

JMP Genomics

Version 5.1

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

JMP Genomics

Copyright © 2011, 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.

JMP Genomics, Version 5.1 - Release Notes

This document describes changes and enhancements from JMP Genomics, Version 5.0 to JMP Genomics, Version 5.1. Processes are described in the order in which they first appear in the JMP Genomics menu¹.

General Features

JMP and SAS Platform Updates

- JMP Genomics 5.1 is built on the latest JMP maintenance release, JMP 9.0.2. For more information about the updates to JMP software that are included in this release, please see the [JMP 9.0.2 Release Notes](#).
- JMP Genomics 5.1 is built on the latest SAS release, SAS 9.3. For more information about the enhancements to SAS analytical software that are included in this release, please see the [SAS 9.3 Release Notes](#).
- Full installation instructions will be available on the JMP Web site. Please note the following when you install JMP Genomics 5.1:
 - You must uninstall JMP Genomics 5.0 and all SAS 9.2 components first using either the **Add or Remove Programs** (Windows 7) or the **Uninstall a program** utility from the Control Panel.
 - If you have already installed SAS 9.3 on your computer before installing JMP Genomics, you might need to run the install twice. Initially, SAS will be updated, and when that installation has completed, you can re-start the installer again to install JMP Genomics.
 - After installing, you will not see a desktop icon for JMP Genomics 5.1. If you want to create one, you can do so by browsing to the location of the JMP executable (typically

1. **Note:** If you have a suggestion, comment, or encounter a bug in JMP Genomics 5.0, please click Send a Comment or a Feature Request under Genomics > Documentation and Help > Additional Resources, or e-mail details to genomics@jmp.com. For bugs, it is especially helpful if you can attach a settings file for the JMP Genomics process in which you encountered the problem, along with a subset of your data that can be used to reproduce the error. If you cannot share a subset of your own data, but can reproduce the problem with one of our sample data sets, please send us a settings file for this so that we can replicate the error. We will make every effort to address the issue promptly. Thank you for taking the time to do this!

C:\Program Files\SASHome\JMP\9) and creating a new shortcut to this file on your desktop. To change the icon used for the shortcut on Windows 7, right-click on the shortcut and choose **Properties**. Click **Change Icon** and browse to the JMP Genomics icon in C:\Program Files\SASHome\JMP\9\LifeSciences\Documentation\Icons\JMPGenomicsApp.ico.

Viewing Results Generated Using Prior Versions of JMP Genomics in JMP Genomics 5.1²

- A script is available from [JMP Technical Support](#) to modify results scripts and drill-down buttons created in earlier versions of JMP Genomics so that they can be run in JMP Genomics 5.1.

Online Documentation *New!*

- The JMP Genomics User Guide and Help System has been extensively redesigned.
- Click the **User Guide** button, located at the bottom of every dialog, to open an online User Guide that provides details about every individual process, parameter and type of output available in JMP Genomics.
- Click the **?** button, located next to each parameter in the dialogs, to open an HTML page containing comprehensive information about what the parameter does, how to specify the parameter, and the types and functions of all available options.

Tabbed Reports

- When closing a tabbed report, a Close Results window is now displayed asking if you want to close all associated graphics and results tables. This simplifies the process of closing tables that are hidden by default when tabbed reports are opened. Please note that if tables are modified and not saved, clicking this button will close them without saving your changes. If you want to save changes, you must first either click **File > Save** or click the **Save** icon (📁) on the taskbar. A new preference is available under **File > Configure Life Science Settings** that will suppress this prompt and close all associated graphs and tables.
- A new **Create Report** button, located at the bottom left corner of most tabbed reports, enables you to create either an RTF or a PDF report of all figures in the tabbed report.

2. Unfortunately, due to the extensive changes to the platform, we cannot guarantee that all results scripts generated in previous versions run correctly in JMP Genomics 5.1. Please contact technical support if you need assistance converting a script.

Import and Experimental Design

Getting Started Wizard

- The Wizard now accommodates newly supported summary file types from high throughput sequencing experiments. It launches the appropriate import process for creating an analysis-ready data set.
- The Wizard now differentiates between regular expression CEL and exon/whole transcript CEL files.

Create Design File Template

- A new Save As SAS Data Set check box option is available on the Options tab. This option is unchecked by default. When it is checked, the newly created EDF is saved as a SAS data set and closed.

Affymetrix Cytogenetics CHP (CYCHP) Input Engine

- An auto-launch button that loads the imported data into the JMP Genomics Copy Number Partition process (for partitioning of intensities or copy number calls) is now available in the output tabbed report for this input engine.

Affymetrix Expression CEL Input Engine

- An option for specifying a key variable for merging reference set information with input data when performing Loess normalization during import of CEL files is now available under the Advanced section of the Normalization tab.

Affymetrix SNP Cel Input Engine

- An option for specifying a key variable for merging reference set information with input data when performing Loess normalization during import of CEL files is now available under the Advanced section of the Normalization tab.

Affymetrix Cytogenetics CEL Input Engine

- An option for specifying a key variable for merging reference set information with input data when performing Loess normalization during import of CEL files is now available under the Advanced section of the Normalization tab.

Affymetrix Tiling CEL Input Engine

- This process now supports Calvin-formatted tiling CEL files.
- An option for specifying a key variable for merging reference set information with input data when performing Loess normalization during import of CEL files is now available under the Advanced section of the Normalization tab.

Agilent Input Engine

- When working with single color arrays, you are no longer required to specify an intensity column in your experimental design file. In this case, the gProcessedSignal column is used by default.

Illumina Expression Input Engine

- Multiple expression and miRNA Final Report files from Illumina BeadStudio or GenomeStudio and their associated sample files can now be imported and combined. When multiple files are specified, files are merged and a combined data set and combined design file are generated.

Illumina miRNA Input Engine

- Multiple expression and miRNA Final Report files from Illumina BeadStudio or GenomeStudio and their associated sample files can now be imported and combined. When multiple files are specified, files are merged and a combined data set and combined design file are generated.

Illumina SNP Input Engine

- Multiple SNP Final Report or Full Data tables can now be imported simultaneously using the same map file and combined into a single SAS data set.

Imputed SNP (Tall Format) Input Engine

- This process now creates an additional output data set, the expected genotype output data set, which can be given a custom name on the Options tab. The major and minor alleles for each SNP are determined, and the expected genotype is calculated as $2 * P(BB) + P(AB)$ where **B** is the minor allele based on the probabilities in the input file. This output data set is in wide format, and can be used as input to other association testing processes that take numeric genotypes as input, e.g. when running trend tests.

Imputed SNP (Wide Format) Input Engine

- This process now creates an additional output data set, the expected genotype output data set, which can be given a custom name on the Options tab. The major and minor alleles for each SNP are determined, and the expected genotype is calculated as $2 * P(BB) + P(AB)$ where **B** is the minor allele based on the probabilities in the input file. This output data set is in wide format, and can be used as input to other association testing processes that take numeric genotypes as input, e.g. when running trend tests.

Generate Counts from SAS (SAM Input Engine)

- This process imports aligned reads in SAM format, parsing CIGAR strings within the specified .sam file(s) to generate output SAS data sets.
- Counts can be generated at every genomic position. Alternatively, binning criteria can be specified to summarize counts either within equally spaced bins, or within exon or genes if an existing gene model in UCSC, text, or BED file format is provided that provides positional

information for genomic features. In addition to raw counts, RPM and RPKM values are now also output for bins, exons, or gene features.

- To bin read counts using known gene models on the Annotation tab, you must download a text file from the UCSC table browser or create a gene model table in UCSC format. To download an existing file for use in this process, or to use as a template to create your own UCSC-formatted file, click on the **Table Browser** link on the [UCSC home page](#). Select the genome build that corresponds to your genome of interest, and choose an available isoform table for that genome, and save the resulting output as text.
- PCR duplicates are assumed to be aligned reads from the same chromosome that start at the same position. It has been suggested that failing to remove PCR duplicate reads might lead to bias when coverage is low or moderate. A new check box has been added to the Options tab to address this problem. Check this option to remove PCR duplicates. **Note:** Selecting this option might not be appropriate for your experiment, so we encourage you to research the issue before selecting it. You should not select this option when performing high-coverage targeted resequencing, where duplicate reads might be expected to occur that are not the result of PCR duplication.
- A button to auto-launch the **Gene Model Summary** process, with your output from this input engine preloaded as input, has been added to the output Results window. You can launch this process to summarize counts using existing gene model information generated for individual positions or bins without specifying a gene model annotation file during import.

Generate Counts from BAM (BAM Input Engine) *New!*

- This process uses samtools 0.1.12 to convert aligned reads in BAM format to SAM format and parse CIGAR strings within the file(s) to generate output SAS data sets containing raw counts, RPM, and RPKM values. To use this process, you must download samtools 0.1.12 from the [SAMtools Web site](#) and save the executable files in the C:\Program Files\SASHome\JMP\9\Genomics\ThirdPartyAnnotation\NextGen directory.
- Counts can be generated at every genomic position. Alternatively, binning criteria can be specified to summarize counts either within equally spaced bins, or within exon or genes if an existing gene model in UCSC, text, or BED file format is provided that provides positional information for genomic features. In addition to raw counts, RPM and RPKM values are now also output for bins, exons, or gene features.
- To bin read counts using known gene models on the Annotation tab, you must download a text file from the UCSC table browser or create a gene model table in UCSC format. To download an existing file for use in this process, or to use as a template to create your own UCSC-formatted file, click on the **Table Browser** link on the [UCSC home page](#). Select the genome build that corresponds to your genome of interest, and choose an available isoform table for that genome and save the resulting output as text.
- PCR duplicates are assumed to be aligned reads from the same chromosome that start at the same position. It has been suggested that failing to remove PCR duplicate reads can lead to bias when coverage is low or moderate. A new check box has been added to the Options tab to address this problem. Check this option to remove PCR duplicates. **Note:** Selecting this option might not be appropriate for your experiment, so we encourage you to research the issue before selecting it. You should not select this option when performing high-coverage

targeted resequencing, where duplicate reads might be expected to occur that are not the result of PCR duplication.

- A button to auto-launch the **Gene Model Summary** process, with your output from this input engine preloaded as input, has been added to the output Results window. You can launch this process to summarize counts using existing gene model information generated for individual positions or bins without specifying a gene model annotation file during import.

Generate Counts from Eland (Eland Input Engine) *New!*

- This process imports aligned reads in Eland files, parsing Descriptor strings within the specified file(s) to generate output SAS data sets containing raw counts, RPM, and RPKM values.
- Counts can be generated at every genomic position. Alternatively, binning criteria can be specified to summarize counts either within equally spaced bins, or within exons or genes if an existing gene model in UCSC, text, or BED file format is provided that provides positional information for genomic features. In addition to raw counts, RPM and RPKM values are now also output for bins, exons, or gene features.
- To bin read counts using known gene models on the Annotation tab, you must download a text file from the UCSC table browser or create a gene model table in UCSC format. To download an existing file for use in this process, or to use as a template to create your own UCSC-formatted file, click on the **Table Browser** link on the [UCSC home page](#). Select the genome build that corresponds to your genome of interest, and choose an available isoform table for that genome and save the resulting output as text.
- PCR duplicates are assumed to be aligned reads from the same chromosome that start at the same position. It has been suggested that failing to remove PCR duplicate reads can lead to bias when coverage is low or moderate. A new check box has been added to the Options tab to address this problem. Check this option to remove PCR duplicates. **Note:** Selecting this option might not be appropriate for your experiment, so we encourage you to research the issue before selecting it. You should not select this option when performing high-coverage targeted resequencing, where duplicate reads might be expected to occur that are not the result of PCR duplication.
- A button to auto-launch the **Gene Model Summary** process, with your output from this input engine preloaded as input, has been added to the output Results window. You can launch this process to summarize counts using existing gene model information generated for individual positions or bins without specifying a gene model annotation file during import.

Gene Model Summary

- In JMP Genomics 5.0, a **Track Gene Text** setting was used to specify a gene model file in this process. In JMP Genomics 5.1, the text file that contains the gene model information is required as input on the Anno1 tab instead. To download an existing file for use in this process, or to use as a template to create your own gene model table in UCSC format, click on the **Table Browser** link on the [UCSC home page](#). Select the genome build that corresponds to your genome of interest, and choose an available isoform table for that genome. Save the resulting file as a .txt file.
- A secondary Anno2 tab has been added to allow the merge of additional annotation information that is not in the gene model file. This additional annotation information must be con-

tained in a second SAS data set. For example, if you are working with a gene model file from UCSC, you can have a secondary annotation file that enables you to match the UCSC identifiers with gene names or annotation information from a second source. Key files that match UCSC gene identifiers and other common gene identifiers (e.g., Genbank, UniGene, and so on) can be generated using tools like DAVID (<http://david.abcc.ncifcrf.gov/conversion.jsp>).

Call Variants with SAMtools *New!*

- This process uses samtools/bcftools (version 0.1.12) to call SNPs/INDELs from BAM files, generating VCF (variant call format) files as output. You can then import the VCF files into a SAS data set using the **VCF Input Engine**. Please note that to use this process, you must download samtools from the [SAMtools Web site](#) and save the executable files in the C:\Program Files\SASHome\JMP\9\Genomics\ThirdPartyAnnotation\NextGen directory.
- Please note that for large BAM files, the process of creating VCF files can be computationally intensive. If you observe a “File in Use” error message, this indicates that bcftools is still updating the VCF file. Please wait until the output VCF file has been finalized before attempting to open and import into SAS data set format using the **VCF Input Engine** auto-launch button. If you encounter this error, we recommend that before attempting to import VCF files into SAS data set format, you open the finalized VCF file in a text editor and shorten the column headers by removing the file path information that bcftools includes in the column name. If no error is encountered, this step will be performed automatically by JMP Genomics and you may proceed to import the VCF files without modification.

CLC Bio Input Engine *New!*

- This process imports SNP and indel summary files from CLC bio software. By default, files are imported as tall data sets, with information about features in rows and sample in columns. The tall data sets can be used as input to the new **IBS Sharing Regions** process. Optionally, wide data sets can also be created for further analysis in JMP Genomics processes, which take this format as input. Output data sets can be filtered to include only SNPs or indels where less than a specified number of genotypes are missing in the input files.
- Please note that depending on the size and file type(s) being imported, it might be necessary to increase the value specified in the Number of Rows to Scan parameter on the Options tab so that character values of chromosome identifiers (e.g., X, Y) are detected when occurring late in the file. If the Number of Rows to Scan is not set at a high enough number, you will likely see missing values for chromosome in the output data set.

Complete Genomics Input Engine *New!*

- This process imports text files from Complete Genomics' bioinformatics pipeline. Supported files include variant (var), annotated variants within known genes (gene), dbSNP (dbSNPannotated), and gene variant (gene) summary (geneVarSummary) files from file format v1.3, software v1.8.0. By default, all files are imported as tall data sets, with information about features in rows and samples in columns. You can click on the **IBS Sharing Regions**

auto-launch button in the output tabbed report to perform further analysis on tall data sets imported from var, gene, and dbSNPannotated formats.

- Optionally, wide versions of imported var, gene, and dbSNP annotated data sets can also be created for further analysis in JMP Genomics processes, which take wide data sets as input. Output data sets can be filtered to include only SNPs or indels where less than a specified number of genotypes are missing in the input files. Please note that creating wide data tables can be extremely time-consuming, and might not be possible on a 32-bit desktop computer. An alternative is to reduce the size of the data set to be transposed by filtering or running IBS Sharing Regions first on the tall data sets and then selecting only interesting areas of the genome to transpose.
- Please note that depending on the size and file type(s) being imported, it might be necessary to increase the value specified in the Number of Rows to Scan parameter on the Options tab, so that character values of chromosome identifiers (e.g., X, Y) are detected when occurring late in the file. If the Number of Rows to Scan is not set at a high enough number, you will likely see missing values for chromosome in the output data set.

VCF Input Engine *New!*

- This process imports VCF v4.0 SNP data files such as those used in the 1000 Genomes Project. By default, files are imported as tall data sets, with information about features in rows and samples in columns. The tall data sets can be used as input to the new **IBS Sharing Regions** process. Optionally, wide data sets can also be created for further analysis in JMP Genomics processes, which take this format as input.
- Please note that depending on the size and file type(s) being imported, it might be necessary to increase the value specified in the Number of Rows to Scan parameter on the Options tab so that character values of chromosome identifiers (e.g., X, Y) are detected when occurring late in the file. If the Number of Rows to Scan is not set at a high enough number, you will likely see missing values for chromosome in the output data set.

Import a Designed Experiment from Text, CSV, or Excel Files

- A new Select Key Variable to Merge Files option has been added to the General tab. In prior releases of JMP Genomics, you could use row ID variables for merging multiple files (as in previous versions). To do this, however, required that the row ID variables be located in the data set. In JMP Genomics 5.1, these row ID variables are no longer required when this new option to use the row number as the identifier variable is selected. **Note:** The option to use the row number as the identifier variable should only be selected if all the data sets are ordered consistently.
- The experimental design file (EDF) columns Array and Intensity columns are no longer required for the import process. Instead, a new Count column, which lists the name of the column variable in each text file that holds count values for that file, is required.

Workflows

Basic RNA-Seq Workflow *New!*

- This new workflow can be used both to analyze pre-summarized count data that has been imported into SAS data sets from text files and to assess SAS data sets consisting of counts generated in JMP Genomics from SAM, BAM, or Eland files.
- In addition to offering quality control and analysis options similar to those offered by the **Basic Expression Workflow**, new options address the use of count data in this workflow. The Input data is log transformed option on the General tab enables you to specify whether your input data has been transformed. The Shifted log₂ Transformation for QC option on the QC and Normalization tab enables you to view log-scale QC plots when working with raw intensities or counts.
- This workflow features two normalization methods, TMM (Robinson and Oshlack 2010) and a variation called KDMM (Chu, unpublished), that are especially appropriate for count data. These methods use kernel density information to normalize read counts across a set of samples.
- Either Continuous or Count data can be used when performing row-by-row modeling analysis in the workflow. Please note that count or continuous refers to the type of distribution assumed for numeric values for each feature (e.g., gene, exon, bin) across all samples in the input data set. Count data that has already been *log*-transformed can be treated as continuous and analyzed using PROC MIXED, while non-transformed count data can be treated as counts and analyzed with PROC GLIMMIX. If Count is selected, the distribution of the data must also be specified below using the Distribution of Data pull-down menu. Two commonly used distributions for read counts are Poisson and Negative binomial, but other options can be selected. When the Distribution of Data field is left empty, the default distribution assumed for count data will be Poisson. The link function appropriate to the selected data distribution is automatically selected by PROC GLIMMIX.

Basic miRNA/miRNA Seq Workflow

- This process has been renamed since it can be used to analyze both intensity data generated from microarrays as well as count data summarized from next-gen sequencing experiments. In addition to offering quality control and analysis options similar to those offered by the **Basic Expression Workflow**, new options have been added to this workflow to address the use of count data.
- The Input data is log transformed option on the General tab enables you to specify whether your input data has been transformed.
- The Shifted log₂ Transformation for QC option on the QC and Normalization tab enables you to view log-scale QC plots when working with raw intensities or counts. When this option is selected and a normalization method (applied to raw data) is also specified, QC plots generated for data before and after normalization will both be displayed on the *log* scale. Please note that if you do not select the option to shift *log*₂-transform for QC, but elect to perform *log*-transformation prior to normalization, your pre- and post-normalization QC plots will be on different scales. We therefore recommend that if you elect to perform log

transformation prior to normalization, you check the Perform shifted log₂ transformation check box.

- You can specify whether to perform a shifted log transformation on raw data prior to normalization or prior to ANOVA modeling. Two new normalization options have been added to the QC and Normalization tab that incorporate kernel density information: Kernel Density Loess (Hsieh, Chu, Lin, and Wolfinger 2011) and Kernel Density Quantile. In this workflow, all data sets will be analyzed on the log scale using PROC MIXED. If you want to perform modeling using count data directly, this option is available in the Basic RNA-seq Workflow.
- Please note that if your miRNA data set has many missing values, you might want to run a process such as **Filter Intensities** or **Statistics for Rows** to remove rows with many missing values. We have observed that, because of their small size, miRNA data sets often must be filtered heavily to remove rows with missing values before analysis. This is especially true when attempting to use count data for row-by-row modeling using count data options that call PROC GLIMMIX.

Basic Exon/Alternative Splicing Workflow

- Either Continuous or Count data can be used when performing alternative splicing analysis on data sets that have been summarized at the exon level. Please note that when selecting this option, count or continuous refers to the type of distribution assumed for numeric values for each feature (e.g., gene, exon, bin) across all samples in the input data set. Count data that has already been log-transformed can be treated as continuous and analyzed using PROC MIXED, while non-transformed count data can be treated as counts and analyzed with PROC GLIMMIX. If Count data type is selected, the distribution of the data should also be specified below with the Distribution of Data pull-down menu. Two commonly used distributions for read counts are Poisson and Negative binomial, but other options can be selected. When the Distribution of Data field is left empty, the default distribution assumed for count data will be Poisson. The link function appropriate to the selected data distribution is automatically selected by PROC GLIMMIX.

Genetics Rare Variant Workflow

- This workflow codes rare variants according to a dominant model, and, optionally, combines rare variants within a gene (or other pre-defined set of SNPs) into a single locus to perform association tests. Alternatively, the workflow can be used to perform combined tests on all common and rare variants within a gene or SNP set.
- The workflow now includes three rare variant tests. Tests include the Cohort Allelic Sums Test (CAST, Morgenthaler and Thilly 2007), which has been added in this release, as well as the Combined Multivariate and Collapsing (CMC) method (Li and Leal, 2008) and the Rare Variant Test 2 (RVT2, Morris and Zeggini, 2010), which were supported previously.
- For more information about the tests performed in this workflow and others implemented in the new Rare Variants Association process, please see the **Rare Variant Tutorial** under the Genetics > Other Association Tests section of the menu.

Genetics

Genetics Utilities

Recode Genotypes

- A new option to append a prefix to the original SNP name when recoding has been added. This check box option, *Append Prefix to Current Marker Name*, is disabled if *Genotype String (A/A A/B B/B)* is selected as the type of recoding to be performed.
- Output data sets from this process now include both the minor allele and major allele when a marker is biallelic.

Flip Strand *New!*

- This new process helps reconcile strand differences by comparing major and minor alleles for SNPs in an annotation data set to a reference annotation data set and flips the strand (i.e. switches both alleles to their complements) for non-A/T and non-C/G SNPs that have the complementary alleles in the reference set. A/T and C/G SNPs can optionally be flipped as well. A *Status* column is included in the output annotation data set to show which SNPs have been flipped, and also identifies other issues such as missing major or minor alleles and incompatible genotypes between the annotation data set and reference (more than 2 alleles for a SNP).

Relatedness Measures

Relationship Matrix

- A new parameter, *Proportion of Alleles Identical by Descent Threshold*, has been added to the *Analysis* tab. You can specify a cutoff value using this new parameter when *Identity by Descent* has been selected as the *Relationship Matrix to Compute*. If the value specified is greater than 0, a data set with the suffix *_prs* will be created that lists pairs of individuals that share greater than or equal to that proportion of alleles identical by descent..

IBS Sharing Regions *New!*

- This process identifies regions of consecutive SNPs that are shared *Identical By State (IBS)* between a set of affected individuals.
- You can specify columns of parents or founders from the input data set to examine only loci for which all parents are heterozygous.
- Sample genotypes can be in a separate data set for each chromosome, or in a single data set.
- "Runs" (denoted "streaks" by Leibon, Rockmore, and Pollak (2008)) can be determined in terms of a set of at least a certain number of individuals (offspring) sharing an allele (Thomas et al. 2007), only those sharing at least one variant allele, or only sharing two variant alleles.

Population Measures

- A *Fst* option has been added to the Measure of Genetic Distance drop-down menu. on the Analysis tab. Please note that when this option is selected, use of an Annotation By group variable is not allowed, and *FST* is calculated over all markers.

Genetic Marker Statistics

Marker Properties

- The *p*-value ("Marker Statistics") output data set now includes the minor allele as well as the major allele when a marker is biallelic.
- New fields have been added to the output tabbed report that enable you to specify cutoffs for minor allele frequency, percentage of missing genotypes, and HWE test statistics. After viewing the distributions of each of these parameters and setting your desired cutoffs, you can use a new action button to automatically launch the **Subset and Reorder Genetic Data** process to filter your data set.

Linkage Disequilibrium

- An option to output data sets containing the correlation coefficient column *CorrCoeff2* in matrix format has been added to the Output tab. This option should not be used when working with genome-wide data sets.

Association Testing

The former **Association Testing** submenu section has been divided into two new sections, **GWAS Testing** and **Other Association Testing**. Processes for use on genome-wide data sets are found under the **GWAS Testing** submenu, including association testing for case-control designs, tests that use various linear modeling methods, those that support data from complex survey designs or imputed genotypes, as well as TDT variations and the new process **GWAS Meta-Analysis**. The **Other Association Testing** submenu contains more computationally intensive processes such as those intended for rare variants association, multiple SNP association, pleiotropic association, SNP interaction assessment and testing, Q-K mixed model analysis, and association testing methods that support multiallelic markers.

GWAS Testing

GWAS Meta-Analysis *New!*

- This new process performs meta-analysis of genome-wide association studies by combining *p*-values or effects for a SNP from multiple studies and calculating a combined *p*-value. Results can be combined using one of two methods: the first uses *p*-values from the studies, converts them to *z*-scores and then combines the *z*-scores signed by the selected Effect Variable and weighted by the square root of the sample size (Stouffer et al. 1949); the second method weights effects, such as regression coefficients, from the studies by the inverse vari-

ance, calculated as the inverse square of the Standard Error variable, and calculates a z -score based on the weighted effect and its standard error (Wang and Bushman 1999).

Other Association Testing

Rare Variant Tutorial *New!*

- This new tutorial details the various methods implemented in both in the **Genetics Rare Variants Workflow** and **Rare Variant Association** processes. Click on different buttons within the tutorial to view detailed information about the publications and the approaches as implemented in JMP Genomics.
- Click on the action button, located below the text detailing each method, to launch the appropriate process with the correct options selected.

Rare Variant Association *New!*

- This process tests for association of a trait or disease with rare variants and, optionally, common variants, that occur in the same gene or pathway. The options provided in this process are designed to accommodate several models for analyzing rare variants, aligning with the unified framework for rare variant tests described in Hoffmann, Marini, and Witte (2010).
- The methods include the weighted sum method described by Madsen and Browning (2009), a variable-threshold approach (Price et al. 2010) with or without weights such as PolyPhen-2 scores, models taking the direction of the variant's effect into account, similar to Han and Pan's data-adaptive sum model (2010), as well as combinations of these approaches. Three other methods are available for binary traits: the KBAC (Liu and Leal 2010), the CMAT (Zawistowski et al. 2010), and the C-alpha test (Neale et al. 2011).
- For more information about the tests performed in this process and others implemented in the **Genetics Rare Variants Workflow**, please see the **Rare Variant Tutorial** under the Genetics > Other Association Tests section of the menu.

Haplotype Analysis

Haplotype Estimation

- A new check box, available on the Output tab under the Phase Assignment Data Set Options section, enables you to output only the most probable haplotype pair to the output data set.
- The haplotype-trait association results, which were previously displayed only in the HTML report, are now collected in a new output table with the file suffix `_hta`.

Haplotype Trend Regression

- A new Create HTML Output check box on the output tab enables you to view the SAS HTML output from PROC Haplotype. The regression results, which were previously displayed only in the HTML report, are now collected in a new output table. A table summariz-

ing global tests has file suffix `_htg`, while a separate table of single tests is output with the suffix `_hrs`.

- This process now accommodates Class covariates.

Linkage Maps and QTL

The former **QTL Mapping** submenu section has been expanded and renamed.

Recombination and Linkage Groups *New!*

- This new process uses experimental cross design information to help identify linkage groups within a set of markers for a biparental experimental population. It calculates a matrix of pairwise recombination rates based on an experimental cross design and clusters the matrix to identify groups. It also calculates segregation ratios for each marker.

Linkage Map Order *New!*

- This new process generates marker order solutions within linkage groups using **Multidimensional Scaling** (MDS) or optimization methods from SAS/OR. MDS is used to find a one-dimensional ordering that best represents the structure found in the matrix of recombination frequencies. A residual plot of the MDS fit can help to determine and flag outlier markers that do not fit well in the ordering. The MDS algorithm is a simple and quick way to find approximate ordering and works well with small, well-formed linkage groups. Alternatively, PROC OPTMODEL from SAS/OR can be used to apply a sophisticated map order optimization algorithm based on directed graph theory. Marker ordering is a version of the famous Traveling Salesman problem and the algorithms implemented in the OPTMODEL procedure find a globally optimal marker order to build the shortest genetic map possible.

Note: To use the SAS/OR-enabled option, you must license this SAS module separately from SAS and combine the module with your JMP Genomics SAS license. Please contact sales@jmp.com for more information.

- Linkage Map Order determines the probable order of genetic markers within linkage groups based on recombination frequencies and calculates the genetic distances between markers to produce a linkage map. The input to this process can be the output data set from the **Recombination and Linkage Groups AP** containing a symmetric matrix of pairwise recombination rates between markers and a variable whose values determine the linkage group to which each marker belongs. Marker ordering is done only within linkage groups.
- You can use the action buttons in the tabbed report results to drill down and visualize pairwise recombination rates for adjacent markers using interactive triangle plots.

Linkage Map Viewer *New!*

- This process enables you to visualize a linkage map containing markers ordered within linkage groups in both 2-D and 3-D representations.
- You can input a map created in JMP Genomics or other mapping software applications. For the latter, you must format the data set in the tall format required by this process. To see an example of how it should be formatted, click the **Load** button at the bottom of the AP dialog

to load an example setting, and after it loads click the **Open** button next to the Input SAS Data Set field.

Single Marker Analysis

- The analysis and output has been reorganized for this process. Unlike **IM and CIM Analysis**, this AP scans all markers for QTL association via regression modeling without using linkage information. The output has been changed to be more consistent with other genetic processes, such as **SNP-Trait Association**, and includes new features such as a summary chart view of significant QTLs across chromosomes, a genome-wide view of genotype tests for association with the QTL(s), and volcano plot views of the genotype effect estimate on the QTL. When multiple traits are specified, the **Venn Diagram** action button now enables you to drill down to find loci that are significant across all traits.

Multiple QTL Analysis APs

- The genotype notation in the dialog and help have changed from the previous notation of (MM, Mm, mm, M-, m-, missing) to (AA, AB, BB, A- [not BB] , B- [not AA], missing).
- The numeric genotype coding for genotypes (AA, AB, BB, A-, B-, missing) has changed from (2, 1, 0, -3, -2, -1) to (0, 1, 2, 3, 4, .).
- If the data that you are using are numerically coded in the previous genotype format, you can select the check box option on the Options tab. Use the QTL Data Numeric Coding from Previous JMP Genomics Versions parameter located on the Options tab, to convert the data to the new format.
- You can click the help **?** next to this new parameter to see a table comparing the previous numeric coding with the current one. Affected APs include **Single Marker Analysis**, **Build Genotype Probability Data Set**, and **IM and CIM Analysis**.

Breeding Analysis *New!*

Phenotype Summary

- A new link to this existing process has been included in the new **Breeding Analysis** section of the menu since this process can be useful for creating summary plots for categorical and quantitative phenotypic variables.

GxE Interaction *New!*

- This new process analyzes the performance of different genotypes in a multi-environment trial, displaying stability measures, genotype and GxE least square means from linear-bilinear models, PCA biplots, and heritabilities.

Modifications to Multiple Genetics APs

- A new check box option is available in several APs to perform tests for only SNP Interaction effects, not SNP main effects. Affected APs include **SNP-Trait Association**, **Pleiotropic**

Association, Survey SNP-Trait Association, Q-K Mixed Model, and the Genetics Q-K Analysis Workflow.

- When the trend test is run in **SNP-Trait Association**, **Q-K Mixed Model**, **SNP-SNP Interactions**, **Survey SNP-Trait**, or **Imputed SNP-Trait Association**, the output p -value data set now includes the following columns: Estimate_Trend and StdErr_Trend. If any class interaction variables are specified, these columns will be missing. Non-class interaction effects should have values in these columns.
- A new SampleSize column is included in the output p -value data sets for the following APs: **Case-Control Association**, **PCA for Population Stratification**, **SNP-Trait Association**, **Imputed SNP-Trait Association**, and **Survey SNP-Trait Association**. Exceptions include:
 - when **Survey SNP-Trait Association** is either run with a non-continuous trait where only Rao-Scott tests selected, or run with a continuous trait.
 - when **Case-Control Association** is performed with Dominant and/or Recessive asymptotic tests only, and the Fast version is not being performed.

Copy Number Analysis

Copy Number Partition

- A new interactive summary graphic is created on the output Chromosome Segmentation Summary tab. This graphic displays the summarized mean value for small segments of the chromosome. To drill down to view chromosome segmentation details, select points in the summary graphic and click the **View Chromosome Segmentations** action button.
- New data sets are also produced and listed under the Output SAS Data Set section of the tabbed report, including a stacked data set that displays mean values for each identified segment.

Expression

Quality Control

Correlation and Principal Components

- The dialog for this process has been renamed **Correlation and Principal Variance Components Analysis** to reflect that both principal components analysis and variance components analysis are performed in the workflow.
- The Analysis tab has been renamed PCA. On this tab, the Number of Principal Components parameter has been renamed to Maximum Number of Principal Components to Apply. Pre-

viously, you could specify either a number of principal components or a cumulative proportion of variation to decide how many components to calculate, but now it is possible to specify both criteria. When both criteria are specified, the condition met first will be used as a stopping point for calculating additional components.

- A new tab called VCA has been added. This tab contains features related to Variance Components Analysis. On this tab, the parameter Number of the First Component in Model has been renamed to Number of the First Principal Component for VCA. The parameter Number of the Last Principal Component in Model has been removed.

Correlation and Grouped Scatterplots

- A new option to specify whether scatterplot groups should be displayed horizontally or stacked vertically has been added.

Normalization

TMM Normalization *New!*

- TMM (Trimmed Mean of M component) is a scaling normalization method for RNA-seq data (Robinson and Oshlack 2010). This process takes as input a tall data set with summarized read counts for features (bins, exons, or genes) in rows, with one sample per column. The counts in the input data set can be generated in JMP Genomics from SAM, BAM, or Eland files, or created in another program, then imported from text files into SAS data sets.

KDMM Normalization *New!*

- KDMM (Kernel Density of Mean of M component) is a scaling normalization method for RNA-seq data similar to TMM (Robinson and Oshlack 2010). The data set is preprocessed and summarized into bins, exons, or genes with each row containing data from a unique individual bin, exon, or gene across samples (columns). The M and A components between the target sample (under normalization) and reference sample are calculated for estimating a two-dimensional kernel density and applying the density for the weighted mean of the M component as the scaling factor corresponding to the target sample.

Caution: This process can be computationally intensive for large data sets.

Row-by-Row Modeling

ANOVA

- A new Data Type tab has been added. The options on this tab enable selection of Continuous or Count as the type of data being analyzed. If Continuous is selected, PROC MIXED is

called to run the analysis, as in previous versions. If Count is selected, PROC GLIMMIX is called instead, triggering several new options and changes to the output of ANOVA:

- New options that enable you to set advanced parameters like the distribution of the count data and the desired link function (both parameters required by GLIMMIX) become active.
- A multiplicative overdispersion parameter (triggered by the statement `random _residual _;`) is added to the GLIMMIX procedure call. Although this statement is optional when GLIMMIX is run independently, it is set as a default in JMP Genomics. For more information about this parameter, please see the PROC GLIMMIX documentation.
- The output tabbed report for ANOVA includes a histogram displaying the Generalized Chi-Square Divided by Degrees of Freedom instead of residual variance.
- Studentized Residual Plots are displayed instead of Standardized Residual Plots.
- Since JMP does not provide modeling options similar to those available in PROC GLIMMIX, the **Fit Model** and **Plot LS Means** drill down buttons have been updated to rerun PROC GLIMMIX in SAS and display the corresponding Model Fitness and LS Means plots as SAS ODS graphics in an html file.
- The default estimation method for computing degrees of freedom has been changed from CONTAINMENT to RESIDUAL. The residual method is faster and can be more liberal in unbalanced designs with random effects; that is, the number of significant results can be greater for a particular p -value cutoff. As a result, there may be some differences in mixed model results from the same data analysis between these two software versions.
- A new Filter to Include Observations parameter that enables you to specify rows to be included in an analysis based on values in the input data set has been added to the General tab.
- A new Compute Multiple Testing Adjustment Separately for Each Test option has been added to the Test tab. Select this option to perform calculation of p -value cutoffs for significance based on individual difference tests rather than by calculating a single global p -value cutoff across all tests.
- A new Additional Filter for Tests section has been added to the Test tab. Use these options to specify an X -axis filter that will be applied (along with any p -value cutoff specified) to decide which tests are identified as significant. Simply specify a cutoff and the desired direction (positive, negative, or absolute value), and this cutoff will be used to filter the `_sig` table and draw vertical cutoff lines in volcano plots.
- On the Residuals tab, the default value for the Filtration Method for Data Points with Large Residuals parameter has been changed from False Positive Rate to None to avoid errors that result when running ANOVA when the data type is specified as Count on the Data Type tab.
- A new auto-launch button for **Gene Set Enrichment** is available in the output tabbed report.

Mixed Model Analysis

- On the Residuals tab, the default value for the Filtration Method for Data Points with Large Residuals parameter has been changed from False Positive Rate to None. This change was made to accommodate new modeling options that use count data.

Difference Chooser

- An **Open** button is now available next to the result folder path in the JMP Genomics Message window.

Survival Analysis

- New options have been added to the Model tab to enable the specification of LSMMeans effects. Corresponding options are available on the Test tab to select LSMMeans difference sets for display in volcano plots, create custom difference lists with the **Difference Chooser**, and add estimate statements using the JMP Genomics **Estimate Builder**.
- The existing Covariates field has been updated to allow class and continuous covariates as well as interactions to be specified as covariates when running row-by-row survival analysis.

Expression Utilities

Combine Experiments

- An error message is now printed that identifies data sets that could not be merged due to missing merge key variables. All data sets that contain the specified key variables are merged.

Pattern Discovery

Cross Correlation

- Two new parameters, Multiple Testing Method for Output Test Data Set and $-\log_{10}(\text{p-value})$ Cutoff have been added to the Options tab. Please note (especially when working with large data sets) that specifying a multiple testing method and cutoff will still cause the full cross correlation results data set to be generated before filtering based on this criteria. If you desire simply to reduce the size of the output data set, you might find it preferable to set a simple $-\log_{10} p$ -value cutoff on the Options tab and leave the Multiple Testing Method blank.
- When working with large data sets, you might also want to uncheck the option for display of the cluster heat map on the Analysis tab. The memory required to display the heat map can be greater than your system resources, especially when working on a 32-bit computer with limited RAM.

Modifications to Multiple Pattern Discovery APs

- A new Where Statement parameter is available on the General tab of the following APs: **Principal Component Scoring, Distance Matrix and Clustering, Hierarchical Clustering, KMeans Clustering, and Principal Components Analysis.**

Predictive Modeling

Main Methods

Partial Least Squares

- The Dependent Variable field on the General tab has been simplified to be consistent with other predictive modeling APs. Only one dependent variable is now permitted.

Survival Predictive Modeling

- A new parameter, Maximum Order of Interactions, which enables you to specify the degree of interactions that you wish to examine between predictors, has been added to the Analysis tab. If this parameter has a value of 1, only main effects are examined. Please note that selecting higher order interactions with this option can be computationally intensive.

Modifications to Multiple Predictive Modeling APs

- A new Custom Costs option has been added to the Options tab of all main predictive modeling methods (except for **Survival Predictive Modeling**). This functionality applies only to binary or nominal dependent variables, and enables you to specify different costs of classification. It can be especially useful for diagnostic predictions in which the cost of different misclassifications might vary considerably. When you specify custom costs, the predicted class for each observation is the value with the smallest expected cost, and an Average Expected Cost (AEC) statistic is added to the output.
- A new Lock-In tab has been added to all main predictive modeling methods (except for **Partition Trees**). The options on this tab enable you to lock continuous or class predictors into the predictive model. Predictors selected on this tab are considered for removal in subsequent predictor reduction steps.

Model Comparison

Cross Validation Model Comparison

- A new Harrell's C tab is added to the output tabbed report when the dependent variable is continuous.

Predictive Modeling Utilities

Survival Residuals *New!*

- This new process creates an output data set with an extra column of deviance residuals. The deviance residuals are a transform of the corresponding Martingale residuals after fitting a Cox model with baseline with no other effects specified. You can use the residuals output from this process in other predictive modeling methods under the **Main Methods** section of the **Predictive Modeling** menu.

P-Value Operations

Meta-Analysis

- This new process performs meta-analysis of studies by combining p -values or effects for a SNP from multiple studies and calculating a combined p -value. Results can be combined using one of two methods: the first uses p -values from the studies, converts them to z -scores and then combines the z -scores signed by the selected Effect Variable and weighted by the square root of the sample size (Stouffer et al. 1949); the second method weights effects, such as regression coefficients, from the studies by the inverse variance, calculated as the inverse square of the Standard Error variable, and calculates a z -score based on the weighted effect and its standard error (Wang and Bushman 1999).

Genome Views

Genome Views is now a main menu category. The items under this menu are split into two sub-menus: Genome Browser and Track Creation.

JMP Genomics Browser

- A new Plot Bars check box option has been added to the Output tab. When this option is checked, bars are displayed on the Chromosome Color Plot. These bars indicate the value of the continuous variable of interest for each position in the genome for which values are available.

Note: When examining very large and dense count or intensity data sets (e.g., for the purposes of drilling down to examine coverage within genomic regions) it is recommended that you uncheck this option and use a cytoband or other pre-specified chromosome color theme rather than basing the color theme on the average count or intensity values. Opening a very large file containing count or intensity information and displaying dense bars will likely require more memory than is available on a 32-bit operating system or a 64-bit system with limited RAM.

Annotation Analysis

Gene Set Enrichment

- You can now specify multiple annotation category variables. When multiple variables are specified, a journal is output with links to results for each category variable. The category variable name is used in the results, although please note that this might result in the truncation of the desired data set name if the category variable name is long.
- A High Hit Threshold parameter has been added to the Analysis tab. This option simplifies removal of categories with many members from the data set before performing enrichment analysis. This removal is necessary because when such categories are included, statistical significance can be achieved with only small mean differences between groups, simply because of the number of items within the group.
- A new **Gene List** action button has been added to the output tabbed report. Use this option to generate a list of selected pathways and genes with 0-1 indicator variables to show which genes were significant (1) and not significant (0) in the enrichment test of interest. Please note that in order for gene identifiers or other annotation information to be included in the drill-down data set, you must specify them in either the Variables to Keep in Output field or the By Which to Merge Annotation Data field on the General tab of the dialog.

SAS Data Set Utilities

Tables

Transpose Tall to Wide *New!*

- The Transpose Tall & Wide process has been split into two new processes aimed at data sets with different input structures. This process transposes a tall SAS data set into a wide SAS data set.

Transpose Wide to Tall *New!*

- The Transpose Tall & Wide process has been split into two new processes aimed at data sets with different input structures. This process transposes a wide SAS data set into a tall SAS data set.

Rows

Statistics for Rows

- A new NZERO option has been added to the Statistics to Compute pull-down menu. When selected, a column reporting the number of 0 values for each row is included in the output data set.
- A new Replace Value of Counts with Proportion of Total Counts check box option has been added to the Statistics tab. This parameter becomes active when NMISS and/or NZERO (the number of missing values and number of zero values, respectively) are selected as the Statistics to Compute. Checking this option converts count values into proportions of missing or zero values over all columns. This can be useful when examining patterns for missing data or zero values in RNA-seq data.

Columns

Filter Wide Columns Based on Tall Rows

- This new utility enables you to remove columns from a wide data set based on a filter specified using variables from a corresponding tall data set. You can also specify criteria for filtering observations in the wide data set and specify variables to drop from the wide data set.

