **User Guide** button located at bottom of each JMP Genomics dialog.

Configure Genomics Settings

- A new option for archiving resolved SAS macro code has been added. *Note:* This option is especially useful for SAS programmers because it allows you to see the actual code that is run in one file. *Caution:* Selecting this option can substantially increase run times for jobs that make many macro calls.

Load Genomics Setting *New!*

- A new option for selecting a saved setting and opening the corresponding AP has been added to the File menu.
- Click **File > Load Genomics Setting** to access this option.

All JMP Genomics dialogs

- You can now type or paste the names and paths of input data sets and input/output folders directly into their respective fields.
- A link to the *JMP Genomics User Guide* has been added to the bottom of every dialog. Click this button to open the volume containing information specific to the process you are running.

Data Import and Manipulation

Import

Affymetrix Exon and Whole Transcript Expression CEL Input Engine

- A column that lists the number of probe level intensities summarized to generate the probeset-level value has been added to the output data set.
- An option for specifying the minimum number of probes used to generate the probeset-level value has been added. Summaries not meeting this minimum are not reported in the output data set.
- The ability to select an existing baseline reference data set and a reference EDDS for normalizing arrays against accepted standards, has been added.

Affymetrix Expression CEL Input Engine

- The ability to normalize new arrays to an existing baseline reference data set and a reference EDDS, is now available.

Affymetrix miRNA CEL Input Engine *New!*

- This new process is specifically designed to import a set of Affymetrix miRNA .CEL files and combine them into a single SAS data set.
 - This process also allows users to filter probes during import.
- Please note that although this process is similar to the Affymetrix Expression CEL Input Engine, default settings are optimized for miRNA arrays.

Affymetrix SNP CEL Input Engine

- A field for specifying an MPS file has been added to the dialog. This file, which is required for importing Affymetrix SNP6 files, contains information used to group copy number (CN_) probesets into larger copy number variant regions identified from previous studies. Intensities for individual CN_ probesets are summarized to generate overall intensities for each variant region. A CN_ probeset may be summarized into more than one copy number variant region, since different studies may have identified similar but not identical overlapping variant regions. *Note:* Specific SNP6 MPS files must be downloaded from the Affymetrix NetAffx web site. This can be done within JMP Genomics using the Download NetAffx Files IE (**Genomics > Import > Affymetrix > Download NetAffx Files**).
- An option for removing Affymetrix control SNPs from the output data set has been added. *Note:* control SNPs are generally identified using an AFFX_ prefix.
- The ability to normalize new arrays to an existing baseline reference data set and a reference EDDS, is now available.
- The ability to base the normalization on autosomal data only has been added. If this option is selected, a two-step normalization procedure is applied. First, the observations from non-autosomes (by default, X and Y) are separated from autosomes. The normalization is then applied only to data for autosomes. Normalized data from the nearest smaller and nearest larger values from autosomes are used to interpolate a normalized value for each observation from a non-autosome.

Affymetrix Cytogenetics CEL Input Engine *New!*

- This new input engine imports a set of .CEL files generated from Cytogenetics 2.7M CEL chip files and combines them into a single SAS data set.

Affymetrix Tiling CEL Input Engine *New!*

- This new input engine is specifically designed to import tiling data from a set of Affymetrix .CEL files to generate an output data set listing intensities, an output EDDS, and an output data set listing data from control probes.

Affymetrix Exon CHP Input Engine *New!*

- This new process is specifically tailored for import of exon .CHP files.
- Please note that although this process is similar to the Affymetrix Expression CHP Input Engine, default settings are optimized for exon arrays.

Affymetrix Tiling Bar Input Engine *New!*

- This new input engine imports and combines normalized intensities from a set of Affymetrix BAR files, which are output by the Affymetrix Tiling Array software (TAS), into a single SAS data set

Illumina miRNA Input Engine *New!*

- This new *experimental* process is specifically tailored for import of a text (.txt) file containing data from a Sample Gene Profile table in Illumina's BeadStudio 3.1 or 3.2 Gene Expression modules into a SAS data set.
- This process also creates a corresponding experimental design data set and annotation data set.

Agilent Input Engine

- A ProbelD column listing the Probe Index value for each row has been added to the output data set.

Arraytrack Input Engine

- A ProbelD column listing the Probe Index value for each row has been added to the output data set.

GenePix Input Engine

- A ProbelD column listing the Probe Index value for each row has been added to the output data set.

Quantarray Input Engine

- A ProbelD column listing the Probe Index value for each row has been added to the output data set.

Scanalyze Input Engine

- A ProbelD column listing the Probe Index value for each row has been added to the output data set.

Imputed SNP (Tall Format) Input Engine *New!*

- This new process imports a set of *tall* text files generated by a SNP imputation program (IMPUTE or BEAGLE, for example).
- Two genotype data sets are generated: a wide data set containing the most likely genotypes that meet or exceed a specified threshold and a stacked data set with genotype probabilities.

Imputed SNP (Wide Format) Input Engine *New!*

- This new process imports a set of *wide* text files generated by a SNP imputation program (MACH, for example).
- Two genotype data sets are generated: a wide data set containing the most likely genotypes that meet or exceed a specified threshold and a stacked data set with genotype probabilities.

Imputed SNP Import Tutorial *New!*

- This interactive JSL program guides you through selecting an imputed SNP IE and specifying parameters for importing files.

Data Set Utilities

Sort Rows

- An option to remove rows with duplicate By Variables has been added.

Workflows

Basic Genetics Workflow

- An option to specify trait values for filtering individuals for inclusion in Hardy-Weinberg Equilibrium tests has been added. For example, users may specify a trait value corresponding to control individuals here to perform HWE tests only on control samples.

Basic Expression Workflow

- An option for specifying a SAS data set containing selected LSMean fixed effects differences has been added. You may either generate this data set beforehand, using the Difference Chooser AP, or click **Difference Chooser** to launch the Difference Chooser AP preloaded with the specified ANOVA parameters.
- An option for specifying the cumulative proportion of variance to be used in the principal components analysis has been added to the QC and Normalization tab.

Basic miRNA Workflow *New!*

- A basic workflow used to import miRNA and perform standard quality control, normalization, and analysis, on miRNA data.

Basic Exon Workflow

- The ability to journal results by chromosomes has been added.
- The number of output plots surfaced when you run this AP has been reduced. Plots not surfaced are still accessible using drill-down options.
- An option for specifying the cumulative proportion of variance to be used in the principal components analysis has been added to the QC and Normalization tab.

Basic Tiling Workflow *New!*

- A basic workflow used to import miRNA and perform standard quality control, normalization, and analysis, on data generated from tiling arrays.

Expression QC Workflow

- An option for specifying variance component parameters has been added.
- An option for specifying the cumulative proportion of variance to be used in the principal components analysis has been added to the QC and Normalization tab.

Expression Statistics Workflow

- An option for specifying a SAS data set containing selected LSMeans fixed effects differences has been added. You may either generate this data set beforehand, using the Difference Chooser AP, or click **Difference Chooser** to launch the Difference Chooser AP preloaded with the specified ANOVA parameters.
- Options for centering and scaling columns and rows for principal components analysis have been added to the Clustering tab.

Genetics

Genetic Data Set Utilities

Check Data Contents

- Options for specifying the first data column and row have been added.

Relationship Matrix *New!*

- This new AP computes a symmetric matrix of pair-wise relatedness measures for rows of the input data set across all SNP loci.
- Three different genome-wide relatedness estimates (Identity by Descent, Identity by State, and Allele Sharing Similarity) can be calculated.
- The output data set can be used in Q-K Mixed Model (see below) or Multidimensional Scaling.

Genetic Marker Statistics

Population Measures *New!*

- The AP represents a significant repurposing of the Population Distance Matrix AP previously found in the Genetic Data Set Utilities submenu in JMP Genomics 4.0.
- In addition to generating a dissimilarity matrix, this AP produces tables listing *F*-statistics both for the individual markers as well as overall for the population.

Marker Properties

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.
- An option to specify trait values for filtering individuals for inclusion in Hardy-Weinberg Equilibrium tests has been added. For example, users may specify a trait value corresponding to control individuals here to perform HWE tests only on control samples.

Missing Genotype by Trait Summary

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Linkage Disequilibrium

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.

LD TagSNP Selection

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.

Association Testing

Case-Control Association

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

PCA for Population Stratification

- A new **Marker Name** field has been added to the **Annotation** tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the **Subset and Reorder Genetic Data AP**.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.
- An option for merging Input data and PCA results into the output data set has been added.

Marker-Trait Association

- A new **Marker Name** field has been added to the **Annotation** tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the **Subset and Reorder Genetic Data AP**.
- Both class fixed effects and interaction effects can be specified for survival traits.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.
- A new option has been added that allows you to select the method used to calculate denominator degrees of freedom when a random effect is specified.

SNP-Trait Association

- A field for specifying an annotation variable listing the name of each column from the input data set has been added to the **Annotation** tab. The values of this variable are compared to the column labels to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the **Subset and Reorder Genetic Data AP**.
- Interaction effects are now allowed in the model for the trend test. Note that odds ratios are not calculated for the interaction effects.
- Both class fixed effects and interaction effects can be specified for survival traits.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.
- A new option has been added that allows you to select the method used to calculate denominator degrees of freedom when a random effect is specified.

Imputed SNP-Trait Association *New!*

- This new process tests for association between various types of traits and each individual SNP taking into account the probabilities of each possible genotype.
- Two types of analyses can be performed: a general test based on the probabilities of each SNP genotype or a regression testing for a linear trend of SNP alleles.
- Adjustments can be made for quantitative covariates and random effects or for some trait types, strata variables.
- *P*-values from these tests, with adjustments applied if requested, are plotted along the marker map.

Survey SNP-Trait Association

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Interaction effects are now allowed in the model for the trend test. Note that odds ratios are not calculated for the interaction effects.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Q-K Mixed Model *New!*

- This new process tests for association between various types of traits and SNP genotypes or alleles from a single SNP at a time while adjusting simultaneously for population structure and family relatedness.
- Two types of analyses can be performed: an ANOVA based on SNP genotypes or a regression testing for a linear trend of SNP alleles.
- *P*-values from these tests, with adjustments applied if requested, are plotted along the marker map.

Quantitative TDT

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

TDT

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Model-free Linkage

Affected Sib-pair Tests

- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Haseman-Elston Regression

- Specifying an annotation group variable with more than one value generates a Manhattan plot.
- *Note:* This AP requires an IBD matrix be specified as input. This matrix, and others, can now be computed using the Relationship Matrix process.

Variance Components

- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Haplotype Analysis

Haplotype Estimation

- A new Marker Name field has been added to the Annotation tab. Use this field to specify the variable in the annotation data set that lists columns of markers from the input data set. This column is used to verify that the order of the columns in the input data set matches the order of the rows in the annotation data set. If mismatches are found, you are prompted to run the Subset and Reorder Genetic Data AP.
- Specifying an annotation group variable with more than one value generates a Manhattan plot.

Expression Analysis

Quality Control

Correlation and Principal Components

- The ability to automatically select the number of principal components to include in variance components analysis, based on a specified cumulative proportion of variance explained, has been added.

Normalization

Loess Normalization

- The ability to normalize new arrays to an existing baseline reference data set and a reference EDDS, is now available.

Quantile Normalization

- The ability to normalize new arrays to an existing baseline reference data set and a reference EDDS, is now available.
- The ability to base the normalization on autosomal data only has been added. If this option is selected, a two-step normalization procedure is applied. First, the observations from non-autosomes (by default, X and Y) are separated from autosomes. The normalization is then applied only to data for autosomes. Normalized data from the nearest smaller and nearest larger values from autosomes are used to interpolate a normalized value for each observation from a non-autosome.

Pattern Discovery

Plot Intensities *New!*

- This new AP allows you to visualize row-level intensity measurements for individual samples or groups of samples via interactive parallel plots.
- This AP may be either be run *de novo* on a tall data set or may be launched from the ANOVA action button window, generated by other APs, to examine a selected subset of interesting results.

Cross Correlation

- A second annotation tab has been added. This tab allows you to specify an annotation data set containing information on the secondary variables.

Row-by-Row Modeling

ANOVA

- Several new multiple testing methods for p-value adjustment, including Adaptive Holm, Adaptive Hochberg, Adaptive FDR, Dependent FDR, have been added.
- An option for specifying a SAS data set containing selected LSMeans fixed effects differences has been added. You may either generate this data set beforehand, using the Difference Chooser AP, or click **Difference Chooser** to launch the Difference Chooser AP preloaded with the specified ANOVA parameters.

Mixed Model Analysis

- An option for specifying a SAS data set containing selected LSMeans fixed effects differences has been added. You may either generate this data set beforehand, using the Difference Chooser AP, or click **Difference Chooser** to launch the Difference Chooser AP preloaded with the specified ANOVA parameters.

Difference Chooser *New!*

- This analytical process allows you to select (and, optionally reverse the order of) the LSMean fixed effect level differences to be included in an ANOVA or Mixed Model analysis. All differences that are included and reversed are saved to a SAS Data Set which you can then input to the ANOVA or Mixed Model APs.
- The Difference Chooser AP may be launched from the Row By Row Modeling submenu, or launched pre-loaded with specified parameters by clicking **Difference Chooser** on the LSMEANS tab found on select APs/workflows.

P-Value Browser

- Users may specify chromosome and position information to generate an overall view of the chromosomes overlaid with p -value significance indicators. Interesting genomic areas may be selected using available zoom and drill-down capabilities to display interactive JMP p -value plots with gene and SNP tracks overlaid.
- Chromosomes may be colored with a custom color theme. Sample settings using cytoband patterns for human, mouse and rat, and other examples of custom color themes are available. If cytoband or other high-level grouping information is not available, genomic bins may be created using position or by summarizing over a specified number of rows, and colored by average $-\log_{10}$ (p -value) of the binned measurements.
- Optionally, when only a single chromosome is specified, a genome may be displayed as circular.
- An option for generating separate overlay plots for each p -value variable has been added.

Predictive Modeling

All Predictive Modeling APs

- Several new multiple testing methods for p -value adjustment, including Adaptive Holm, Adaptive Hochberg, Adaptive FDR, Dependent FDR, have been added.
- When predicting a binary trait, Receiver Operating Characteristic (ROC) curves and statistics have been added.

Survival Predictive Modeling *New!*

- This new process applies a Cox proportional hazards model on survival data with time-to-event variable and optional censor indicator to estimate survival functions and the corresponding median survival time for each row in the input data set.
- A variety of model selection methods are available, including forward, backward, and stepwise.

Cross Validation Model Comparison

- When cross-validating a binary trait, a drill-down button for displaying Receiver Operator Characteristic (ROC) curves has been added.
- The reliability diagrams have been consolidated into one window to allow for easier comparisons.

Annotation

Venn Diagram

- Additional controls have been added to the Venn diagrams. Use these controls to adjust the position, colors, and other attributes of the circles, labels and counts.

Column Enrichment

- Several new multiple testing methods for p-value adjustment, including Adaptive Holm, Adaptive Hochberg, Adaptive FDR, Dependent FDR, FDR Bootstrap, and FDR Permutation, have been added.

Track Gene Text *New!*

- This new process creates a settings (.sas) file that defines a gene display track which can be added to JMP Genomics statistical results like plots of $-\log_{10}(p\text{-values})$ along chromosomes. You can download text files containing track information from sources like the Table Browser of the UCSC Genome Browser for use as input to this process.
- The resulting settings file can then be selected as a track file in any AP containing a Tracks tab (P-Value Browser, for example) for embellishing graphics with depictions of genes.

Track Gene Web *New!*

- This new process creates a settings (.sas) file that defines a gene display track which can be added to JMP Genomics statistical results like plots of $-\log_{10}(p\text{-values})$ along chromosomes.
- Track information within regions selected for drill-down is pulled dynamically from UCSC for each query. *Note:* This AP requires an active internet connection.
- The resulting settings file can then be selected as a track file in any AP containing a Tracks tab (P-Value Browser, for example) for embellishing graphics with depictions of genes.

Track Gene GFF *New!*

- This new process creates a settings (.sas) file, using input from a .GFF file containing genomic information, that defines a gene display track which can be added to JMP Genomics statistical results like plots of $-\log_{10}(p\text{-values})$ along chromosomes. .GFF files are available for many organisms from a variety of public data bases.
- The resulting settings file can then be selected as a track file in any AP containing a Tracks tab (P-Value Browser, for example) for embellishing graphics with depictions of genes.

Track SNP Web *New!*

- This new process creates a settings (.sas) file that defines a display SNP track which can be added to JMP Genomics statistical results like plots of $-\log_{10}(p\text{-values})$ along chromosomes.
- Track information within regions selected for drill-down is pulled dynamically from UCSC for each query. *Note:* This AP requires an active internet connection.
- The resulting settings file can then be selected as a track file in any AP containing a Tracks tab (P-Value Browser, for example) for embellishing graphics with depictions of SNPs.

Chromosome Color Theme *New!*

- This new process creates a settings (.sas) file from a text file that defines a Chromosome Color Theme that can be used for display of the Chromosome Color Plot in the P-Value Browser AP.
- The output settings file defining the input Chromosome Text Data Set and the Color Theme is saved as a .sas file for subsequent use.

Create Weblink *New!*

- This new process creates an annotation web link report with links to various Bioinformatics databases such as GenBank, UniGene, Entrez Gene, Gene Ontology, PubMed, KEGG, Affymetrix, and more.

KEGG Get Gene Identifiers *New!*

- This new process retrieves KEGG-specific identifiers for all genes in the input data set by mapping them to identifiers in the input data set from one of the following databases: GenBank, NCBI GI, NCBI GeneID, OMIM, UniGene, UniProt.
- A new column is added to the output data set containing the matching KEGG identifiers. If multiple KEGG identifiers are found for one gene identifier, then duplicate rows are added to the output table for each unique combination.
- A list of those gene identifiers not returning any KEGG identifiers is also shown.
- *Caution:* Performance of this AP is affected by your internet connection and internet traffic. Queries requesting the return of thousands of KEGG identifiers may take several hours to complete.

KEGG Get Gene Pathways *New!*

- This new process retrieves all of the pathway identifiers for all of the genes in the input data set and creates a new column containing them.
- The new column includes the delimiter " / " to separate the pathway identifiers and is suitable for use with the Column Enrichment AP or general searching.

KEGG Color Pathways *New!*

- This new process opens an .html page for each KEGG pathway you specify and colors the genes on each pathway page according to numeric experimental variables.