

# JMP Genomics

Release 5.0

## Getting Started with JMP Genomics

*“The real voyage of discovery consists not in seeking new  
landscapes, but in having new eyes.”*

Marcel Proust

JMP, A Business Unit of SAS  
SAS Campus Drive  
Cary, NC 27513

## **JMP Genomics - Getting Started with JMP Genomics**

Copyright © 2010, SAS Institute Inc., Cary, NC, USA

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

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a Web download or e-book:** 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.

1st printing, October, 2010

JMP<sup>®</sup>, SAS<sup>®</sup> 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.

<b>Introduction</b> .....	<b>1</b>
<b>1 Introduction to JMP Genomics</b> .....	<b>3</b>
<b>2 Using JMP to Design New Experiments</b> .....	<b>19</b>
<b>3 Configuring JMP Genomics Settings</b> .....	<b>37</b>
<b>4 Using JMP Genomics in Client/Server Mode</b> .....	<b>43</b>
<hr/>	
<b>Appendices</b> .....	<b>51</b>
<b>5 Using JMP Genomics Graphics in Presentations and Publications</b>	<b>53</b>
<b>6 The SAS WHERE Expression</b> .....	<b>55</b>
<b>7 JMP Genomics Files Are Identified by Suffixes</b> .....	<b>63</b>
<b>8 JMP Genomics Processes Call SAS PROCS</b> .....	<b>79</b>
<b>9 Troubleshooting</b> .....	<b>87</b>
<b>10 Glossary</b> .....	<b>97</b>
<b>11 References</b> .....	<b>109</b>
<b>Index</b> .....	<b>115</b>



---

# **Introduction**

---



# Chapter 1

## Introduction to JMP Genomics

---

Welcome to JMP Genomics, a powerful desktop software system for integrated statistical analysis of genetic marker, microarray, and spectral (proteomics and metabolomics, for example) data. JMP Genomics consists of more than 100 independent analytical procedures (APs). The purpose of this manual is to provide you with informative examples of how to use JMP Genomics to extract the maximum amount of useful information from genomics data. You should be familiar with the terminology and technology associated with modern genomics analyses and standard JMP functionality. The *JMP Introductory Guide* provides information about getting started with JMP.

This chapter provides an overview of the primary functional aspects of the JMP Genomics system, descriptions of some important differences between standard JMP functionality and JMP Genomics functionality, and descriptions of the included sample data sets.

---

### About this Documentation

The JMP Genomics documentation is presented as a series of individual *manuals*. Each manual is characterized by one or more specific themes. This general theme of each book is described, as follows:

#### **I. Getting Started with JMP Genomics**

This is the manual that you are currently reading. It lists installation instructions and requirements, provides a brief introduction to JMP Genomics and this documentation. This manual also provides a trouble shooting guide, various appendices, a comprehensive glossary, index and a complete reference list for all of the manuals.

Depending on the nature of your experiments and the types of analysis that you want to perform, this manual might be the only one you refer to on a regular basis.

#### **II. Importing Data**

This manual is divided into two parts. The first part, *Experimental Design*, describes how to use JMPs DOE functions to design your experiment and generate an experimental design file, which is the first step in your analysis. The second part, *Importing Data into JMP Genomics* provides detailed descriptions of the different types of data sets used by JMP Genomics and how to import data from different sources .

#### **III. Workflows**

This manual provides detailed descriptions and instructions for setting up and using the basic JMP Genomics workflows

#### **IV. Genetic Analyses**

This manual provides detailed descriptions for manipulating genetic data sets, determining specific marker statistics, carrying out various association tests, assessing linkage, mapping quantitative trait loci, and performing haplotype analyses.

**V. Copy Number Analysis**

This manual provides detailed descriptions for normalizing and analyzing copy number data.

**VI. Spectral Preprocessing and Analysis**

This manual describes the JMP Genomics processes that are useful for analyzing two dimensional and three-dimensional spectra.

**VII. Expression Analysis**

This manual lists and describes a multitude of procedures for performing quality control on your data, normalizing it, fitting statistical models to the observations and comparing the aggregated results.

**VIII. Pattern Discovery and Predictive Modeling**

This first part of this manual lists and describes a multitude of procedures for discovering patterns in your data. The second part of this manual details a number of different predictive models as well as procedures for comparing the relative efficacy of the different models with your data.

**IX. Further Analysis**

This manual provides detailed descriptions for examining and characterizing the statistical significance of your data and for incorporating biological meaning with your statistical results.

**X. Utilities**

This manual provides you with detailed descriptions of the many tools in JMP Genomics designed to help you view and manipulate your data and transform your data sets into formats appropriate for specific analyses.

**XI. The JMP Genomics Programming Guide**

This manual provides an introduction to building and programming your own analytical processes. These processes can be added to the JMP Genomics menu.

---

## About Each Chapter

The manuals are divided into many short chapters. Each chapter corresponds to a separate AP. The manuals are organized so that chapters generally occur in the same order in which the APs appear in the JMP Genomics menu.

Chapters are divided into the following sections:

**What does it do?**

This section provides a brief overview of the AP, telling what the AP does and when and why you might use it.

**What do I need?**

This section describes any requirements for running the AP. It lists the type of input data sets needed, the format of those data sets, and the type of information that must be contained within them. Any prerequisite modifications to the data sets are described here. Additional requirements (such as accounts with online databases, and so on) are also described.

**The Dialog**

This section describes the dialog for the AP. All tabs are illustrated. All inputs (buttons, fields, text boxes, check boxes, and so on) and their purposes are described and defined.

**Example**


This section describes an example analysis using the AP. The input data is described. In most cases the input data sets are included with JMP Genomics so you can follow along with the example. Any recoding, prerequisite analysis, filtering, or other manipulations of the input data are fully described or referenced. In those case where you might need to download input data sets or other information from other providers, the location and procedure for accessing the data is described. Finally, all input settings are illustrated and described.

**Results**

The data sets, graphics and other output of the example analysis are illustrated and described. If this step is intermediate in an analysis, subsequent APs are suggested.

The text conventions illustrated in **Table 1.1** are used throughout this manual.

**Table 1.1** Text Conventions.

Symbol/Font/Style	Used to designate:
	Instruction or task to be performed
A > B > C	Navigation path from A to B to C, used for paths through nested directories
<b>A &gt; B &gt; C</b>	Navigation path from A to B to C, used for paths through nested menus
<b>Choose</b> or <b>Run</b>	Buttons or Commands
General or Options	Names of data tables, column headings, and other text generated by JMP are set in a different font
General or Options	Text to be typed by the user.

---

## Genomics Main Menu

JMP Genomics is a fully functional version of JMP plus a collection of analytical process dialogs in the **Genomics** main menu (**Figure 1.1**). It provides access to more than 100 analytical processes.

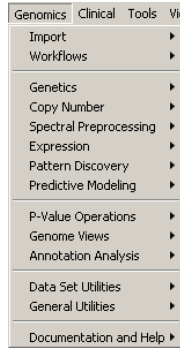


Figure 1.1 The JMP Genomics main menu is organized into submenus.

## The Genomics Starter

A new feature in JMP Genomics 5.0 is the Genomics Starter window (Figure 1.2). This dialog enables you to quickly view and access all JMP Genomics input engines, workflows, and analytical processes.

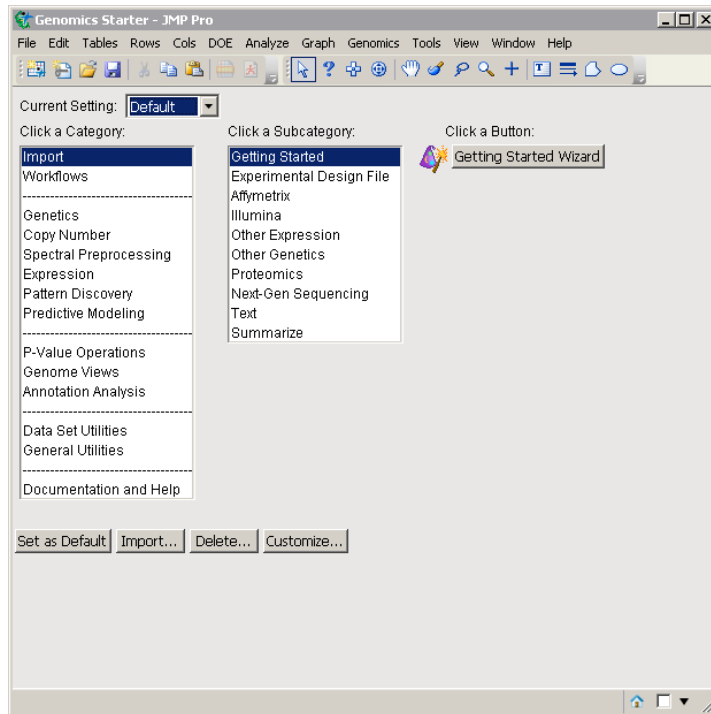


Figure 1.2 The Genomics Starter dialog

To use the Genomics Starter, first select one of the items in the Click a Category field.

When you select a category, the subcategories that it contains are displayed in the Click a Subcategory field.

☞ Select one of the items in the Click a Category field.

When you select a subcategory, specific buttons for the processes listed in that subcategory become available.

☞ Click on one of the buttons to open the desired process.

---

## Some Important Differences between JMP and JMP Genomics

### JMP Genomics Dialogs

JMP Genomics dialogs function differently from standard JMP dialogs. Standard JMP dialogs invoke calculations in compiled code, whereas JMP Genomics dialogs generate a SAS program (with suffix `.sas`), execute it in the background, and then return results. The results typically consist of SAS data sets (also known as SAS data tables, with suffix `.sas7bdat`) along with a JMP scripting language file (with suffix `.jsl`) that automatically invokes standard JMP platforms.

An important distinction of most JMP Genomics dialogs is that they do not process open JMP data tables. Instead, they prompt you to specify one or more SAS data sets that have been created and saved in your file system. This characteristic enables you to work with very large data sets without having to open them as JMP data tables and to specify multiple SAS data sets in one process.

The creation and use of JMP Genomics data sets is described more fully in [Data Sets Used by JMP Genomics](#).

---

## Deciding Which Processes to Run

An initial challenge in using JMP Genomics is deciding which processes to run and in what order they should be run. The software does not provide detailed guidance on constructing a workflow, and there are a wide variety of possible workflow combinations depending on your discovery objectives.

The Genomics menu organizes the JMP Genomics processes into groups. The groups are organized in an order that is typically used by bioinformaticians, statisticians, and data analysts. However, you are free to rearrange the menus to your liking. Refer to the [Programming Guide](#) for details about customizing menus.

JMP Genomic processes are modular, so they can be run in any order. Over time, you develop expertise with the system and form favorite workflows. The sample case studies, outlined in the *JMP Genomics User Guide*, illustrate some typical, frequently used workflows.

## Running a Process

To run a JMP Genomics process, select the process from one of the JMP Genomics menus, specify the parameters on all tabbed panes in the process dialog, and then click **Run**. The following example, which invokes the ArrayTrack Input Engine, illustrates a typical JMP Genomics process.

- ☞ Select **Genomics > Import > Other Expression > ArrayTrack**. The dialog shown in Figure 1.3 opens:

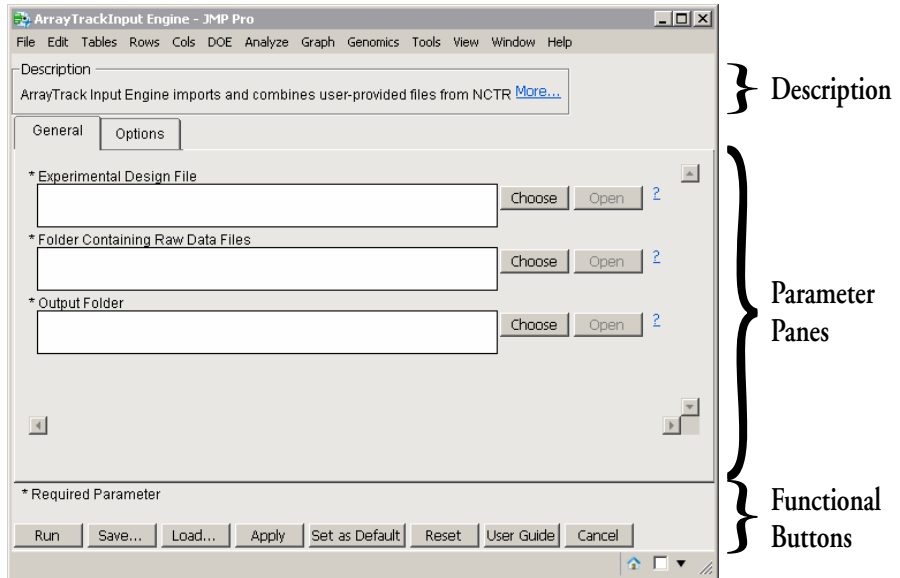


Figure 1.3 A typical JMP Genomics dialog

Each dialog has three main sections: a description box, one or more tabbed parameter panes, and functional buttons (illustrated in Figure 1.3). The description box on the top of the dialog describes the purpose of the process. The tabbed panes are the main area to specify input parameters.

The eight functional buttons, common to all of the JMP Genomics dialogs, are described in Table 1.2

Table 1.2 Functional Buttons.









Functional Button	Used to:
	Run the process using the specified parameters
	Save the specified parameters
	Load selected, saved parameters into the dialog <sup>a</sup>


Table 1.2 Functional Buttons.

Functional Button	Used to:
	Apply the specified parameters as default settings to all relevant JMP Genomic dialogs
	Save the specified parameters as the default setting.
	Clear all the parameter settings and return the dialog to its default state
	Open the specific volume of the <i>JMP Genomics User Guide</i> that contains information about the AP.
	Cancel the process and close the dialog

a. You can also load settings using the **File > Load Genomics Settings** command.


Use these buttons to load, save, or clear specified parameters, run the process using the specified parameters, or apply those parameters to other JMP Genomics processes.

---

*Note:* Clicking  opens the *JMP Genomics User Guide* to the title page of the specific volume containing information about the dialog that is open and in focus. You must still scroll down from the title page to the specific chapter describing the AP.

---

There is a defined order to the specification of some parameters. Such parameters are disabled and grayed out until their dependency requirements are fulfilled. Many processes contain multiple tabbed panes with numerous optional parameters. As you develop expertise with particular processes, be sure to investigate the often rich collection of parameters available.

🔗 Click  to the right of any parameter entry field to obtain help about its specification.

The **General** tab for each dialog typically contains the most important parameters for the process. For example, most processes require specific types of input files or data sets and an output folder. For our example, we want to open the `AT_exp2.txt` file. This *Experimental Design File*, which contains information about the experiment, is needed to import raw data into JMP Genomics.

🔗 Click **Choose** (circled in Figure 1.4)

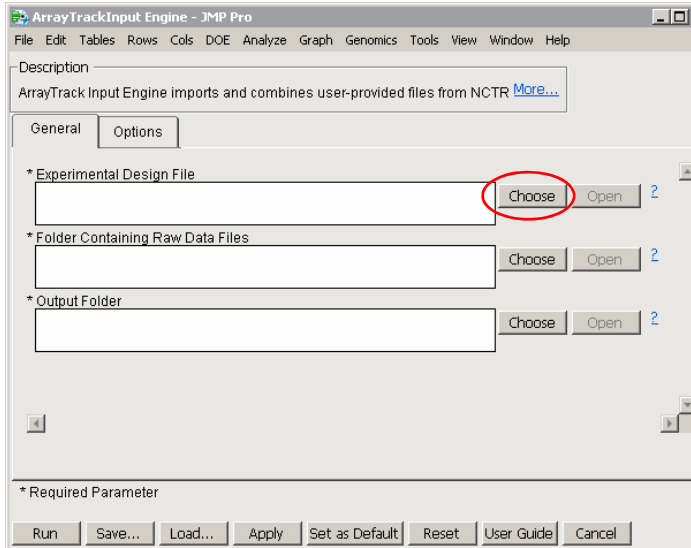


Figure 1.4 Click **Choose** to select a file or folder.

- ☞ When you installed JMP Genomics, a folder named **Sample Data** was also installed. Navigate to this folder and then to a file named **AT\_exp2.txt** by following the path **Sample Data > Microarray > ArrayTrack**.
- ☞ Click on the **AT\_exp2.txt** file.
- ☞ Click **Open** to select the file (circled in Figure 1.5)

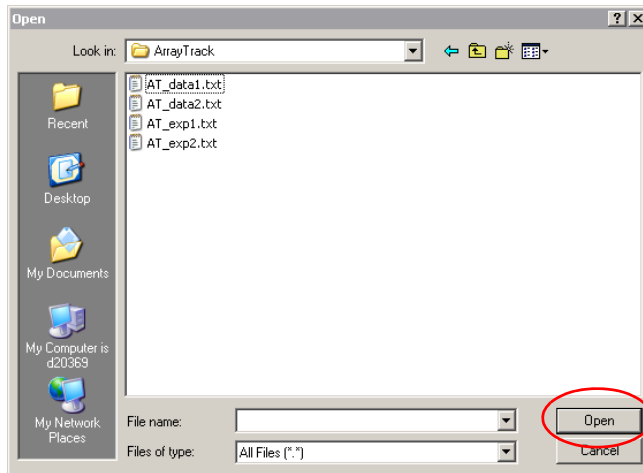


Figure 1.5 Click **Open** to select the file.

The file is added to the dialog, as shown in Figure 1.6.

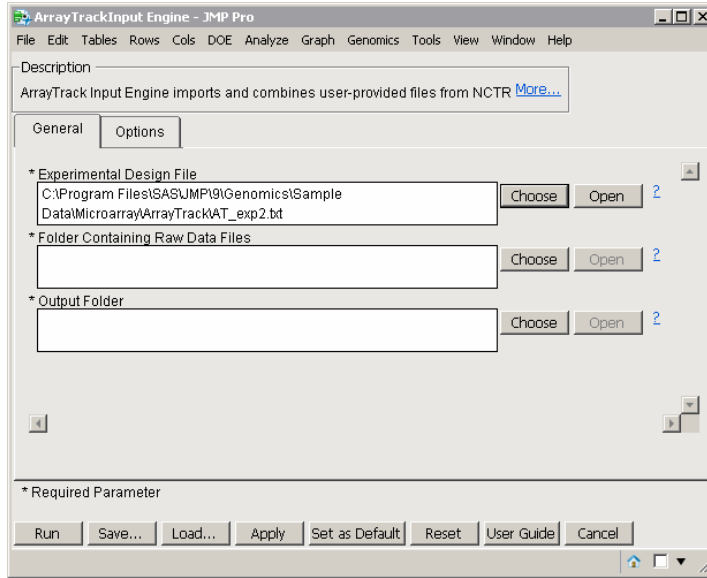


Figure 1.6 The Experimental Design File has been specified.

Our next step is to select the folder containing the raw data files.

☞ Click **Choose** (circled in Figure 1.7).

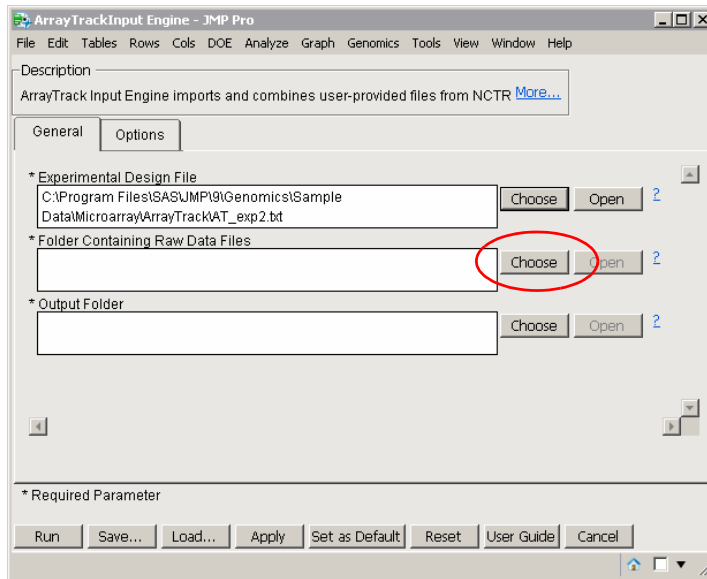
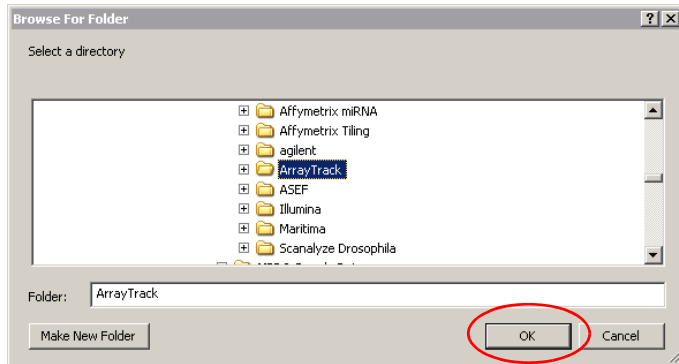


Figure 1.7 Click **Choose** to select a file or folder.

☞ Navigate to the **Sample Data** folder and then to a folder named **ArrayTrack** by following the path **Sample Data > Microarray > ArrayTrack**.

☞ Click **OK** (circled in Figure 1.8) at the bottom of the **Browse for Folder** window.

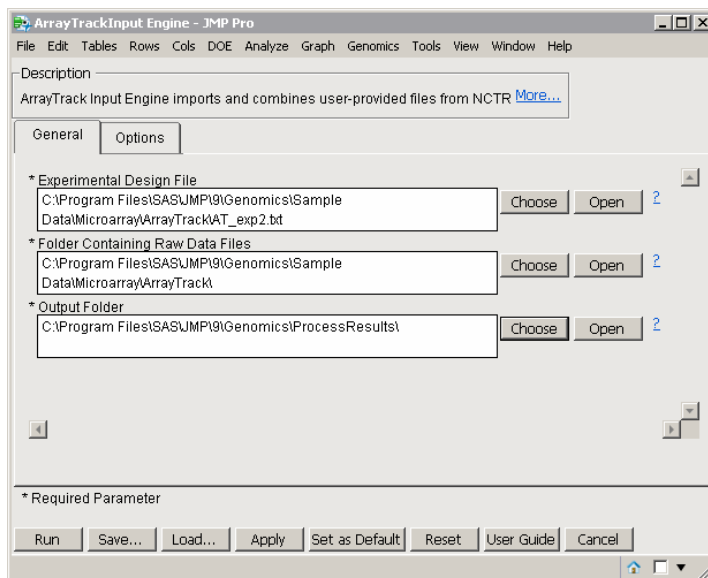


**Figure 1.8** Selecting the ArrayTrack folder

The next step is to choose a folder in which to place and store output. You can choose any folder that you like. For this example, select the **ProcessResults** folder that came with JMP Genomics.

☞ Repeat the selection process to specify the output folder.

The completed dialog is shown in **Figure 1.9**.



**Figure 1.9** The completed ArrayTrack Input Engine dialog

☞ Once you have specified the parameters for a process, click **Save** to save the parameters for later recall, if needed.

☞ Click **Run** to run the process.

JMP Genomics dialogs generate and run a SAS program each time you click **Run**. Depending on the size of your data sets and capacities of your computer, some analyses can take several minutes or, for very large and complex runs, several hours. While a program is running, the message **SAS Running Processes** window is displayed (**Figure 1.10**).

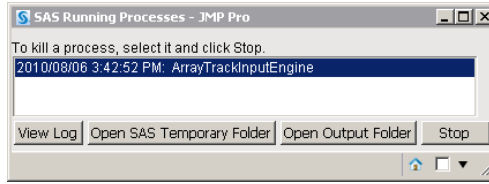


Figure 1.10 Display during analysis

While a process is running, it is a good idea to monitor progress using an application that displays statistics such as CPU, memory, and disk usage, like the Windows Task Manager. This can be informative for troubleshooting a hung process.

You can only run one process at a time. If you attempt to run a second process while another one is running, you are prompted to either disconnect from SAS and stop the current process, to view the current SAS log, or to wait until it completes.

The location of each SAS data set generated by your analysis is listed in a new window (shown in Figure 1.11). You can view each of the data sets by clicking **Open**.

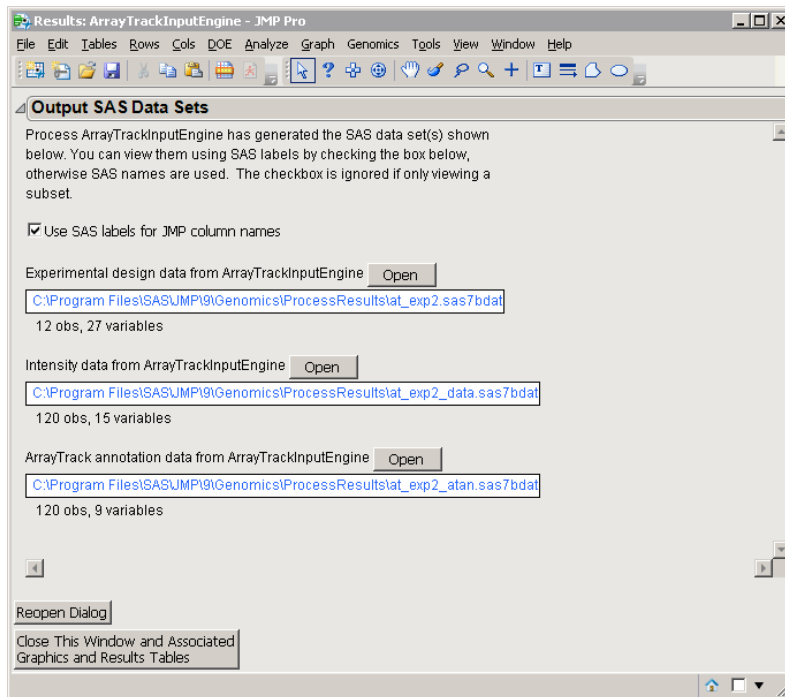


Figure 1.11 The SAS Message generated by our analysis

## Tabbed Reports

If the output from a process includes graphs and charts in addition to the data sets, these output are organized and accessed through a tabbed report (Figure 1.12).

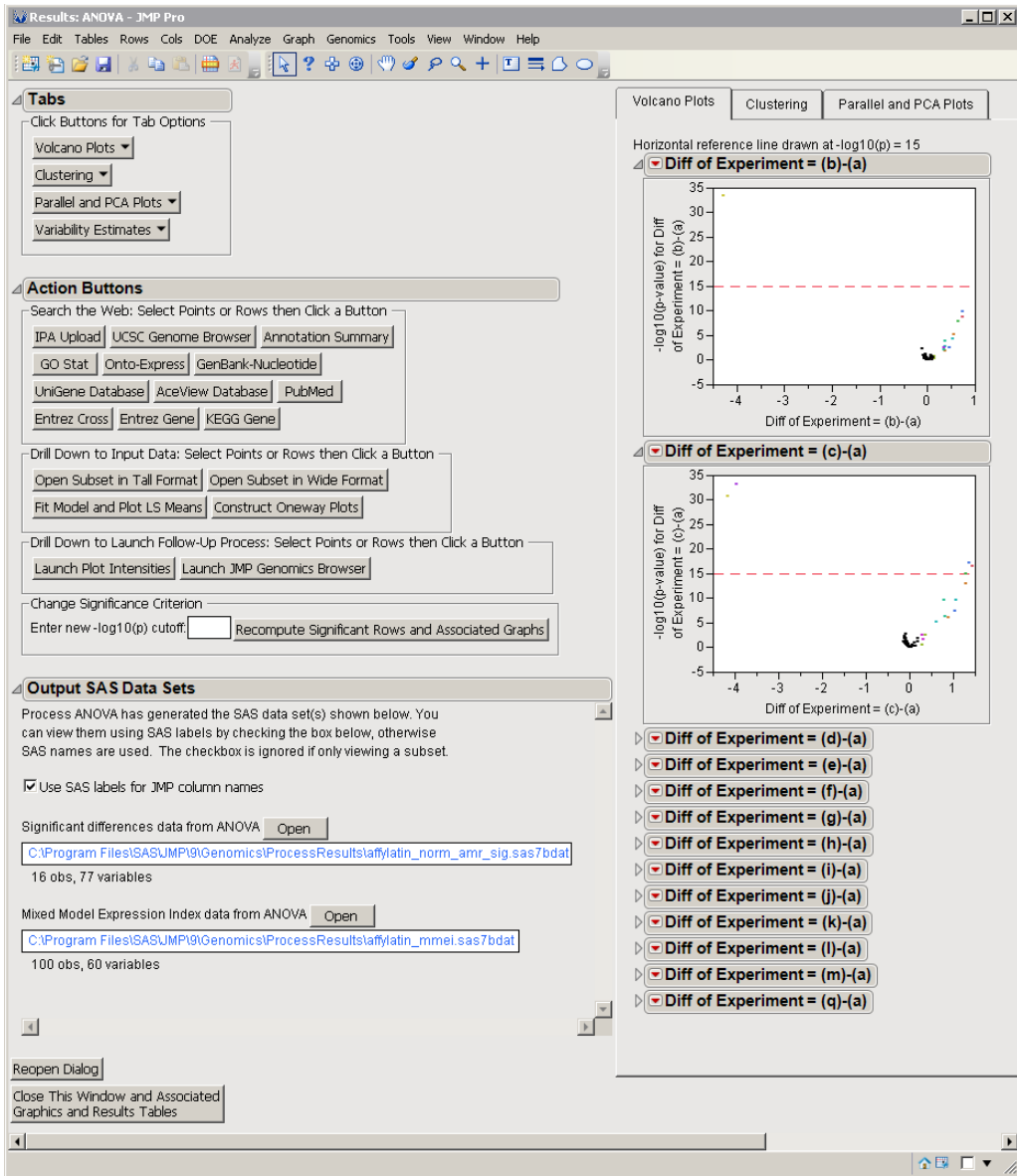


Figure 1.12 A tabbed report generated by the ANOVA process

Plots and relevant data are accessed using the drop-down menus in the Tabs pane (Figure 1.13).

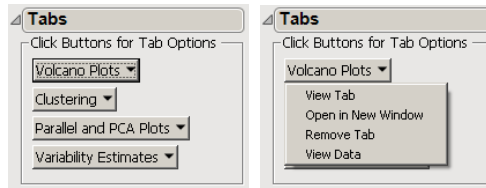


Figure 1.13 The Tabs pane (*left*). Output can be accessed using drop-down menus (*right*).

Data sets can be accessed from the Output SAS Data Sets pane (Figure 1.14).

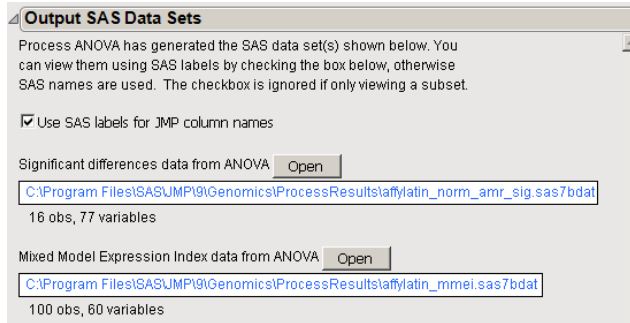


Figure 1.14 The Output SAS Data Sets pane

Finally, you can drill-down into your data or launch follow-up processes from the Action Buttons (Figure 1.15) or Launch Follow-Up Processes (not shown) panes.

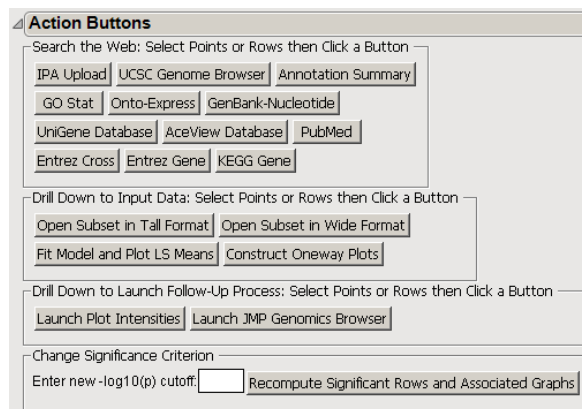


Figure 1.15 The Action Buttons pane

## Stopping a Process

To stop any JMP Genomics process currently running, click **Run** a second time to bring up a Warning:SAS is Running window. Click **Stop** to terminate the process.

---

*Note:* You cannot use the **Stop** option to stop either the [Affymetrix Tiling CEL Input Engine](#) or the [Affymetrix Tiling Bar Input Engine](#). To stop these processes, simultaneously click <Ctrl>, <Alt>, and <Delete> to open the Windows Security window. Click **Task Manager**, click **Processes**, highlight the TilingReader.exe routine and click **End Process**.

---

---

## Saving and Loading Settings

JMP Genomics dialogs enable you to save and load parameter settings. This enables you to save, recall, modify and exchange analyses without having to re-enter specifications each time you run a process. You can save and load settings using the **Save** and **Load** buttons at the bottom of each dialog. Most of the processes in JMP Genomics come with one or more example settings that use the example data sets that come with the system. A good way to learn about a new process is to load one of the example settings, study its parameter values, run the process, and explore the results.

---

## SAS Variable Names and Labels

Each variable/column in a SAS data set must have a unique name. SAS variable names must conform to the following conventions:

- 1 The first character must be a letter (A, B, C, ...) or underscore (\_).
- 2 Subsequent characters can be letters, numeric digits (0,1,2 ...) or underscores (\_).
- 3 Blank spaces are not allowed.
- 4 Special characters, except for underscore, are not allowed.
- 5 Names must not exceed 32 characters.

SAS variable names are not case-sensitive.

SAS variables can be either character or numeric. In either case, a fixed length is assigned to store each observation of that variable.

Optionally, SAS variables can have a *label*. Labels have much less restrictive creation rules. For example, SAS labels can be up to 256 characters in length and can contain blanks and special characters. When JMP opens a SAS data set, it reads the labels (when they exist) and uses them as JMP data table column names.

If you want information about the variable names and labels for a SAS data set, run the **Column Contents** process under the **Data Set Utilities** menu. There are other processes available for changing SAS variable names, labels, and lengths.

## Sample Case Studies

The following data sets, included with JMP Genomics, which are detailed below, enable you to work through most of the analytical processes in JMP Genomics. In addition to the data sets, each case study includes experimental design files and other needed files. These case studies are referred to throughout this manual.

### ***Drosophila* Aging Experimental Data**

This data set represents a small subset of the *Drosophila* aging experiment data from (Jin, Riley *et al.* 2001). The experiment consisted of 24 two-color cDNA microarrays, 6 for each experimental combination of 2 lines (Oregon and Samarkand), 2 sexes (Female and Male), and 2 ages (1 week and 6 weeks). The *Cy3* and *Cy5* dyes were flipped for two of the 6 replicates for each genotype and sex combination. The design is a split-plot, with Age and Dye as subplot factors, and Line and Sex as whole-plot factors. A total of 4256 clones were spotted on the arrays, but this example uses a subset containing 100 randomly selected genes from the original data set.

### **Affymetrix Latin Square Data**

The spike-in data set used in this example was originally generated by Affymetrix Corporation to develop and validate their U95A GeneChip and Microarray Suite (MAS) 5.0 algorithm over a range of known concentrations (Affymetrix, 2001). The experiment consists of 59 arrays. There are 14 experimental groups, designated with letters, a, b, c, d, e, f, g, h, i, j, k, l, m, and q. (Group m and group q each have 4 within-chip replicates, group m replicates were originally designated n, o, and p and group q replicates were originally designated r, s, and t. The extra letters are not needed because they are replicates of m and q, respectively.)

Each experiment was repeated in triplicate using Affymetrix chips cut from different wafers. The last four digits of the wafer numbers are 1521, 1532 and 2353. Wafer 2353, chip c was defective so is not included in the data set. For wafers 1521 and 1532, 20 .cel files were generated, and for wafer 2353, 19 .cel files were generated. Each group contains a pool of non-specific RNA as well as a set of 14 distinct human transcripts spiked in at known concentrations of 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM.

### **Prostate Cancer Biomarkers**

This data set was obtained by surface-enhanced laser desorption/ionization (SELDI). This method allows an investigator to detect and resolve multiple proteins bound to protein chip arrays (Merchant and Weinberger 2000). This approach was used by Qu, *et al.* (2002) to discriminate prostate cancer from non-prostate cancer patients. The promise of this approach is that a panel of multiple biomarkers can be used to distinguish important phenotypes such as cancer status. However, great care must be taken to pre-process and analyze the data appropriately to ensure generalizability of results.

The example data set consists of serum samples collected from 165 men, 84 of whom had prostate cancer. The remaining 81 men are considered to be controls. The primary goal is to determine differences in protein expression between these groups.

## Sample Genetic Marker Data

These data are computer-simulated. The data are in wide form, with the 1000 rows corresponding to individuals and 130 columns corresponding to various data on these individuals. These data contain family, genotype, and phenotype information. The disease column contains the binary trait of primary interest, with 1 indicating individuals affected with the disease and 0 indicating unaffected individuals. There are also four quantitative traits and sixty markers, with two possible alleles (designated 1 and 2), per marker, for each individual. The marker data occur in pairs, so that the genotype at the first marker comprises columns `ma1` and `ma2`, `ma3` and `ma4` the second marker genotype, and so on. The analyses performed on this data set are aiming to locate the gene or genes that affect susceptibility to this disease.

Accompanying this data set is a map data set that provides information about the 60 markers, which are spread across two hypothetical candidate gene regions. The variable representing on which candidate gene the marker resides can be used to group analyses, and the Location variable is useful for accurately displaying distances in base pairs between markers along the  $x$ -axis of plots containing various association  $p$ -values.

## Affected Sib-Pair (ASP) Data

Two hundred families, each containing an affected sib-pair and the siblings' parents, were genotyped at 20 markers from a single chromosome in simulated data provided by Gonçalo Abecasis at the University of Michigan Center for Statistical Genetics. MERLIN was used to estimate identical-by-descent (IBD) allele-sharing probabilities at these markers for all pairs of related individuals. The 400 offspring are also measured for a quantitative trait of interest.

## Additional Data Sets

Some of the examples discussed in this manual make use of data sets, not included with JMP Genomics, nor described here. Where applicable, these additional data sets are described in relevant chapters. We have tried, wherever possible, to use publicly available data sets and, as part of the description, have included instructions on where and how you can obtain these data.

## Using JMP to Design New Experiments

---

JMP offers a wide range of Design of Experiments (DOE) functionality, including classical designs. Most of these functions are grouped under the **DOE** menu (Figure 2.1).

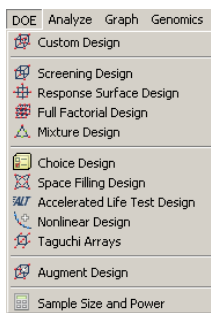


Figure 2.1 JMP's DOE menu

This chapter guides you through some of the features in JMP that plan efficient experiments. For scientific discovery and genomics purposes, we focus only on the first item in this menu: **Custom Design** (Figure 2.1). Refer to the *JMP Design of Experiment Guide* for in-depth details and background on all items in this menu.

---

*Note:* Some of the terminology in the *JMP Design of Experiments* guide is derived from the statistical and engineering literature, which chronicles a long, rich, and successful history of highly efficient experimental designs. Many of the best designs are not widely known or utilized in genomics research, but JMP enables you to rapidly find and customize them for your laboratory's needs.

---

The examples illustrated on the following pages demonstrate how easy it is to use JMP's DOE functions in planning gene expression and other studies.

---

### Example 1: A Two-Way Design for Single Channel Instrumentation

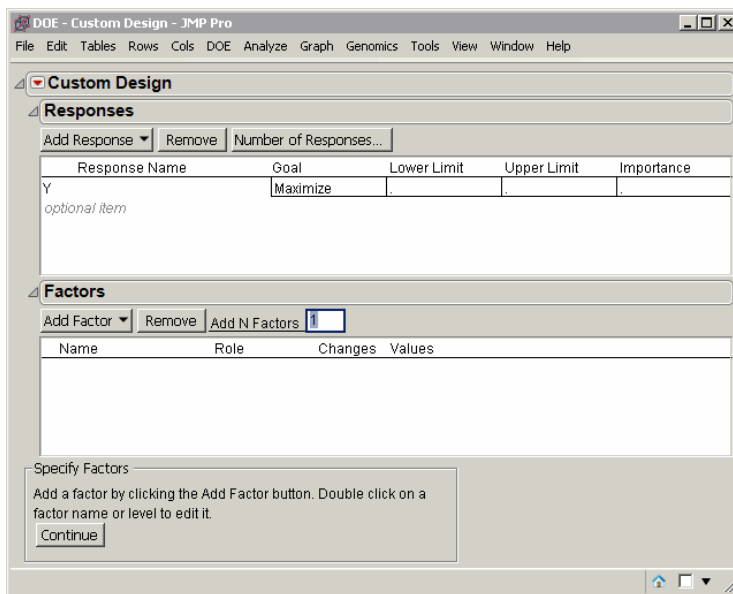
This example uses 12 biological samples to study the effects of a chemical agent versus a chemical control. The study examines the expression of a large set of genes, proteins, or metabolites at 1 hour, 6 hours, and 24 hours after dosing the samples with the chemical. Because of the destructive nature of the

expression protocol, each sample can be treated with only one chemical and observed at only one time. Expression is measured with a single-channel instrument, which excludes two-channel microarrays, considered later in this chapter. A standard two-way design is appropriate in this case.

The JMP Custom Design platform allows you to interactively create designs of any complexity, but let's begin with this simple case.

☞ Select **DOE > Custom Design**.

The dialog shown in **Figure 2.2** opens.



**Figure 2.2** JMP's DOE-Custom Design dialog

The two main fields of the DOE-Custom Design dialog are **Responses** and **Factors**. These fields allow entry of responses and factors specific to your experiment.

*Responses* are the numerical measurements taken during the experiment. In Genomics research, thousands of responses are collected simultaneously, so JMP Genomics has special conventions for loading large response data files. These conventions are explained later.

☞ For now, make no changes to the **Response** field.

*Factors* are the variables that are controlled during the experiment. They are the effects of interest. For our two-way experiment, we have two factors: **Treatment** and **Time**. **Treatment** has two levels: **Agent** and **Control**. **Time** has three levels: **1h**, **6h** and **24h**.

To add the first factor to the design, complete the following steps:

☞ Select **Add Factor > Categorical > 2 Level** using the drop-down menu, as shown in **Figure 2.3**.

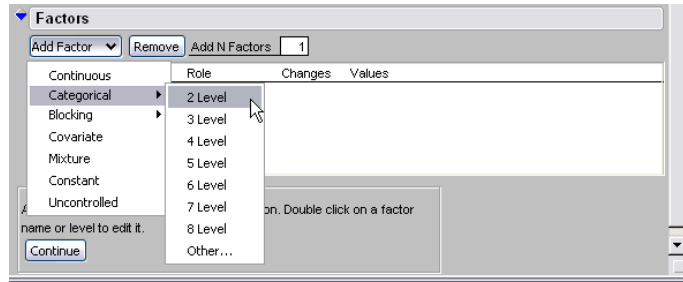


Figure 2.3 Adding a two level categorical factor.

A two-level categorical factor, termed X1, is added (Figure 2.4).

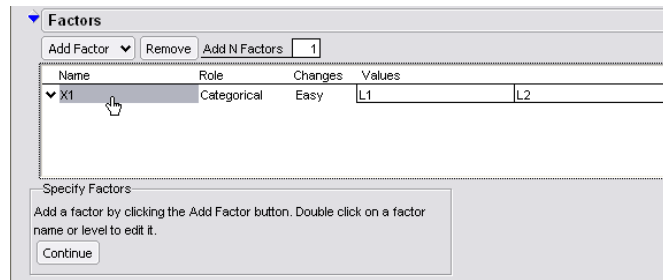


Figure 2.4 The first factor has been added to the design.

To specify the X1 factor as **Treatment**, complete the following steps:

- ☞ Double-click on X1 and type **Treatment**, as shown in Figure 2.4.
- ☞ Double-click on L1 (under **Values**) and type **Agent**.
- ☞ Double-click on L2 (under **Values**) and type **Control**.

To add the second factor to the design, complete the following steps:

- ☞ Select **Add Factor > Categorical > 3 Level** using the drop-down menu.

A three-level categorical factor, termed X2, is added. To specify the X2 factor as **Time**, complete the following steps:

- ☞ Double-click on X2 and type **Time**.
- ☞ Double-click on L1 (under **Values**) and type **01h**.
- ☞ Double-click on L2 (under **Values**) and type **06h**.
- ☞ Double-click on L3 (under **Values**) and type **24h**.

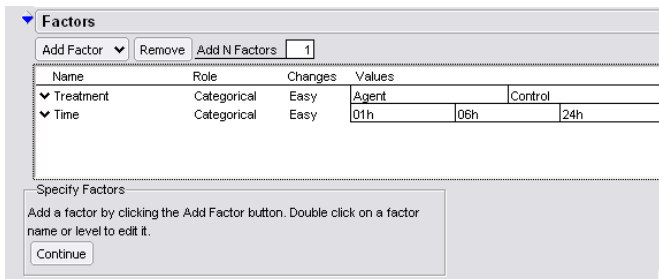
---

*Tip:* Using zero-padding to code numerically-based values with varying lengths ensures alphabetical sorting order matches numerical order during later analytical processing in SAS.

---

*Note:* You could optionally define Time as a Continuous factor if you plan to directly model linear or quadratic trends over time. For this example, we define Time as categorical in order to allow each time level to have an arbitrary mean response.

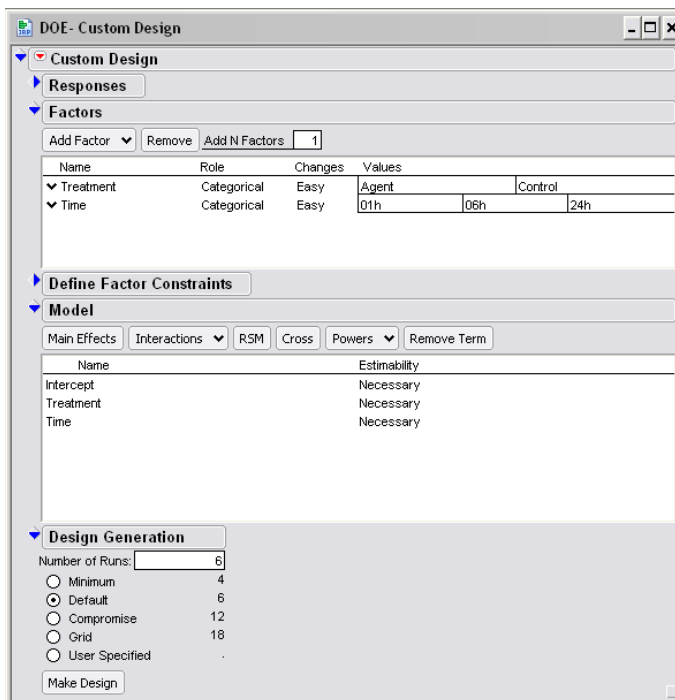
The completed Factors field is shown in **Figure 2.5**.



**Figure 2.5** The completed Factors field

☞ Click **Continue** to proceed to the next design step.

The updated DOE-Custom Design dialog is shown in **Figure 2.6**.



**Figure 2.6** JMP's DOE-Custom Design Dialog window (part II)

There are no constraints to be defined

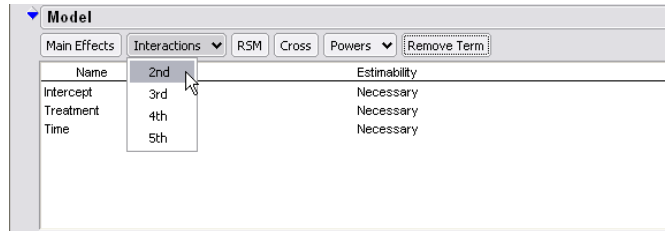
☞ Skip the Define Factor Constraints section.

The **Model** section allows specification of the design that enables estimation of interactions between individual factors, such as **Treatment** and **Time**.

☞ Examine the **Design Generation** section at the bottom of the dialog shown in **Figure 2.6**.

In the current design, with no interaction terms specified, the default number of runs is set to 6. To specify the interaction of factors in the design, complete the following step:

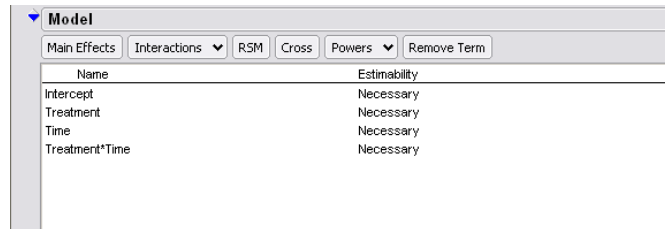
☞ Select **Interactions > 2nd** using the drop-down menu as shown in **Figure 2.7**.



**Figure 2.7** Adding two-level interactions

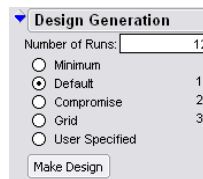
☞ Examine the **Model** section of the dialog.

A **Treatment\*Time** row has been added to the model, as shown in **Figure 2.8**.



**Figure 2.8** The completed model

The default number of runs listed in the **Design Generation** section (**Figure 2.9**) has changed to 12.



**Figure 2.9** The updated Design Generation section

---

*Note:* A *run* is one specific combination of factors applied to obtain one set of responses.

---

☞ Since there are 12 samples budgeted for the runs, leave this field as is and click **Make Design** to generate the design shown in **Figure 2.10**.

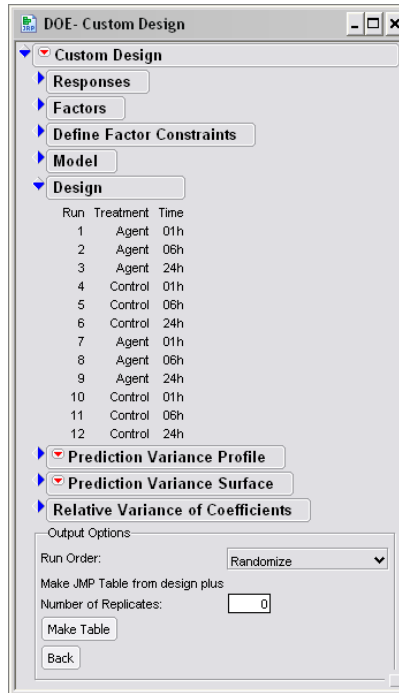


Figure 2.10 JMP's DOE-Custom Design Dialog window (part III)

☞ Examine the DOE-Custom Design Dialog window shown in **Figure 2.10**.

The **Design** section sequentially lists the 12 runs.

Whenever possible, it is always a good idea to randomize the order in which you collect experimental data. This helps avoid any unwanted trends that may creep into the data over time. If you are unable to randomize the order of one or more factors, you should consider more complex designs such as Randomized Block or Split-Plot designs, described later in this chapter. To do this randomization, complete the following steps:

☞ Make sure that **Randomize** is selected in the Run Order drop-down menu in the Output Options box (**Figure 2.11**).

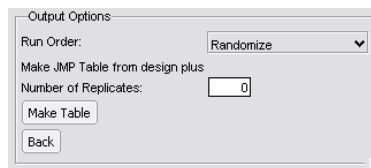


Figure 2.11 The Output Options box

☞ Leave the number of replicates set to 0 because all 12 available samples are used in the design.

☞ Click **Make Table**.

The Experimental Design table shown in **Figure 2.12** is generated.

Design	Treatment	Time	Y	
1	Agent	06h		•
2	Agent	24h		•
3	Control	24h		•
4	Agent	01h		•
5	Agent	06h		•
6	Control	06h		•
7	Agent	01h		•
8	Control	24h		•
9	Control	01h		•
10	Agent	24h		•
11	Control	01h		•
12	Control	06h		•

Figure 2.12 The Experimental Design table

The treatment to which each biological sample is subjected and the time at which each sample is observed are both listed in Figure 2.12. Note that the run order (1-12) may be different than Figure 2.12 because of the random number generator used to generate this design.

This is an example of a completely randomized design. The levels of both Treatment and Time are arranged in a random order.

When collecting expression data on only one gene, protein, or metabolite, simply enter the data in the Y column and then analyze them directly in JMP using any number of different methods. But to work with thousands of expression measurements simultaneously, JMP Genomics requires you to construct a table like this one as a way to link the experimental design information to a collection of raw response data files, each of which contains thousands of measurements. Construction of this table, known as an *Experimental Design File*, requires adding two columns to this table, called File and Array, that are described more fully in [The Experimental Design File \(EDF\)](#). The File column lists the names of the raw data files containing the expression measurements corresponding to the factor levels for the run in its same row. The Array column contains a unique index for each array in the experiment.

For now, the table is ready to use in the lab to run the design in the random order specified.

## Example 2: Incorporating Blocking Factors in a Designed Experiment

Experimental designs are often difficult to conduct in a completely randomized fashion because of the presence of one or more additional factors that can induce correlation in the observed responses. In these situations you should define one or more blocking factors to better control unwanted experimental variation. Examples of blocking factors include: batch, animal, day of processing, technology lot number, machine, location, laboratory, technician, or operator. Blocking factors are typically considered random because they can be viewed as arising from a population of effects having a probability distribution, usually a normal distribution.

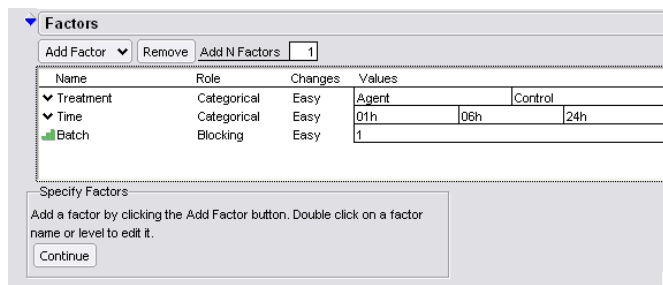
To continue the two-way design example, illustrated above, suppose that the 12 samples are not totally independent, but that 3 samples each were taken from 4 distinct batches. The batches could consist of any number of things, including the day of initial sample collection or the mode of processing them. In

this case, it is important to control for the effect of batches on the experimental outcomes. To do this, add **Batch** as a blocking factor to the design. Defining such a blocking factor lets you model a correlation between samples from the same batch and provides a more accurate assessment of true batch-to-batch variability. Ignoring the batch effect when it is significant leads to biased conclusions about expression differences.

To add **Batch** to the previous design, complete the following steps:

- ☞ Select **DOE > Custom Design** to begin a new design.
- ☞ Define **Treatment** and **Time** factors, as previously described.
- ☞ Click **Factor > Blocking > 3 runs per block**. Double-click on X3 and change it to **Batch**.

The **Factors** section should now appear as shown in **Figure 2.13**.



**Figure 2.13** Design with a blocking factor

- ☞ Click **Continue** to specify which terms need to be modeled.
- ☞ Click **Interactions > 2nd**.
- ☞ Click **Continue** in any message windows.
- ☞ Click **Make Design** to generate the Custom Design shown in **Figure 2.14**.

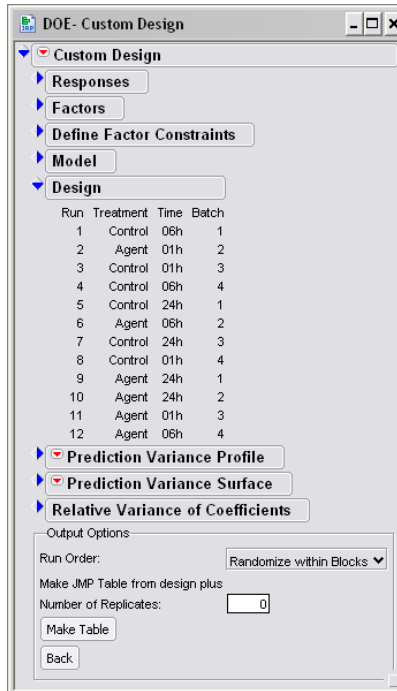


Figure 2.14 Custom Design dialog with 1 blocking factor

☞ Examine the Custom Design dialog shown in Figure 2.14.

Note that the Batch factor has four levels, with three runs for each level.

☞ Click **Make Table** to generate the Experimental Design table shown in Figure 2.15.

Run	Treatment	Time	Batch	Y
1	Control	06h	1	*
2	Control	24h	1	*
3	Agent	24h	1	*
4	Agent	06h	2	*
5	Agent	01h	2	*
6	Agent	24h	2	*
7	Agent	01h	3	*
8	Control	01h	3	*
9	Control	24h	3	*
10	Control	06h	4	*
11	Agent	06h	4	*
12	Control	01h	4	*

Figure 2.15 Experimental Design table (with 1 blocking factor)

This is an example of an *Incomplete Block Design*. The blocks corresponding to **Batch** are incomplete because not all combinations of treatment and time are observed within a block; however, there is a form of partial balance in the experiment, because each unique combination of treatment and time is observed exactly twice across the whole experiment.

---

*Note:* Good designs often have some form of balance in terms of number of treatment combinations observed. Balancing the number of factor levels helps break confounding among factors and ensures approximately equal information gain on all relevant differences.

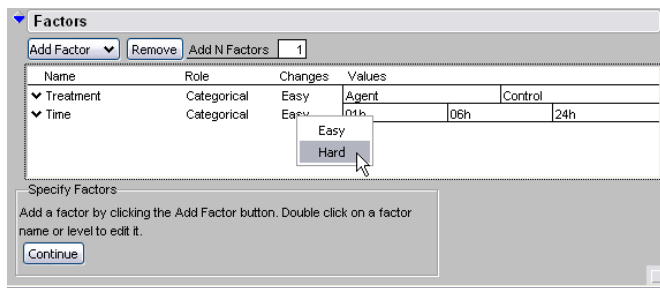
---

### Example 3: Generating a Split-Plot Design

Continuing our two-way experiment example with factors Treatment and Time on a single-channel instrument, suppose that instead of the need for the batch blocking factor, the actual constraint is that samples need to be processed immediately after collection at the 1 hour, 6 hour, and 24 hour time points. In other words, it is not feasible to conduct the experimental runs in a completely randomized fashion; rather, they must be processed in time order.

This is a situation calling for a *Split-Plot Design*, in which certain factors are easy to change in the lab and others are hard to change. You can easily generate a split-plot design in JMP DOE by changing values in the Changes column in the Factors section.

- ☞ Select **DOE > Custom Design** to begin a new design.
- ☞ Define Treatment and Time factors, as previously described.
- ☞ In the Changes column, click Easy in the Time row.
- ☞ Select Hard from the menu that appears, as shown in **Figure 2.16**.



**Figure 2.16** Changing Time from Easy to Hard

- ☞ Click **Continue** to define the model.
- ☞ Click **Interactions > 2nd** in the Model section.
- ☞ Click **Make Design** to proceed to the next step (shown in **Figure 2.17**).

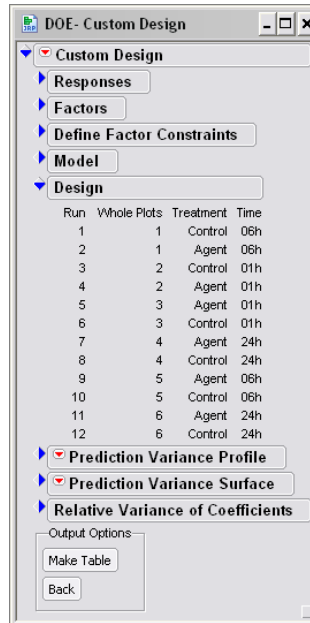


Figure 2.17 DOE-Custom Design dialog (split-plot design)

Notice the automatic creation of the Whole Plots column in the Design section.

---

*Note:* The term *Whole Plot* derives from agricultural field research where split-plot designs were originally popularized. Imagine a two-way design in a field trial in which the effects of plant variety and different fertilizers are to be studied. The fertilizers can only be applied to large sections of the field via large machinery or airplanes, but varieties can be planted in smaller sections. The split-plot design consists of dividing the field into fertilizer-level sections called *Whole Plots*, and the varieties are planted in subplots within each whole plot.

---

There are six whole plots in this design. Note that levels of Time are constant within any particular whole plot. In contrast, the Agent and Control levels of Treatment change within whole plots. Treatment is known as a *subplot factor*, and Time as a *whole-plot factor*. By their nature, split-plot designs provide more precision in estimating effects of subplot factors than they do for effects of whole-plot factors. This is perhaps intuitive given the constraints placed on the whole-plot factors.

Many experimenters employ split-plot designs without realizing it when they process samples in a grouped order, but then analyze the data as if they were completely randomized. This practice can lead to badly biased conclusions, especially when the whole-plot effect is substantial. The appropriate way to analyze a split-plot design involves specifying whole-plots as a random effect in the analysis, thereby modeling a correlation among measurements taken within the same whole plot.

---

## Example 4: Split-plot Design for Two-Channel Microarrays

Two-channel microarrays are characterized by the fact that two measurements for each gene are obtained from each microarray. This is because two different samples are tagged with different dyes, competitively hybridized to one array, and then measured under two different laser frequencies. This technology therefore offers an additional layer of complexity for experimental design beyond the one-channel designs described previously.

Several papers have discussed different two-channel design options in detail, including Kerr and Churchill (2001) and Dobbin and Simon (2002). Arguably the most popular design is the *Reference Sample Design*, in which a common reference sample (typically a pool of samples that is not of direct experimental interest) is tagged with one dye and hybridized on every array, while the various treated samples are tagged with the other dye. This design is easy to set up and effectively reduces design considerations to the single channel that is changing.

However, the reference sample design can be two to four times less efficient than designs that hybridize samples of interest directly together on microarrays. The keys to higher efficiency are to pair samples together on arrays in a way that optimizes experimental interests and then to make sure the analysis of the data is conducted appropriately.

The previous discussion of blocking factors and split-plot designs has direct bearing here. If we narrow our focus to all the data from a single gene, and assume there is only one spot for that gene on each array, then the data come in pairs corresponding to the two measurements from each array. Each array can therefore be considered as a block of size two. Alternatively, in a split-plot scenario where certain factors are hard to change, you desire more precise information on some factors versus others. In these situations, you can consider arrays to be whole plots and assign certain factors to change within whole plots (subplot factors) and others to stay constant on the whole plots (whole-plot factors).

Here we use the *Drosophila* aging experiment described by Jin *et al.* (2001) as an example to consider for experimental design options for two-channel microarrays. A subset of these data is included with your JMP Genomics installation and is described in [Introduction to JMP Genomics](#). This design has three experimental factors with two levels each: Age (1 week, 6 weeks), Sex (Female, Male), and Line (Oregon, Samarkand).

---

*Note:* For higher-level factorial arrangements, experimental design experts often use exponential notation as a shorthand description. The *Drosophila* example would be called a  $2^3$  design, which designates 3 factors with 2 levels each.

---

The primary experimental factor of interest is **Age**, and for this experiment it was desirable to obtain more precise information on the effects of **Age** at the expense of the **Sex** and **Line** effects. The latter two are still included to provide a higher degree of generalization for conclusions. These considerations call for a split-plot design.

☞ Select **DOE > Custom Design**.

☞ Define the three categorical factors and a fourth factor indicating the **Channel**. Specify **Sex** and **Line** as **Hard** in the **Changes** column, and leave **Age** and **Channel** as **Easy**.

The completed dialog should appear as shown in **Figure 2.18**.

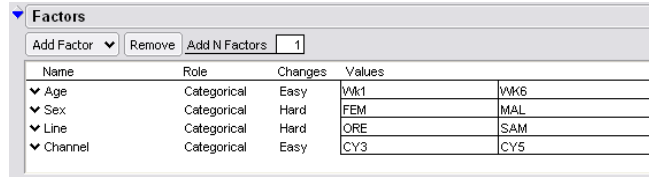


Figure 2.18 The factors for the split-plot design have been specified

There are 24 assays available for experimentation.

☞ Specify 24 in the Number of Whole Plots box in the Design Generation section (Figure 2.19).

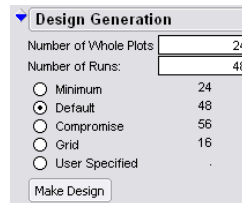


Figure 2.19 The completed Design Generation section of the dialog

☞ Click **Make Design** and then **Make Table** to generate the table shown in Figure 2.20.

	Whole Plots	Age	Sex	Line	Channel	Y
1	1	Wk6	FEM	ORE	CY3	.
2	1	Wk1	FEM	ORE	CY5	.
3	2	Wk6	FEM	SAM	CY5	.
4	2	Wk1	FEM	SAM	CY3	.
5	3	Wk1	FEM	SAM	CY5	.
6	3	Wk6	FEM	SAM	CY3	.
7	4	Wk1	MAL	ORE	CY3	.
8	4	Wk6	MAL	ORE	CY5	.
9	5	Wk1	FEM	ORE	CY5	.
10	5	Wk6	FEM	ORE	CY3	.
11	6	Wk6	FEM	ORE	CY5	.
12	6	Wk1	FEM	ORE	CY3	.
13	7	Wk1	MAL	ORE	CY5	.
14	7	Wk6	MAL	ORE	CY3	.
15	8	Wk1	MAL	SAM	CY3	.

Figure 2.20 The Experimental Design table

Note how Age and Channel change within whole plots, whereas Sex and Line stay constant for each whole plot.

To convert this table to a valid JMP Genomics *Experimental Design File* (EDF), change the name of the Whole Plot column to Array by double-clicking on the column header and typing in Array as the new column name. Also, delete the Y column, since it will be replaced by a column named File.

To compare this design with the original design in Jin *et al.* (2001), open the AgingExperimentTable.txt This file, which is included with JMP Genomics, is located in the Sample Data folder. Note the run order and randomization schemes are different, but the designs are similar in terms of their split-plot structure.

## Example 5: Randomized Block Design for Two-Channel Microarrays

Suppose that instead of the split-plot design just considered, equal information about the Age, Sex, and Line factors is needed and they need to be randomly allocated to the arrays in a randomized block design. A somewhat different approach in JMP illustrates a few more of its features.

☞ Select **DOE > Custom Design**.

☞ Define Dye as a 2-level categorical factor and Array as a 2-runs-per-block blocking factor as shown in Figure 2.21.

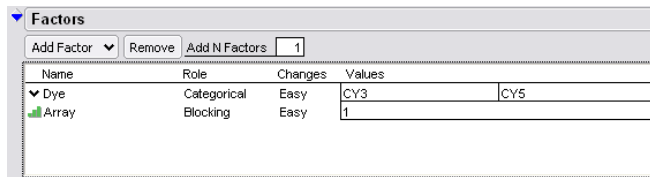


Figure 2.21 The completed Factors box

☞ Click **Continue**.

☞ Enter 48 runs in the Design Generation box as shown in Figure 2.22.

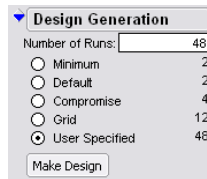


Figure 2.22 The completed Design Generation box

Click **Make Design** and then **Make Table** to generate the table shown in Figure 2.23.

Design	Custom Design	Dye	Array	Y
1	CY5	1		
2	CY3	1		
3	CY5	10		
4	CY3	10		
5	CY5	11		
6	CY3	11		
7	CY5	12		
8	CY3	12		
9	CY3	13		
10	CY5	13		
11	CY5	14		
12	CY3	14		
13	CY3	15		
14	CY5	15		
15	CY5	16		

Figure 2.23 The Experimental Design table

The experimental design table establishes the static portion of the design and ensures that Cy3 and Cy5 always appear once in each array.

- ☞ Make sure that the experimental design table is open and in focus.
- ☞ Select **DOE > Custom Design** to open a new DOE-Custom Design dialog.
- ☞ Click **Add Factors > Covariate** in the Factors section
- ☞ Select Dye, and click **OK**, as shown in Figure 2.24.

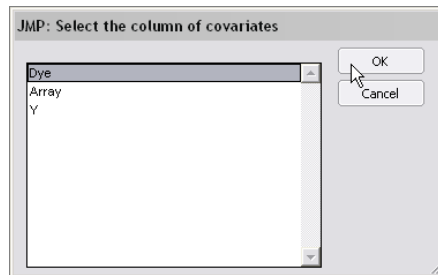


Figure 2.24 Selecting the first covariate

- ☞ Click **Add Factors > Covariate** again.
- ☞ Select Array, and click **OK** to generate the Factors section shown in Figure 2.25.

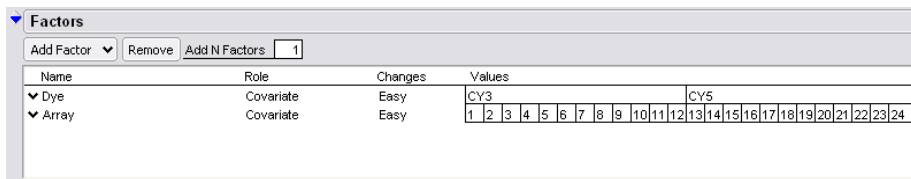


Figure 2.25 Both covariates have been specified

*Note:* JMP considers a covariate to be a factor describing fixed characteristics of the samples that do not change. Also note the levels of the two covariates Dye and Array are automatically read from the active JMP table because it is a previously created JMP table. In addition to loading factors from an active JMP table, you can save and load factors by clicking on the small red triangle beside Custom Design.

Next, we must define the three experimental factors Age, Sex, and Line. Since all three of these factors have two levels, they can be added to the design at the same time.

- ☞ Type 3 in the Add N Factors box.
  - ☞ Click **Add Factor > Categorical > 2 Level**.
- Three new rows are generated in the Factors section of the dialog.
- ☞ Change each row to match the Factors section shown in Figure 2.26.

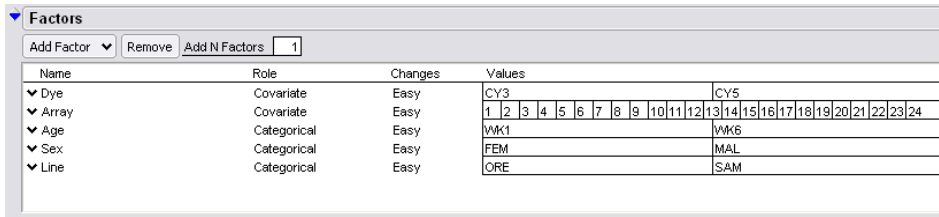


Figure 2.26 The completed Factors section

- ☞ Click **Continue**.
- ☞ Highlight Age, Sex, and Line in the Factors section.
- ☞ Select **Interactions > 3rd** in the Model section to produce the Model section shown in Figure 2.27.

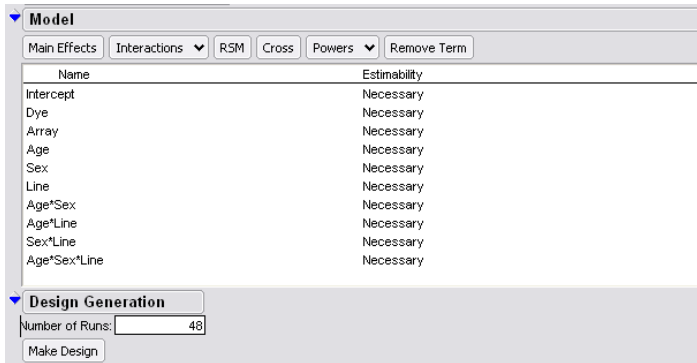


Figure 2.27 The Model section with all factors and interactions defined

- ☞ Click **Make Design** and then click **Make Table** to generate the Experimental Design table shown in Figure 2.28.

Custom Design		Dye	Array	Age	Sex	Line	Y
Design	Custom Design						
Criterion	D Optimal	1	CY5	1	Wk1	MAL	ORE
Model		2	CY3	1	Wk6	MAL	SAM
Columns (6/0)		3	CY5	10	Wk1	MAL	ORE
Dye *		4	CY3	10	Wk6	FEM	SAM
Array *		5	CY5	11	Wk6	FEM	ORE
Age *		6	CY3	11	Wk6	MAL	ORE
Sex *		7	CY5	12	Wk6	FEM	SAM
Line *		8	CY3	12	Wk1	FEM	SAM
Y *		9	CY3	13	Wk1	FEM	ORE
Rows		10	CY5	13	Wk1	MAL	SAM
All rows	48	11	CY5	14	Wk1	FEM	ORE
Selected	0	12	CY3	14	Wk6	MAL	SAM
Excluded	0	13	CY3	15	Wk6	MAL	SAM
Hidden	0	14	CY5	15	Wk1	MAL	SAM
Labelled	0	15	CY5	16	Wk6	MAL	SAM

Figure 2.28 The randomized block design

This randomized block design allocates 2 of the 8 possible treatment combinations to each array. This previous design is also known as a kind of loop design (Kerr and Churchill, 2001), and is illustrated in Figure 2.29.

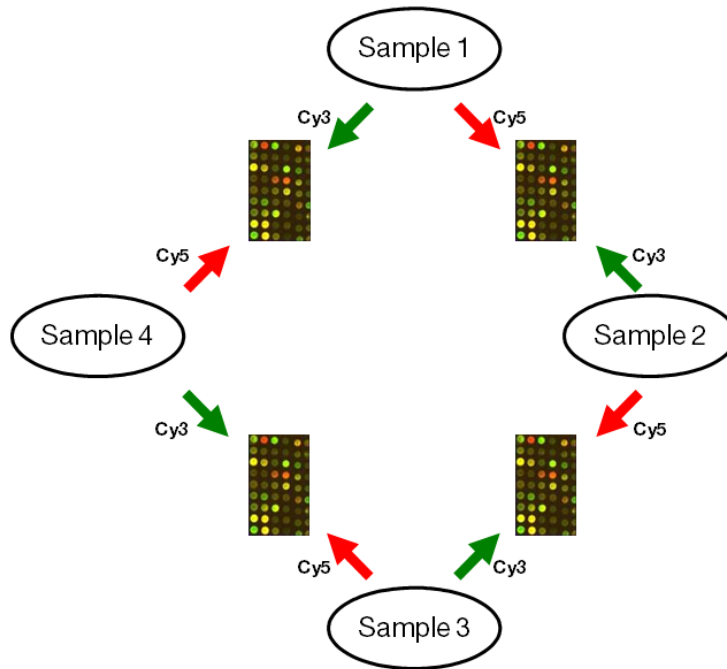


Figure 2.29 Loop design for 4 experimental conditions

The term *loop* derives from the fact that the design can be depicted as nodes indicating samples treated with one particular experimental factor combination. Aliquots of RNA from each sample are labeled either with the CY3 (green) or CY5 (red) florescent dyes. Two labeling reactions are required for each sample. Pairs of alternately labeled samples are pooled and hybridized to identical arrays. Each spot is probed with each sample, labeled with either dye, allowing the experimenter to control for confounding biases resulting from either dye or array effects.

## Microarrays with Three or More Channels

With microarrays having three or more channels, the previous discussion for two-channel designs can be extended. For incomplete block designs, set the number of runs per block equal to the number of channels and set up the other factors as usual. For split-plot designs, set the number of whole plots equal to the number of budgeted arrays.

---

## Microarrays with More than One Spot per Gene on Each Array

Some microarrays, often those manufactured in your local lab, have multiple spots per gene on the array. Such two-color arrays pose no additional concerns from a new experimental design perspective because the samples are applied to the entire array. However, the existence of multiple spots does make a difference during subsequent data analysis, when random effects caused by such things as the nesting of identical spots within an array or differences in dye effects among multiple arrays should be considered, in addition to the usual Array random effect.

---

## Choosing the Overall Number of Runs in a Design

Selecting the number of runs in a design is always a trade-off between cost of the experiment versus the desired information gain, precision, or power. The latter can be difficult to quantify, considering that tens of thousands of genes or proteins are measured simultaneously.

One rule of thumb is to use three biological replicates for each distinct combination of factors. A biological replicate is a biologically unique sample from the population of samples considered for experimentation. This is to be distinguished from a technical replicate, which is a repetitive measurement from biological material already used in a previous run. Biological replicates tend to be much more variable than technical replicates, but they also provide the best means to make appropriate conclusions about the population of interest.

A more statistical concept for evaluating size of designs is *degrees of freedom for error*. This represents the fraction of the data that is used to estimate noise instead of signal. It is computed by subtracting the total number of factor combinations from the total number of runs. Another rule of thumb requires at least 10 degrees of freedom for error in the design in order to be able to obtain an accurate estimate of noise and accompanying standard errors for effect differences.

A rigorous statistical approach for determining the number of replicates in a design is to use sample size and power calculations. These require some prior knowledge about anticipated magnitudes of effect sizes as well as desired false positive rates. Some common methods are available under **DOE > Sample Size and Power**, and a few advanced ones are under **Genomics > Power and Sample Size**. Refer to the *JMP Design of Experiments Guide* for additional information.

## Configuring JMP Genomics Settings

---

### What does it do?

The Configure Genomics Settings process enables you to define the specific parameter settings and defaults used by JMP Genomics processes, to configure JMP Genomics for grid computing, and to specify the name of the proxy server through which your organization can access with the Internet.


---

*Caution:* Changing the default settings can potentially impact the performance and outcome of JMP Genomics processes. You should contact your information systems administrator before running this process.

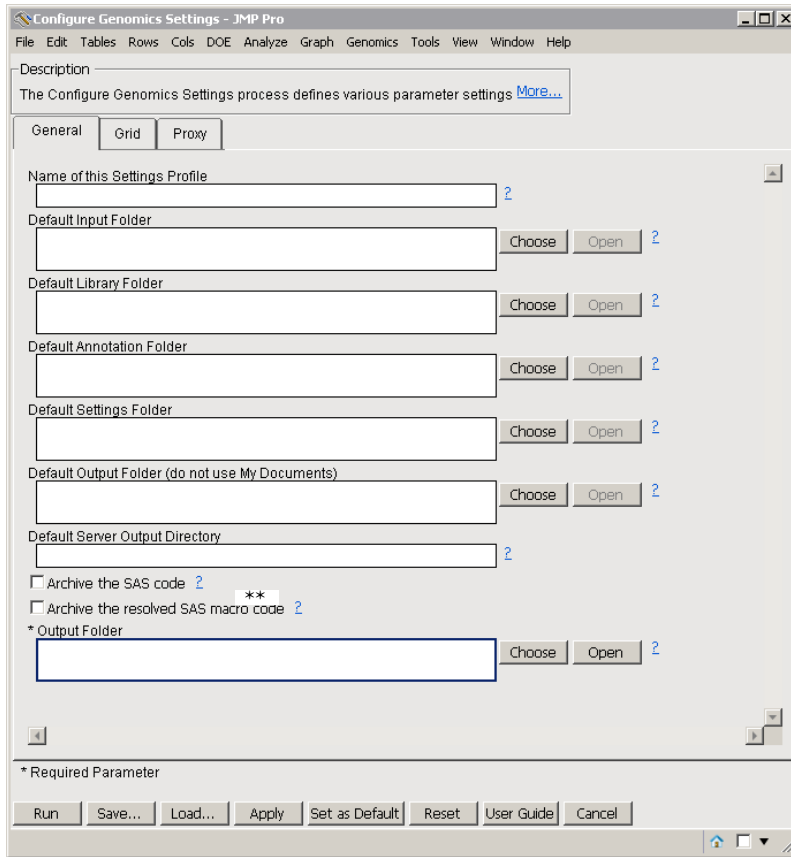
---

---

### The Dialog

The dialog for the Configure Genomics Settings process consists of three tabs **General**, **Grid**, and **Proxy**. The fields and buttons on each tab are described. The  buttons provide additional information for specific features. Fields and selections in which input is required for the AP to run successfully are labeled (*required*).

The **General** tab, shown in **Figure 3.1**, is used to specify the names and locations of input, output and settings files and folders, data files and the folder where all of the output will be placed. You have a great deal of latitude in specifying the names and locations of the default folders and you are free to choose the names and locations that best serve your needs.



**Figure 3.1** The General tab

#### Name of this Settings Profile field

- Use this field to specify a name by which this profile is saved. You can subsequently load this setting by clicking **Load**.

#### Default Input Folder field

- Use this field to specify a default location for input data sets.
- Either click **Choose** to browse to the folder or type the path to and name of the folder in the field.

#### Default Library Folder field

- Use this field to specify a default location for library files needed by various input engines.
- Either click **Choose** to browse to the folder or type the path to and name of the folder in the field.

#### Default Annotation Folder field

- Use this field to specify a default location for annotation data sets.
- Either type the path to and name of the folder in the field or click **Choose** to browse to the folder.

#### Default Settings Folder field

- Use this field to specify a default location in which to save settings.
- Either type the path to and name of the folder in the field or click **Choose** to browse to the folder.
- *Note:* You must have write permissions for this folder.

#### Default Output Folder field

- Use this field to specify a default location in which to save all of the output from various JMP Genomics processes.
- Either type the path to and name of the folder in the field or click **Choose** to browse to the folder.
- *Note:* You must have write permissions for this folder.
- *Note:* You should not specify the My Documents folder as the output folder, although you can specify subfolders within My Documents.

#### Default Server Output Directory field

- Use this field to specify a default location on a server in which to save all of the output from various JMP Genomics processes. This default applies only when you are connected to a SAS metadata server.
- Type the path to and name of the folder in the field or click **Choose** to browse to the folder.
- *Note:* You must have write permissions for this folder.

#### Archive the SAS code check box

- This option enables you to generate a single SAS program containing all of the SAS code used every time you run a process that uses SAS code.
- A new file is generated every time you run a process that uses SAS code. The resulting SAS program is deposited in the specified output folder.

#### Archive the resolved SAS macro code check box

- This option enables you to generate a single SAS program containing all of the resolved SAS macro code used.
- A new file is generated every time you run a process that uses SAS code. The resulting SAS program is deposited in the specified output folder.
- *Caution:* Checking this option can substantially increase computational times.

Output Folder field (*required*)

- The field shows the path to the folder where the output data set is to be placed.
- The **Choose** button enables you to browse to the selected folder.
- *Note:* The default output folder is C:\Documents and Settings\\*\*\Local Settings\Application Data\JMPG\, where \*\* is your user ID.

The Grid tab, shown in **Figure 3.2**, is used to configure the JMP Genomics for grid computing. You must specify all of the parameters on this tab for any to be saved.

**Figure 3.2** The Grid tab

---

*Note:* To run JMP Genomics on a grid you must have access to and authorization for a SAS metadata server running SAS Grid Manager software. You should contact your information systems administrator for assistance in specifying the grid settings if you plan to run JMP Genomics on a grid.

---

**Metadata Server Name field**

- Use this field to specify the name of the SAS metadata server running SAS Grid Manager software.

**Port Number field**

- Use this field to specify the port number for the SAS metadata server.

**Repository Name field**

- Use this field to specify the name of the SAS metadata repository.

**User Name field**

- Use this field to specify the name of a user authorized to access the SAS metadata server.

**Password field**

- Use this field to specify the password associated with the authorized user name specified above.
- You can leave this field blank. If left blank, you must enter the password each time you run a JMP Genomics process on the grid.

### Shared Grid Folder field

- Use this field to specify the path to a shared folder for which all nodes on your grid have read and write access. The pathname for this folder must be recognizable by all nodes. A UNC path specification (for example, \\machine\folder\subfolder) is recommended.
- *Caution:* If the shared path is on a Windows XP machine, then you are only able to use a maximum of 8 grid nodes because of a limitation on the number of users that can simultaneously access a shared folder. If you want to use more than 8 grid nodes, specify a shared path on a server.

The Proxy tab, shown in **Figure 3.3**, is used only if your organization accesses the Internet through a proxy server. Use this tab to specify the name and your proxy port number. You must specify both parameters for this setting to be used. You should contact your information systems administrator for the name and port number for your proxy server.

**Figure 3.3** The Proxy tab

### Proxy Server Name field

- Use this field to specify the name of your proxy server. Type the name in the form `server-name.domain.com`.
- If your computer does not access the Internet through a proxy server, you must leave this field empty (Blank).

### Proxy Port Number field

- Use this field to specify the proxy port number.
- If your computer does not access the Internet through a proxy server, you must leave this field empty (Blank).

The action buttons at the bottom of the dialogs are described in [Introduction to JMP Genomics](#).

---

## Results

- ☞ Specify all of the parameters that need to be reconfigured.
- ☞ Click **Run**.

A new settings file is generated and placed in the specified folder. All affected parameters are set to the new defaults..



## Using JMP Genomics in Client/Server Mode

---

JMP Genomics was originally designed to run as a stand-alone desktop application. Both the JMP and SAS components are installed on a single machine and all analytical procedures and computations are performed on that one machine. Even though desktop PCs are becoming increasingly powerful, it might make sense to use a remote SAS server to alleviate the computational load on your local machine, especially for CPU- and I/O-intensive calculations.

With the release of JMP Genomics 4.1, the ability to operate over a client/server configuration has been added experimentally to JMP Genomics. In this mode, SAS is installed on a server while the JMP components are installed on one or more client machines. Data files are either stored on the server or are copied to the server from the client during the execution of the analytical process. Jobs set up on the client are sent to the server where they are run; results are surfaced back to the client. The advantages to this configuration are numerous: server machines typically have greater computing capabilities than the typical desk top machine and jobs can run in much less time; hard disk space is usually much more plentiful; the analytical procedures, which can tie up considerable amounts of resources, are carried out elsewhere, thus freeing up desktop machines for other purposes; and, finally, multiple users can use the same SAS installation.

Instructions for installing and configuring JMP Genomics in client/server mode are provided below. If you are not comfortable setting this up on your own, you should contact your system administrator for assistance.

---

*Note:* The addition of client/server functionality does not affect the ability of JMP Genomics to run as a stand-alone desktop application. However, the SAS server-side components must be licensed separately. The required products on the server are: Base SAS, Integration Technologies, SAS/STAT, SAS/Genetics, SAS/GRAPH, SAS/IML, and Access to PC File Formats.

---

---

### Prerequisites

In order to run JMP Genomics in Client/Server mode, you must have access to fully functional SAS 9.2 metadata server. You also need write access to a directory on the server machine. This folder is used to hold the results of analysis runs performed on the server.

---

*Note:* If you want to keep the results for different runs in separate folders on the server, change the server output path prior to each run.

---

Consult your systems administrator to obtain the necessary authentication credentials for accessing the metadata server.

## Configuring JMP Genomics to Run in Client/Server Mode

Before you can set up your metadata server profile, you must configure JMP Genomics to run in client/server mode.

First, on the server machine, you need a working directory to which you have write permissions. This directory will be used by JMP Genomics both for copying over all required files and as a work space during calculations. You should use network tools, such as Telnet or ssh, for example, to connect to a unix-based machine and create a directory.

*Note:* You should contact your system administrator if you need assistance in setting up this directory.

Next, on your client machine, you must save the path to the directory on the server in JMP Genomics.

☞ Open JMP Genomics.

☞ Select **File > Configure Genomics Settings** to open the dialog shown in Figure 4.1.

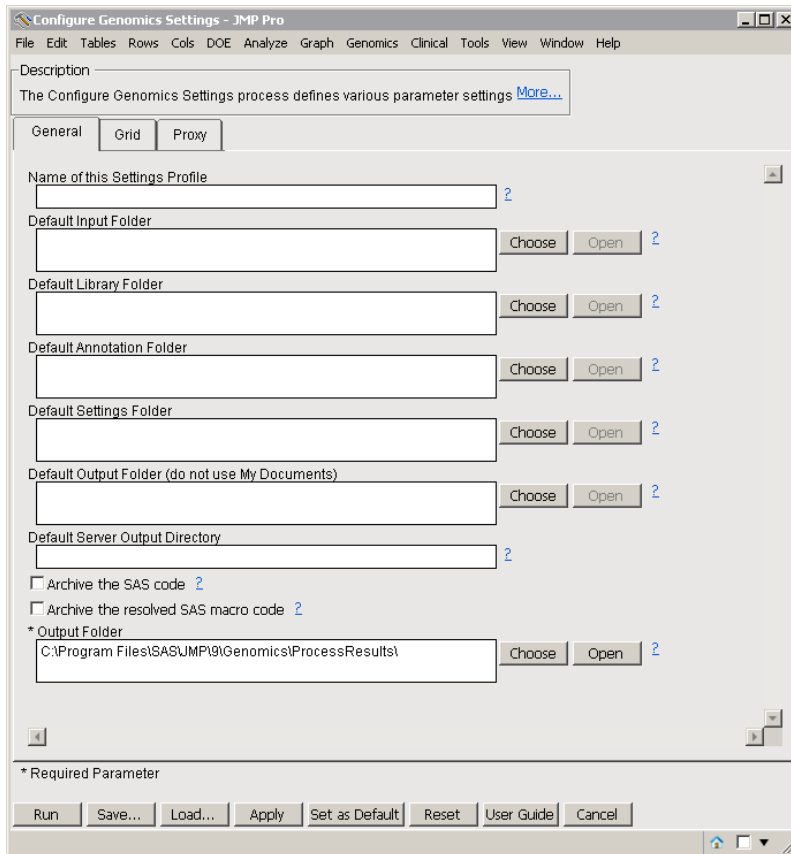


Figure 4.1 The Configure Genomics Settings dialog

☞ Type the name of the directory created above in the Default Server Output Directory field, as shown in Figure 4.2.

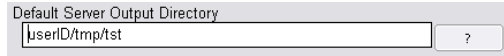


Figure 4.2 The completed Default Server Output Directory field

*Note:* Running Configure Genomics Settings with a Default Server Output Directory specified results in this directory name being automatically filled in whenever you run in experimental client/server mode and open a new dialog.

- ☞ Click **Run**.
- ☞ Close the Configure Genomics Settings dialog and message windows.

## Setting Up Your Metadata Server Profile and Connecting to the Metadata Server

Before you can connect to the SAS metadata server, you must set up a metadata server profile.

- ☞ Select **File > SAS > Server Connections** to open the SAS Server Connections window shown in Figure 4.3.

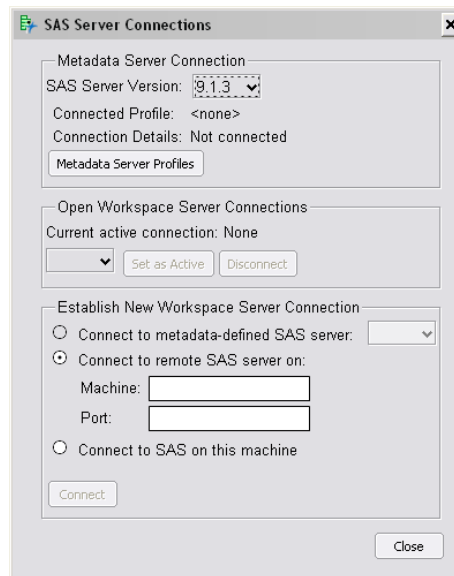
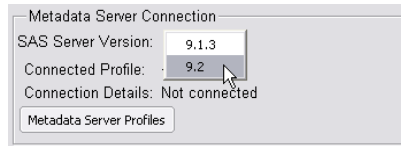


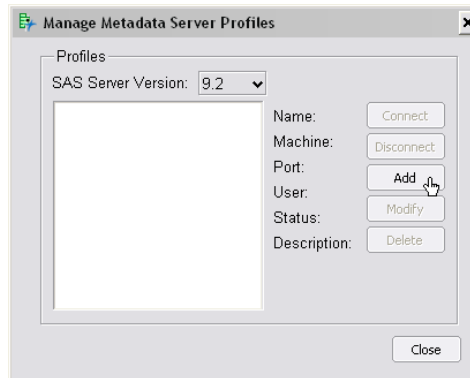
Figure 4.3 The SAS Server Connections window

- ☞ Select SAS version 9.2 from the SAS Server Version drop-down menu shown in Figure 4.4.



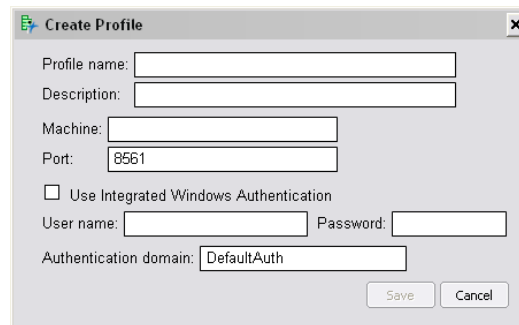
**Figure 4.4** Selecting SAS version 9.2

- Click **Metadata Server Profiles** to open the Manage Metadata Server Profiles window shown in **Figure 4.5**.



**Figure 4.5** The Manage Metadata Server Profiles window

- Click **Add** to open the Create Profile window shown in **Figure 4.6**.



**Figure 4.6** The Create Profile window

- Specify a name and description for your metadata server profile in the Profile name field.
- Type the name of the server in the Machine field.

---

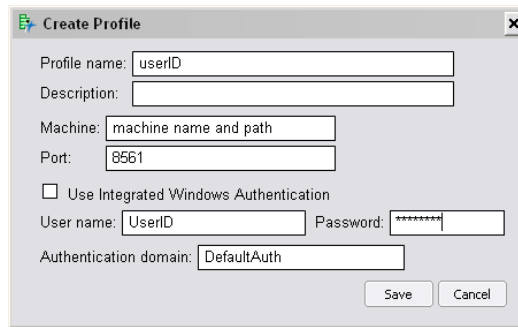
*Note:* The name of the server should contain no spaces. Individual identifiers are usually separated by dots (MachineName.unx.sas.com, for example).

---

- Type the port number in the Port field
- Type your authenticated user name and password for the server in their respective fields.

*Note:* If you do not have an authenticated user ID and password, contact your system administrator.

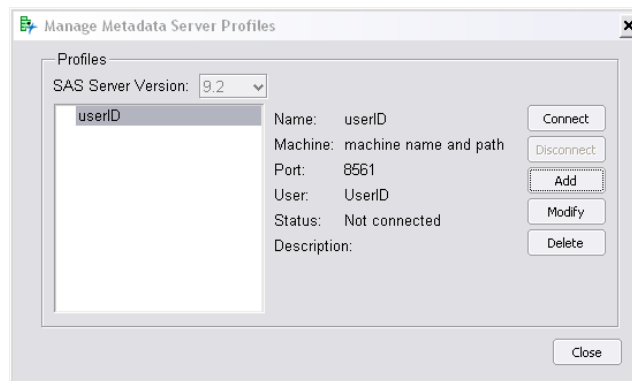
The completed Create Profile window appears as the example shown in **Figure 4.7**.



**Figure 4.7** The completed Create Profile window

Click **Save** to save your profile and close the Create Profile window.

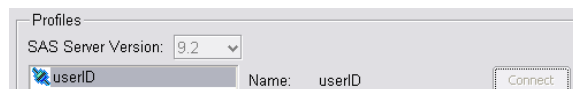
The completed Manage Metadata Server Profiles window appears as shown in **Figure 4.8**.



**Figure 4.8** The completed Manage Metadata Server Profiles window

Click **Connect**.

A blue *Connection Belt* symbol (**Figure 4.9**) appears next to your profile name when you have successfully connected to the metadata server. If it does not appear, modify your profile accordingly or contact your system administrator for assistance.



**Figure 4.9** A successful connection has been established.

Close the Manage Metadata Server Profiles and SAS Server Connections windows.

You are now in Client/Server mode. As long as the connection is maintained, all JMP Genomics APs will run SAS on the metadata server.

To return to standard local mode:

- ☞ Select **File > SAS > Server Connections**.
- ☞ Click **Metadata Server Profiles** and click **Disconnect**.

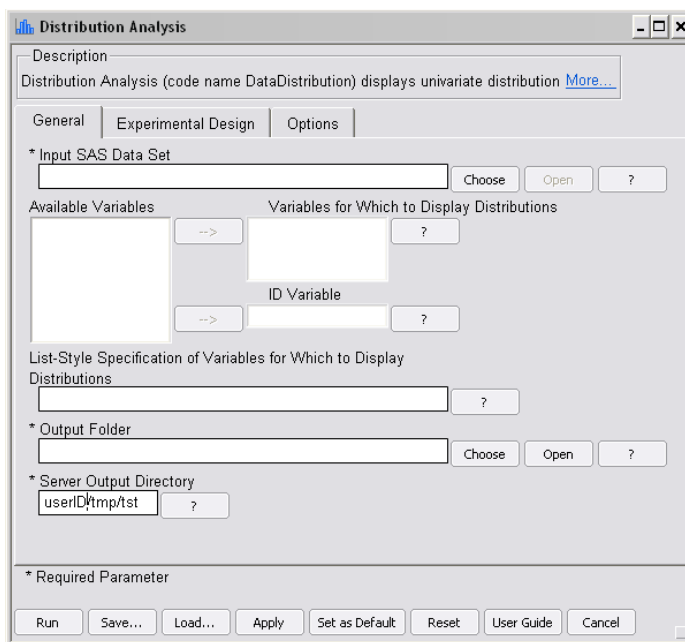
---

## Running an AP in Client/Server Mode

Before you can run a JMP Genomics AP in client/server mode, you must be connected to the SAS metadata server, as described in the previous section.

- ☞ Open the AP (Distribution Analysis, for example) you want to run.

Whenever an active connection to the metadata server is open, a new **Server Output Directory** field listing the server output directory specified in the Configuration step above is shown at the bottom of the dialog on the **General** tab (Figure 4.10).



**Figure 4.10** The Distribution Analysis dialog displaying the Server Output Directory field

To run the AP in client/server mode:

- ☞ Load your setting or specify the input data sets and the analysis parameters as usual.

---

*Note:* When you click run, all of the input files are copied to the server. If the input files that you need are already on the server, you can type or paste server paths directly into input data set and path fields.

---

- ☞ Click **Run**.

The AP runs on the server and returns the results to your desktop client machine.

---

*Note:* the first time you load settings or run an AP, during a JMP Genomics session, you are prompted to select a SAS work space server. Select the appropriate work space server (for example SASApp) from the window. This work space server is then used for the remainder of your session.

---

A copy of raw data files, data sets, SAS code, and results remain on the server in your server directory. You might wish to check this directory from time to time and delete unwanted files to save disk space.



---

# Appendices

---


The information in the following appendices is provided to make your use of JMP Genomics easier and more informative. They include a table listing the suffixes used to identify all of the different data sets output by JMP Genomics, a glossary and a Troubleshooting guide and a comprehensive bibliography.



## Using JMP Genomics Graphics in Presentations and Publications

---

When you need JMP Genomics graphics or data tables for use in another program, such as Microsoft Word or PowerPoint, you can copy and past all or part of the graphic or table into the other program.

- Click the selection tool ()
- Click and drag (or **Shift-click**) to select items in a JMP graphics window or data table.

---

*Note:* Clicking near the edge of the window selects the entire window.

---

- Click **Edit > Copy** to copy the selected item(s) in JMP Genomics.
- Click **Edit > Paste Special** in the target application, select Windows metafile (.wmf) as the format and click **OK** to paste the selected item into the other program.



## The SAS WHERE Expression

---

### What does it do?

A WHERE expression is a type of SAS expression that allows you to filter and select observations meeting one or more specific defined criteria. A WHERE expression can be as simple as a single variable name. A WHERE expression can contain a SAS function, or it can be a sequence of operands and operators that define one or more conditions for selecting observations.

---

### The WHERE expression

A SAS WHERE expression contains the *WHERE* keyword, one or more *operands* and one or more *operators*, and takes the following general form:

```
WHERE operand operator operand
```

---

*Note:* Blank spaces must be entered between each element in a WHERE expression.

---

The WHERE keyword alerts SAS to subset the data set.

An *operand* is an object to be operated on. An operand can be a variable (or column in a SAS data set), a SAS function (the result of a computation or other manipulation), or a constant.

An *operator* is a symbol that requests a comparison, a logical operation, or arithmetic computation.

WHERE expressions may either be *simple* or *compound*. Simple WHERE expressions contain only one condition that must be satisfied. Compound WHERE expressions contain more than one condition that must be satisfied, with each condition being separated by Boolean terms, such as *and* or *or*.

For example, to filter only diseased individuals from a data set containing a mixed population of diseased (sick) and healthy (healthy) individuals (as indicated in a column named `DiseaseStatus`), you could use the following simple WHERE expression:

```
WHERE DiseaseStatus = 'sick'
```

To simultaneously filter diseased individuals who also possess the genotype A/A (as indicated in a column named `Marker1`), you could use the following compound WHERE expression:

```
WHERE DiseaseStatus = 'sick' and Marker1 = 'A/A'
```

---

## Using a WHERE expression in JMP Genomics

Recall that JMP Genomics runs SAS in the background for manipulation and analysis of genomics data sets. Any statements or commands, including WHERE expressions, that can be used in SAS can also be used in JMP Genomics. However, because JMP dialogs function as the front end for generating and running the underlying SAS code, you don't need to understand all of the syntax for writing WHERE expressions. Instead, you are prompted by specific data entry fields (identified by the title: Filter to Include...) to specify the relevant operands and operators.

The Filter to Include Observations field (Figure 6.1), for example, found on many JMP Genomics dialogs, simplifies the specification of a WHERE expression.

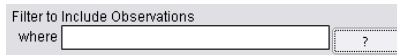


Figure 6.1 The Filter to Include Observations field

Note that the WHERE keyword has already been specified. There is no need to type the WHERE keyword. All you need to do is type the operands and the operators in the field.

☞ To use the previous example, type `DiseaseStatus = 'sick'`, as shown in Figure 6.2, to retain only the diseased individuals from the input data set.

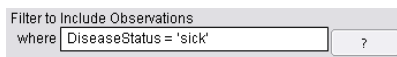


Figure 6.2 A WHERE expression to retain diseased individuals

The filter is applied when you click **Run**.

---

## Specifying an Operand

An operand is defined as the object to be operated on or the condition to be met. Operands can be either variables, SAS functions, or constants.

The variable names used in this statement (e.g., `Geneld`, `Position`) must exactly match those in the Input Data Set.

### Variables

A *variable* is a column in a SAS data set. Each variable has a name and a type (numeric or character) and contains values.

You must type the name exactly as it appears in the data set.

The values listed in numeric variables must be numeric. Character variables may contain numbers, letters, and symbols. The variable type determines how you specify that variable.

To filter the data based on a numeric variable, type the name of the variable followed by a space, the relevant operator followed by a space, and the condition to be met. For example, to select only those rows whose numeric value in the `FREQ` variable exceeds 0.5, type `FREQ > 0.5` in the `Filter` field.

SAS treats numeric values of zero (0) or that are missing (.) as false. All other values are true.

To continue the example listed above, to filter out those frequency observations that are not reported (indicated by the missing value symbol (.)), type `FREQ` in the `Filter` field.

To filter your data using a character variable, you must enclose the character string defining the condition to be met within quotes. For example, to filter only the diseased individuals, type `DiseaseStatus = 'sick'` in the `Filter` field. You may use either single quotes ('x') or double quotes ("x"), however, you must be consistent in the quotes you use. If you open the condition with a single quote you must close it using a single quote, and vice-versa. You cannot mix single and double quotes ('x', for example) within the same expression.

## SAS Functions

A SAS function returns a value from a computation or system manipulation. Most functions use arguments that you supply. To use a SAS function in a WHERE expression, type the name of the function followed by the argument. The argument must be contained within parentheses.

The `SUBSTR` function examines the character strings in a variable for a specific substring and retains only those observations containing the specified substring. The argument for this function specifies the variable containing the substring, the position within the string at which the substring starts, and the length of the substring. For example, typing the estimation `substr (GeneName, 1, 3) = "cyt"` in the `Filter` field subsets the input data set so that only observations in which the character value in the `GeneName` column begins with *cyt*, such as *cytosolic* or *cytochrome* are retained. Others, such as *lymphocyte*, are not.

Many additional, legitimate SAS functions can be used in WHERE estimations; for example `TODAY`, which returns the current date, however, most of these tend not to be very useful in genomics analysis and so, are not discussed further here.

Please refer to Base SAS documentation for further details about use of SAS functions in WHERE expressions, if you require more information.

## Constants

A constant is the fixed value within a variable for which you are searching. The value is either numeric or character. Constants are also called literals. For example, a constant could be a specific genotype.

As with character and numeric variables, if the constant is a character, you must type either single or double quotation marks before and after the value. Remember, you cannot mix single and double quotes ('x', for example) within the same expression. Do not use quotes if the constant is numeric.

---

## Specifying an Operator

An *operator* is a symbol that requests a comparison, a logical operation, or arithmetic computation. When writing the WHERE expression, you should place the operator between two operands.

## Comparison Operators

*Comparison* operators (also called binary operators) compare a variable with a value or with another variable. Comparison operators propose a relationship and ask SAS to determine whether that relationship holds. Only observations that meet the condition(s) specified are included in the analysis.

The comparison operators, available to you in JMP Genomics, are shown in **Table 6.1**.

**Table 6.1** Comparison Operators

Symbol	Definition	Example
= or EQ	<ul style="list-style-type: none"> <li>equal to</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Gene EQ "EGF"</code> subsets the input data set, retaining only those observations where the value in the <code>Gene</code> column is <code>EGF</code>.</li> </ul>
^= or NE	<ul style="list-style-type: none"> <li>not equal to</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>IndividualNum ^= 450</code> subsets the input data set, retaining only those observations where the value in the <code>IndividualNum</code> column is <code>450</code>.</li> </ul>
< or LT	<ul style="list-style-type: none"> <li>less than</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Position &lt; 100000</code> subsets the input data set, retaining only those observations where the position value is less than <code>100,000</code>.</li> </ul>
> or GT	<ul style="list-style-type: none"> <li>greater than</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Position GT 100000</code> subsets the input data set, retaining only those observations where the position value is greater than <code>100,000</code>.</li> </ul>
<= or LE	<ul style="list-style-type: none"> <li>less than or equal to</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Position &lt;= 100000</code> subsets the input data set, retaining only those observations where the position value is less than or equal to <code>100,000</code>.</li> </ul>
>= or GE	<ul style="list-style-type: none"> <li>greater than or equal to</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Position GE 100000</code> subsets the input data set, retaining only those observations where the position value is greater than or equal to <code>100,000</code>.</li> </ul>

You may combine comparison operators into compound WHERE expressions. For example, typing `50000 <= Position <= 100000` includes only those observations whose values listed in the `Position` variable are greater than or equal to `50,000` and are less than or equal to `100,000`.

---

*Note:* The combination of the `>` and `<` symbols (as in `7 > x < 9`, for example) is not supported.

---



---

*Remember:* You do not need to type WHERE in the Filter field.

---

## Arithmetic Operators

*Arithmetic* operators enable you to perform a mathematical operation on values in the specified operand.

The arithmetic operators, available to you in JMP Genomics, are shown in **Table 6.2**.

**Table 6.2** Arithmetic Operators

Symbol	Definition	Example
*	<ul style="list-style-type: none"> <li>• multiplication</li> </ul>	<ul style="list-style-type: none"> <li>• Typing <math>x = y * 0.1</math>, in the Filter field selects and returns observations whose values in the x column are equal to the corresponding y column values multiplied by 0.1.</li> </ul>
/	<ul style="list-style-type: none"> <li>• division</li> </ul>	<ul style="list-style-type: none"> <li>• Typing <math>x = y / 10</math>, in the Filter field selects and returns observations whose values in the x column are equal to the corresponding y column values divided by 10.</li> </ul>
+	<ul style="list-style-type: none"> <li>• addition</li> </ul>	<ul style="list-style-type: none"> <li>• Typing <math>x = y + 1</math>, in the Filter field selects and returns observations whose values in the x column are equal to the corresponding y column values plus one.</li> </ul>
-	<ul style="list-style-type: none"> <li>• subtraction</li> </ul>	<ul style="list-style-type: none"> <li>• Typing <math>x = y - 1</math>, in the Filter field selects and returns observations whose values in the x column are equal to the corresponding y column values less one.</li> </ul>
**	<ul style="list-style-type: none"> <li>• exponentiation</li> </ul>	<ul style="list-style-type: none"> <li>• Typing <math>x = y ** 2</math>, in the Filter field selects and returns observations whose values in the x column are equal to the square of the corresponding y column values.</li> </ul>

---

*Remember:* You do not need to type WHERE in the Filter field.

---

## Additional Operators

Additional operators can be used to select observations matching or containing a specified set of characters, missing values, to concatenate variables, return minimum or maximum values, etc.

The most useful of these operators are the IN, CONTAINS, and IS MISSING operators shown in **Table 6.3**.

Table 6.3 Additional Operators

Symbol	Definition	Example
IN	<ul style="list-style-type: none"> <li>equal to one of a list</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>Window IN (1, 4, 9, 32)</code> subsets the input data set, retaining only those observations where the value in the <code>Window</code> column is equal to either 1, 4, 9 or 32.</li> </ul>
CONTAINS	<ul style="list-style-type: none"> <li>contains a specified set of characters</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>GeneName CONTAINS "cyt"</code> subsets the input data set, retaining only those observations where the value in the <code>GeneName</code> column contains the string <code>cyt</code> (for example all of the cytochrome genes).</li> </ul>
IS NULL or IS MISSING	<ul style="list-style-type: none"> <li>value is missing</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>GeneName IS MISSING</code> subsets the input data set, retaining only those observations where the value in the <code>GeneName</code> column is absent.</li> <li>This operator can be combined with the <code>NOT</code> operator.</li> </ul>

### Prefix Operators

Prefix operators are added to modify existing operators in a WHERE expression. The most commonly used prefix operator is NOT. Placing this prefix before the normal operator selects those observations not matching the specified condition.

Table 6.4 NOT is a prefix operator

Symbol	Definition	Example
NOT	<ul style="list-style-type: none"> <li>value does not match specified characters</li> </ul>	<ul style="list-style-type: none"> <li>Typing <code>GeneName NOT "cyt"</code> subsets the input data set, retaining only those observations where the value in the <code>GeneName</code> column is not <code>cyt</code>.</li> </ul>

The + and - symbols, which typically serve as arithmetic operators, also function as prefix operators when put in front of an open parentheses. Typing `z = -(x + y)` in the Filter field returns observations whose value for z equals the negative of the sum of the values in x and y.

### Compound WHERE Expressions

You might find it useful to combine multiple simple WHERE expressions into a single compound into a single, compound WHERE expression. Expressions can be joined either by using Boolean terms, such as AND and OR, or by grouping arguments within parentheses.

For example, typing `GeneName CONTAINS "cyt" OR "Fed"` in the Filter field selects and returns all rows with these character strings in the `GeneName` columns. Alternatively, typing `GeneName CONTAINS ("cyt", "Fed")` does the same.

### **Additional Information**

Please refer to Base SAS documentation for additional operators, as needed.



# Chapter 7

## JMP Genomics Files Are Identified by Suffixes

---

Typical analyses in JMP Genomics often generate a large number of output files and data sets. This is especially true when input and output files are contained with the same folder, when an AP produces multiple output files, or when the same output folder is used for multiple processes. If any or all of these conditions occur, the number of files in a folder can multiply dramatically. How do you identify and distinguish between the different files? More importantly, how do you prevent overwriting existing files with the output of subsequent APs, particularly when all of the files tend to be similarly named?

JMP Genomics adds a unique suffix to each output file generated by an AP. Suffixes are specific to each AP and are dependent on the type of content contained in the file. These suffixes allow you to identify different output files and to correlate them with specific APs. All of the suffixes used by JMP Genomics are listed alphabetically in the table below. Each suffix is defined both by the AP that generates it and by the contents of the file that carries it.

---

*Note:* Even with the suffixes, it is still possible to overwrite existing files. You should take care to specify different names for output files when doing multiple runs with the same input files. Alternatively, you may specify different output folders.

---

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_aag	<ul style="list-style-type: none"><li>• Create Annotation Analysis Group Variables</li></ul>	Output data set listing the input data plus the annotation analysis group variable
_afreqs	<ul style="list-style-type: none"><li>• Marker Properties</li></ul>	Output data set containing allele frequencies
_agc	<ul style="list-style-type: none"><li>• Correlation and Grouped Scatterplots.</li></ul>	Output data set consisting of the input data merged with the annotation data set
_alp	<ul style="list-style-type: none"><li>• Affymetrix Cytogenetics CHP Input Engine</li></ul>	Output data set
_amr	<ul style="list-style-type: none"><li>• ANOVA</li></ul>	Output data set listing annotation information and various statistics for the most significant markers

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_anm	<ul style="list-style-type: none"> <li>ANOVA Normalization</li> </ul>	Output data set with the same structure as the input data set but with normalized response variables
_api	<ul style="list-style-type: none"> <li>Pseudolmage</li> </ul>	Output data set listing the X- and Y-coordinates, probe set IDs, and intensities for each spot on the specified array
_ars	<ul style="list-style-type: none"> <li>ANOVA</li> <li>One-Way ANOVA</li> </ul>	Output data set listing standardized residuals for each observation in the input data set
_as	<ul style="list-style-type: none"> <li>Relationship Matrix</li> </ul>	Output data set listing the relationship matrix generated using the allele sharing similarity estimate.
_asf	<ul style="list-style-type: none"> <li>Allele Specific Expression Filter</li> </ul>	Output data set containing the filtering results.
_asp	<ul style="list-style-type: none"> <li>Affected Sib-Pair Tests</li> </ul>	Output data set containing affected sib-pair test statistics and .jsl file for plotting <i>p</i> -values
_asr	<ul style="list-style-type: none"> <li>Allele Specific Expression Filter</li> </ul>	Output data set containing the <i>log</i> -ratio of RNA intensity data for estimated homozygous genotypes
_atan	<ul style="list-style-type: none"> <li>ArrayTrack</li> </ul>	Annotation file created by ArrayTrack
_bin	<ul style="list-style-type: none"> <li>Copy Number Bin</li> </ul>	Output data set listing the bins into which the probe sets have been grouped. This data set also lists the physical position of the bin start site and the average intensity across the bin.
_bnm	<ul style="list-style-type: none"> <li>Batch Normalization</li> </ul>	Output data set containing the batch-normalized data
_bnp	<ul style="list-style-type: none"> <li>Batch Normalization</li> </ul>	Output data set containing the results of the principal components analysis
_bowa	<ul style="list-style-type: none"> <li>Bivariate One-Way ANOVA</li> </ul>	Output data set containing annotation data, and statistics on individual experiments and pair-wise comparisons
_box	<ul style="list-style-type: none"> <li>Distribution Analysis</li> </ul>	Output data set listing the relative densities of spots on each array grouped by similar responses; used for drawing Box plots
_bsc	<ul style="list-style-type: none"> <li>Batch Scoring</li> </ul>	Output data set containing the batch-normalized data

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_bypathwayid	<ul style="list-style-type: none"> <li>• KEGG Pathway Search</li> </ul>	Output data set listing the various pathways and affiliated genes for each gene entered in the Gene/Protein IDs field
_cca	<ul style="list-style-type: none"> <li>• Case-Control Association</li> </ul>	Output data set containing case-control association test statistics and .jsl file for plotting $p$ -values
_ce	<ul style="list-style-type: none"> <li>• Combine Experiments</li> </ul>	Output data set generated from the merging of multiple tall SAS data sets.
_ceb	<ul style="list-style-type: none"> <li>• Column Enrichment</li> </ul>	Output data set listing the categories that show results displaying significance as specified by the binary significance variables
_cec	<ul style="list-style-type: none"> <li>• Column Enrichment</li> </ul>	Output data set listing the categories that show results displaying significance as specified by the continuous significance variables
_ceexp	<ul style="list-style-type: none"> <li>• Combine Experiments</li> </ul>	Output experimental design data set
_cef	<ul style="list-style-type: none"> <li>• Column Enrichment</li> </ul>	Output data set containing the Fisher Exact test results for the binary significance variables
_cep	<ul style="list-style-type: none"> <li>• Column Enrichment</li> </ul>	Output data set containing PAGE results for the continuous variables.
_chrn	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_cim_out	<ul style="list-style-type: none"> <li>• IM and CIM Analysis</li> </ul>	Output data set listing position and associated LOD scores for each marker
_cml	<ul style="list-style-type: none"> <li>• IM and CIM Analysis</li> </ul>	Output data set listing the control markers used in the analysis
_cor	<ul style="list-style-type: none"> <li>• Correlation and Principal Components</li> <li>• Cross Correlation</li> </ul>	Output data set listing the correlation of all the different variables in the input data
_cov	<ul style="list-style-type: none"> <li>• Correlation and Principal Components</li> </ul>	Output data set listing the covariates of all the different variables in the input data
_cpg	<ul style="list-style-type: none"> <li>• SNP-Trait Association</li> <li>• Imputed SNP Trait Association</li> <li>• Q-K Mixed Model</li> </ul>	Output data set containing covariance parameter estimates for the genotype test. This data set is generated whenever any random effects are specified.

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_cpt	<ul style="list-style-type: none"> <li>• SNP-Trait Association</li> <li>• Imputed SNP Trait Association</li> <li>• Q-K Mixed Model</li> </ul>	Output data set containing covariance parameter estimates for the trend test. This data set is generated whenever any random effects are specified.
_csn	<ul style="list-style-type: none"> <li>• Control Set Normalization</li> </ul>	Output data set containing the normalized data
_csne	<ul style="list-style-type: none"> <li>• Control Set Normalization</li> </ul>	Output experimental design data set
_dap	<ul style="list-style-type: none"> <li>• Append</li> </ul>	A concatenated data set resulting from appending the rows of two different data sets
_data	<ul style="list-style-type: none"> <li>• Exon and Whole Transcript Expression CEL</li> </ul>	SAS data set containing the imported data
_den	<ul style="list-style-type: none"> <li>• Distribution Analysis</li> </ul>	Output data set used for plotting the overlaid kernel density estimate plot
_dge	<ul style="list-style-type: none"> <li>• Distribution Analysis</li> </ul>	Output experimental design data set when using Group variables
_dgm	<ul style="list-style-type: none"> <li>• Distribution Analysis</li> </ul>	Output data set when using Group variables
_dlb	<ul style="list-style-type: none"> <li>• Change Labels</li> </ul>	Output data set in which the variables of the original data set have been relabeled
_dln	<ul style="list-style-type: none"> <li>• Change Lengths</li> </ul>	The output data set in which the lengths of the variables of the original data set have been resized
_dmg	<ul style="list-style-type: none"> <li>• Merge</li> </ul>	The large data set resulting from merging two smaller data sets
_dmt	<ul style="list-style-type: none"> <li>• Distance Matrix</li> </ul>	Output data set listing the measures of dissimilarity between all pairs of observations in the experiment
_drk	<ul style="list-style-type: none"> <li>• Rank Rows</li> </ul>	The output data set in which the individual values within each variable have been replaced with the ranking of that observation within the variable
_drn	<ul style="list-style-type: none"> <li>• Rename</li> </ul>	The output data set in which the primary variables of the original data set have been renamed
_dro	<ul style="list-style-type: none"> <li>• Reorder</li> </ul>	The output data set in which the order of the variables in the original data set has been changed

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_drs	<ul style="list-style-type: none"> <li>Statistics for Rows</li> </ul>	Data set listing row-wise statistics for the primary data set. Rows meeting the criteria specified in the Delete Rows Satisfying this Expression box on the Options tab are excluded from the data set.
_dsa	<ul style="list-style-type: none"> <li>Discriminant Analysis</li> </ul>	Output data set containing all of the data from the input data set plus three additional columns. These include observation number, and two columns listing the predicted values for the dependent class variable.
_dsav	<ul style="list-style-type: none"> <li>Discriminant Analysis</li> </ul>	Lists the names of the variables selected and used for making predictions
_dsm	<ul style="list-style-type: none"> <li>Statistics for Columns</li> </ul>	Data set listing summary statistics for the variables in the primary data set
_dso	<ul style="list-style-type: none"> <li>Distance Scoring</li> </ul>	Lists the statistics and associated predictions for each of the individuals in the input data set
_dsov	<ul style="list-style-type: none"> <li>Distance Scoring</li> </ul>	Lists the names of the variables selected and used for making predictions
_dsp	<ul style="list-style-type: none"> <li>Data Step</li> </ul>	The modified data set resulting from executing SAS data step commands on a data set
_dss	<ul style="list-style-type: none"> <li>Subset</li> </ul>	Output subset data set
_dst	<ul style="list-style-type: none"> <li>Sort Rows</li> </ul>	The output data set in which the variables in the input data set have been sorted according to specified key variables
_dtf	<ul style="list-style-type: none"> <li>Transform</li> </ul>	The output data set resulting from the mathematical transformation of specified variables in a primary data set
_dtr	<ul style="list-style-type: none"> <li>Transform Rectangular</li> </ul>	The data set resulting from the transposition of a block of variables in the original data set
_eig	<ul style="list-style-type: none"> <li>Correlation and Principal Components</li> <li>Relationship Matrix</li> </ul>	Output data set listing the eigen values and associated experimental statistics
_esc	<ul style="list-style-type: none"> <li>Gene Set Scoring</li> </ul>	Output data set containing the scored data
_est	<ul style="list-style-type: none"> <li>Estimate Builder/Compare Means</li> </ul>	Output file containing the estimate statements

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_exp	<ul style="list-style-type: none"> <li>Many different input engines, data set utilities and analytical processes</li> </ul>	Experimental Design Data Set (EDDS)
_fff	<ul style="list-style-type: none"> <li>Feature Flagger</li> </ul>	Output data set in which individual values for the variables have been replaced with "0", if values are within defined parameters, or "1", if values are outside of those parameters
_fin	<ul style="list-style-type: none"> <li>Filter Intensities</li> </ul>	Output data set listing the filtered intensities
_fnm	<ul style="list-style-type: none"> <li>Factor Analysis Normalization</li> </ul>	Output data set with the same structure as the input data set, but with normalized response variables
_gfreqs	<ul style="list-style-type: none"> <li>Marker Properties</li> </ul>	Output data set containing genotype frequencies
_glm	<ul style="list-style-type: none"> <li>General Linear Model Selection</li> </ul>	Lists the statistics and associated predictions for each of the individuals in the input data set
_glmv	<ul style="list-style-type: none"> <li>General Linear Model Selection</li> </ul>	Lists the names of the variables selected and used for making predictions
_gms	<ul style="list-style-type: none"> <li>Gene Model Summary</li> </ul>	Output data set containing the summarized data
_gp	<ul style="list-style-type: none"> <li>Build Genotype Probability Data Set</li> </ul>	Output data set listing the probabilities of observing specific QTL genotypes at each one cM interval along the chromosome(s)
_gse	<ul style="list-style-type: none"> <li>Gene Set Enrichment</li> </ul>	Output data set containing enrichment test results
_gsei	<ul style="list-style-type: none"> <li>Gene Set Enrichment</li> </ul>	Output data set containing 0-1 indicator variables for the categories
_hc	<ul style="list-style-type: none"> <li>Hierarchical Clustering</li> </ul>	Output data set containing all of the variables from the input data set plus results (row means and standard values, the cluster to which each belongs and its place within the cluster) of the hierarchical clustering.
_hest	<ul style="list-style-type: none"> <li>Haplotype Estimation</li> </ul>	.html output files, .jsl file for plotting $p$ -values, and output data set containing association test statistics
_hfr	<ul style="list-style-type: none"> <li>Haplotype Estimation</li> </ul>	Output data set containing haplotype frequency estimates that can be used as input to the htSNP Selection process

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_her	<ul style="list-style-type: none"> <li>• Haseman-Elston Regression</li> </ul>	Output data set containing Haseman-Elston regression test statistics and .jsl file for plotting $p$ -values
_hph	<ul style="list-style-type: none"> <li>• Haplotype Estimation</li> </ul>	Output data set containing phase assignment probabilities. This data set can be used as the input data set for the Haplotype Trend Regression process.
_hti	<ul style="list-style-type: none"> <li>• htSNP Selection</li> </ul>	Output data set that contains all the columns from the annotation data set with an additional htSNP Indicator column
_htr	<ul style="list-style-type: none"> <li>• Haplotype Trend Regression</li> </ul>	.html output files
_hts	<ul style="list-style-type: none"> <li>• htSNP Selection</li> </ul>	.html output files, .jsl file for displaying htSNPs, and output data set containing htSNP information
_hwetest	<ul style="list-style-type: none"> <li>• Marker Properties</li> </ul>	Output data set containing HWE test statistics and other marker properties
_ibd	<ul style="list-style-type: none"> <li>• Relationship Matrix</li> </ul>	Output data set listing the relationship matrix generated using the identical by descent estimate.
_ibs	<ul style="list-style-type: none"> <li>• Relationship Matrix</li> </ul>	Output data set listing the relationship matrix generated using the identical by state estimate.
_igp	<ul style="list-style-type: none"> <li>• Imputed SNP (Wide Format)</li> <li>• Imputed SNP (Tall Format)</li> </ul>	Stacked output genotype probabilities SAS data set containing probabilities for each genotype at a SNP
_igt	<ul style="list-style-type: none"> <li>• Imputed SNP (Wide Format)</li> <li>• Imputed SNP (Tall Format)</li> </ul>	Output genotype threshold SAS data set listing the most likely genotype for each SNP
_iqr	<ul style="list-style-type: none"> <li>• Data Standardize</li> </ul>	Output data set containing the IQR standardized values for the input data set
_ist	<ul style="list-style-type: none"> <li>• Imputed SNP-Trait Association</li> </ul>	Output data set containing association test statistics and .jsl file for plotting $p$ -values
_jnm	<ul style="list-style-type: none"> <li>• Johnson Su Normalization</li> </ul>	Output data set containing $S_U$ normalized data
_jnp	<ul style="list-style-type: none"> <li>• Johnson Su Normalization</li> </ul>	Output data set listing the four estimated parameters for each array used to generate the $S_U$ normalized data
_kc	<ul style="list-style-type: none"> <li>• K Matrix Compression</li> </ul>	Output data set containing the compressed matrix

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_kcv	<ul style="list-style-type: none"> <li>• K Matrix Compression</li> </ul>	Output data set containing the covariance parameter estimates from every model fit during the compression
_kgi	<ul style="list-style-type: none"> <li>• KEGG Get Gene Identifiers</li> </ul>	Output data set containing a column listing the KEGG gene identifiers appended to the input data set.
_kgp	<ul style="list-style-type: none"> <li>• KEGG Get Gene Pathways</li> </ul>	Output data set containing a column listing the KEGG pathway identifiers appended to the input data set.
_kmc	<ul style="list-style-type: none"> <li>• K-Means Clustering</li> </ul>	Lists the center of each cluster and the corresponding statistics
_kmd	<ul style="list-style-type: none"> <li>• K-Means Clustering</li> </ul>	Lists the statistics for each observation, the cluster to which the observation has been assigned and its distance from the cluster mean
_kmx	<ul style="list-style-type: none"> <li>• Kinship Matrix</li> </ul>	Output data set consisting of the input data modified to identify the parents of the different pedigrees and matrix statistics
_knn	<ul style="list-style-type: none"> <li>• K-Nearest Neighbors</li> </ul>	Lists the statistics and associated predictions for each of the individuals in the input data set
_knnv	<ul style="list-style-type: none"> <li>• K-Nearest Neighbors</li> </ul>	Lists the names of the variables selected and used for making predictions
_ld	<ul style="list-style-type: none"> <li>• Linkage Disequilibrium</li> </ul>	.jsl file for displaying LD results
_ldbc	<ul style="list-style-type: none"> <li>• LD Block Creation</li> </ul>	Output data set listing the name, location, LD block and annotation information for each of the SNPs
_ldc	<ul style="list-style-type: none"> <li>• Linkage Disequilibrium</li> </ul>	Output data set containing LD measures used in the Contour Plot and/or the LD Decay plot
_lds	<ul style="list-style-type: none"> <li>• Linkage Disequilibrium</li> </ul>	Output data set containing LD test statistics
_ldts	<ul style="list-style-type: none"> <li>• LD tagSNP Selection</li> </ul>	Output data set containing tagSNP information
_ldts_body	<ul style="list-style-type: none"> <li>• LD tagSNP Selection</li> </ul>	.html output file summarizing the tagSNP information
_len	<ul style="list-style-type: none"> <li>• List Enrichment</li> </ul>	Output data set containing the list enrichment statistics

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_lgr	<ul style="list-style-type: none"> <li>Logistic Regression</li> <li>SNP Interaction Discovery</li> </ul>	Output data set containing all of the data from the input data set plus three additional columns. These include observation number, and two columns listing the predicted values for the dependent class variable.
_lgrv	<ul style="list-style-type: none"> <li>Logistic Regression</li> <li>SNP Interaction Discovery</li> </ul>	Lists the names of the variables selected and used for making predictions
_lnm	<ul style="list-style-type: none"> <li>LoessNormalization</li> </ul>	Output data set containing the loess normalized data
_lnp	<ul style="list-style-type: none"> <li>Loess Normalization</li> </ul>	Output data set containing data used for drawing the M-A plots
_map	<ul style="list-style-type: none"> <li>Imputed SNP (Tall Format) Input Engine</li> </ul>	Output data set containing annotation marker map information
_mdf	<ul style="list-style-type: none"> <li>Multidimensional Scaling</li> </ul>	Output data set listing statics from the multidimensional scaling process
_mdr	<ul style="list-style-type: none"> <li>Multidimensional Scaling</li> </ul>	Output data set listing the statistics for and distances between pairs of observations from the input data set
_mds	<ul style="list-style-type: none"> <li>Multidimensional Scaling</li> </ul>	Output data set listing the multi-dimensional coordinates of each observation
_mea	<ul style="list-style-type: none"> <li>Data Standardize</li> </ul>	Output data set containing the mean standardized values for the input data set
_med	<ul style="list-style-type: none"> <li>Data Standardize</li> </ul>	Output data set containing median standardized values
_meds	<ul style="list-style-type: none"> <li>Data Standardize</li> </ul>	Output data set containing location and scale estimates used to median-center the input values
_mgs	<ul style="list-style-type: none"> <li>Missing Genotype by Trait Summary</li> </ul>	Output data set listing the counts of genotyped cases and controls, the number of missing genotypes for cases and controls, and the statistics for testing for different missing proportions between cases and controls for each marker
_mld	<ul style="list-style-type: none"> <li>Malecot LD Map</li> </ul>	.jsl file for creating plots and output data set containing estimates of LDU between pairs of markers
_mml	<ul style="list-style-type: none"> <li>Malecot LD Map</li> </ul>	Output data set containing estimates for parameters M and L

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_mmp	<ul style="list-style-type: none"> <li>Mixed Model Power</li> </ul>	Output data set listing power values and associated $t$ -statistics for a set of hypothesis tests for a range of $\alpha$ and effects sizes
_mmr	<ul style="list-style-type: none"> <li>Mixed Model Analysis</li> </ul>	Lists annotation and statistical information for each of the groups
_mnm	<ul style="list-style-type: none"> <li>Mixed Model Normalization</li> </ul>	Output data set with the same structure as the input data set but with both the original and the normalized response variables
_mrs	<ul style="list-style-type: none"> <li>Mixed Model Analysis</li> </ul>	Output residual data set
_msd	<ul style="list-style-type: none"> <li>Merge Gene Sets</li> </ul>	Output merged data set
_mtr	<ul style="list-style-type: none"> <li>Merge and Transform</li> </ul>	The output data set resulting from the merging of two sets sharing a common set of variables and the subsequent computation of an arbitrary function for each pair of variables sharing the same name
_mta	<ul style="list-style-type: none"> <li>Marker-Trait Association</li> </ul>	Output data set containing association test statistics and .jsl file for plotting $p$ -values
_mvi	<ul style="list-style-type: none"> <li>Missing Value Imputation</li> </ul>	The complete output data set in which values missing from the input data set have been imputed from the non-missing values in the same row
_names	<ul style="list-style-type: none"> <li>Nexus</li> </ul>	Output data set containing names of markers
_numgeno	<ul style="list-style-type: none"> <li>Marker Properties</li> </ul>	Output data set containing numerically coded genotypes used for the cell plot
_one	<ul style="list-style-type: none"> <li>Single Marker Analysis</li> </ul>	Output data set listing the results of the regression analysis and $-\log_2 p$ -values for association of each of the markers with each of the two quantitative traits, along with annotation information for each of the markers
_owa	<ul style="list-style-type: none"> <li>One-Way ANOVA</li> </ul>	Output data set containing annotation data, and statistics on individual experiments and pair-wise comparisons
_pca	<ul style="list-style-type: none"> <li>Principal Components Analysis for Population Stratification</li> <li>Principal Components Analysis</li> </ul>	Output data set listing the row scores for each of the principal components

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_pcm	<ul style="list-style-type: none"> <li>• Principal Components Analysis for Population Stratification</li> </ul>	Output data set containing the PCA data merged into the input data set.
_peg	<ul style="list-style-type: none"> <li>• SNP-Trait Association</li> <li>• Survey SNP-Trait Association</li> </ul>	Output data set containing the parameter estimates for the genotype test.
_pet	<ul style="list-style-type: none"> <li>• SNP-Trait Association</li> <li>• Survey SNP-Trait Association</li> <li>• Imputed SNP-Trait Association</li> </ul>	Output data set containing the parameter estimates for the trend test.
_pi	<ul style="list-style-type: none"> <li>• Plot Intensities</li> </ul>	Output data set containing raw intensities and row statistics for the input data set.
_plr	<ul style="list-style-type: none"> <li>• SNP Interaction Discovery</li> </ul>	Output data set containing all of the data from the input data set plus three additional columns. These include observation number, and two columns listing the predicted values for the dependent class variable.
_plrv	<ul style="list-style-type: none"> <li>• SNP Interaction Discovery</li> </ul>	Lists the names of the variables selected and used for making predictions
_pls	<ul style="list-style-type: none"> <li>• Partial Least Squares</li> </ul>	Lists the statistics and associated predictions for each of the individuals in the input data set
_plsv	<ul style="list-style-type: none"> <li>• Partial Least Squares</li> </ul>	Lists the statistics associated with the ability of each marker to be used as a predictor for the dependent class variable
_pmd	<ul style="list-style-type: none"> <li>• Population Measures</li> </ul>	Output symmetric matrix of dissimilarities between specified groups within a study
_pmf	<ul style="list-style-type: none"> <li>• Population Measures</li> </ul>	Output data set listing the individual $F$ -statistics for each marker
_pmo	<ul style="list-style-type: none"> <li>• Population Measures</li> </ul>	Output data set listing the overall $F$ -statistics for the population
_pps	<ul style="list-style-type: none"> <li>• Principal Components Analysis for Population Stratification</li> </ul>	Output data set listing each of the markers with selected annotation data and the resulting probabilities the markers are associated with the disease trait, given population stratification

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_pnm	<ul style="list-style-type: none"> <li>Effect Removal via PLS Normalization</li> <li>Partial Least Squares Normalization</li> </ul>	Output data set with the same structure as the input data set but with normalized response variables
_prs	<ul style="list-style-type: none"> <li>Relationship Matrix</li> </ul>	Output data set listing pairs of samples having an identical-by-descent value exceeding a specified threshold.
_psc	<ul style="list-style-type: none"> <li>Phenotype Summary</li> </ul>	Output data set containing counts for categorical variables
_psq	<ul style="list-style-type: none"> <li>Phenotype Summary</li> </ul>	Output data set containing counts for quantitative variables
_ptr	<ul style="list-style-type: none"> <li>Partition Trees</li> </ul>	Output data set listing the input data that has been recursively partitioned to optimal splitting relationships as determined by the specified dependent and predictor variables
_ptrv	<ul style="list-style-type: none"> <li>Partition Trees</li> </ul>	Lists the names of the variables selected and used for making predictions
_pva	<ul style="list-style-type: none"> <li>P-Value Adjustment</li> </ul>	Output data set containing the columns in the input data set with four additional columns listing the adjusted $p$ -values for each set of observations
_pvb	<ul style="list-style-type: none"> <li>P-Value Browser</li> </ul>	Output data set containing the columns in the input data set with four additional columns listing the adjusted $p$ -values for each set of observations
_pvc	<ul style="list-style-type: none"> <li>Correlation and Principal Components</li> </ul>	Output data set containing the variance within each of the principal components that can be attributed to each of the variance components
_qcbp	<ul style="list-style-type: none"> <li>Exon and Whole Transcript Expression CEL</li> </ul>	Output data set that is used to create a pseudoimage of the arrays
_qkm	<ul style="list-style-type: none"> <li>Q-K Mixed Model</li> </ul>	Output data set containing association test statistics and .jsl file for plotting $p$ -values
_qnm	<ul style="list-style-type: none"> <li>Quantile Normalization</li> </ul>	Output data set containing the quantile normalized data
_qnp	<ul style="list-style-type: none"> <li>Quantile Normalization</li> </ul>	Output data set containing the quantile normalized data along with relative rank of each observation and residual data

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_qtdt	<ul style="list-style-type: none"> <li>Quantitative TDT</li> </ul>	Output data set containing Q-TDT test statistics and .jsl file for plotting <i>p</i> -values
_rai	<ul style="list-style-type: none"> <li>Ratio Analysis</li> </ul>	Output data set containing the <i>log</i> intensities of the input data
_rar	<ul style="list-style-type: none"> <li>Ratio Analysis</li> </ul>	Output data set containing the <i>log</i> ratio of the two-channel input data
_rbm	<ul style="list-style-type: none"> <li>Radial Basis Machine</li> </ul>	Lists the statistics and associated predictions for each of the individuals in the input data set
_rbmv	<ul style="list-style-type: none"> <li>Radial Basis Machine</li> </ul>	Lists the names of the variables selected and used for making predictions
_rm	<ul style="list-style-type: none"> <li>Relationship Matrix</li> </ul>	Data set containing the root matrix as computed by single value decomposition (SVD) appended to the input data set.
_s2b	<ul style="list-style-type: none"> <li>2D Bin</li> </ul>	Output data set containing the binned values for each variable
_s2d	<ul style="list-style-type: none"> <li>2D Detrend</li> </ul>	Output data set containing the baseline-adjusted and corrected values
_s2p	<ul style="list-style-type: none"> <li>2D Peakfind</li> </ul>	Output data set identifying each peak found in each trace
s2p_det	<ul style="list-style-type: none"> <li>2D Peakfind</li> </ul>	Output data set providing specific information (upper and lower boundaries, height, area, for example) for each peak identified in each trace
_s2g	<ul style="list-style-type: none"> <li>2D Plot</li> </ul>	Output data set listing trimmed input values
_s3a	<ul style="list-style-type: none"> <li>3D Align</li> </ul>	Output data set listing the <i>X</i> and <i>Y</i> coordinates for each row with the <i>Z</i> values for each sample listed in a separate column
_s3g	<ul style="list-style-type: none"> <li>3D Plot</li> </ul>	Output data set listing the <i>X</i> and <i>Y</i> coordinates for each row with the <i>Z</i> values for each sample listed in a separate column

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_sbg	<ul style="list-style-type: none"> <li>• Case-Control Association</li> <li>• PCA for Population Stratification</li> <li>• Marker-Trait Association</li> <li>• SNP-Trait Association</li> <li>• Imputed SNP-Trait Association</li> <li>• Survey SNP-Trait Association</li> <li>• Q-K Mixed Model</li> <li>• Quantitative TDT</li> <li>• TDT</li> </ul>	Output data set listing the number of significant markers for the selected association tests.
_sdp	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_seq	<ul style="list-style-type: none"> <li>• Affymetrix Tiling CEL Input Engine</li> </ul>	Output sequence data set
_sfs	<ul style="list-style-type: none"> <li>• Surface Summary</li> </ul>	Output data set containing the estimated surface
_sgc	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_sgl	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_smc	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_snl	<ul style="list-style-type: none"> <li>• Affymetrix Cytogenetics CHP Input Engine</li> </ul>	Output data set
_sp	<ul style="list-style-type: none"> <li>• Exon and Whole Transcript Expression CEL</li> </ul>	Output data set which lists single-probes not associated to any probe set
_spmm	<ul style="list-style-type: none"> <li>• Survival Predictive Modeling</li> </ul>	Output data set listing mean survival times
_spms	<ul style="list-style-type: none"> <li>• Survival Predictive Modeling</li> </ul>	Output data set listing survival functions corresponding to the training data set
_spmt	<ul style="list-style-type: none"> <li>• Survival Predictive Modeling</li> </ul>	Output data set listing survival functions corresponding to the test data set
_spmv	<ul style="list-style-type: none"> <li>• Survival Predictive Modeling</li> </ul>	Output data set listing the effects applied in the Cox model

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_sps	<ul style="list-style-type: none"> <li>Exon and Whole Transcript Expression CEL</li> </ul>	Output data set which lists single-probe-sets not associated to any transcript cluster
_sr	<ul style="list-style-type: none"> <li>Subset/Reorder Genetics Data</li> </ul>	The output data set resulting from the reordering/removal of markers or individuals from either a genotype or annotation data set
_sst	<ul style="list-style-type: none"> <li>Survey SNP-Trait Association</li> </ul>	Output data set association test statistics
_sta	<ul style="list-style-type: none"> <li>SNP-Trait Association</li> </ul>	Output data set containing association test statistics and .jsl file for plotting $p$ -values
_std	<ul style="list-style-type: none"> <li>Data Standardize</li> </ul>	Output data set with the same structure as the input data set but response variable values that have been standardized using one or more different methods
_stk	<ul style="list-style-type: none"> <li>Stack</li> </ul>	Output stacked data set
_sva	<ul style="list-style-type: none"> <li>Survival Analysis</li> </ul>	Output data set listing survival statistics
_svd	<ul style="list-style-type: none"> <li>Calculate Square Root of Matrix</li> </ul>	Output data set
_tal	<ul style="list-style-type: none"> <li>Transpose Tall and Wide</li> <li>Unstack</li> </ul>	The tall data set resulting from either the transposition of a wide data set or the unstacking of a stacked data set. <i>Note:</i> These processes also generate an associated EDDS
_tdt	<ul style="list-style-type: none"> <li>TDT</li> </ul>	Output data set containing TDT statistics and .jsl file for plotting $p$ -values
_tdtt	<ul style="list-style-type: none"> <li>TDT</li> </ul>	Output data set containing a pair of allelic transmission scores for each individual at each marker locus.
_tst	<ul style="list-style-type: none"> <li>Correlation and Principal Components</li> <li>Cross Correlation</li> </ul>	Output data test data set containing the $p$ -values for the test that the correlation equals zero
_vc	<ul style="list-style-type: none"> <li>Correlation and Principal Components</li> <li>Batch Normalization</li> </ul>	Output data set containing variance component estimates computed from the principal components

JMP Genomics Files have suffixes: Identification of output files by suffix, AP, and contents.

File Suffix	APs Generating the File	Function/Contents of the File
_vcp	<ul style="list-style-type: none"> <li>Variance Components</li> </ul>	Output data set listing each of the markers with selected statistics for the <i>chi</i> -square statistics and <i>p</i> -values for the likelihood ratio test or the <i>z</i> -scores and <i>p</i> -values for the Wald test that test for linkage between each marker and the quantitative trait
_vg	<ul style="list-style-type: none"> <li>Verify Gender of Samples</li> </ul>	Output data set listing the genetic sex of the individuals
_wid	<ul style="list-style-type: none"> <li>Transpose Tall and Wide</li> </ul>	The wide data set resulting from the transposition of a tall data set
_z	<ul style="list-style-type: none"> <li>Surface Summary</li> </ul>	Output coordinates data set

## JMP Genomics Processes Call SAS PROCs

JMP Genomics makes extensive use of SAS routines (called PROCs) in order to carry out all of its different functions and analyses. The table below lists the PROCs called by each process. It does not include the following PROCs, which are called extensively by every process: CONTENTS, DATASETS, EXPORT, MEANS, PRINT, REGISTRY, SORT, and TRANSPOSE. Processes are listed alphabetically.

### JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
ABI Analyst Input Engine	APPEND, IMPORT, SQL, STDIZE
Affected Sib-Pair Tests	ALLELE, APPEND, FORMAT, PSMOOTH
Affymetrix Annotation CSV Files	IMPORT
Affymetrix CNAT Input Engine	APPEND, IMPORT, SQL, STDIZE
Affymetrix CN CHP Input Engine	APPEND, IMPORT, SQL, STDIZE
Affymetrix Cytogenetics CEL Input Engine	APPEND, FORMAT, G3D, IMPORT, KDE, LOESS, REG, SQL, STDIZE, SURVEYSELECT, UNIVARIATE
Affymetrix Cytogenetics CEL Input Engine	APPEND, IMPORT, SQL, STDIZE
Affymetrix Exon and Whole Transcript Expression CEL Input Engine	APPEND, FORMAT, G3D, IMPORT, KDE, SQL, STDIZE, UNIVARIATE
Affymetrix Expression CEL Input Engine	APPEND, FORMAT, G3D, IMPORT, KDE, SQL, STDIZE, UNIVARIATE
Affymetrix Expression CHP Input Engine	APPEND
Affymetrix miRNA CEL Input Engine	APPEND, FORMAT, G3D, IMPORT, KDE, LOESS, REG, SQL, STDIZE, SURVEYSELECT, UNIVARIATE
Affymetrix SNP CEL Input Engine	APPEND, FORMAT, G3D, IMPORT, KDE, SQL, STDIZE, UNIVARIATE

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Affymetrix SNP CHP Input Engine	APPEND, IMPORT, SQL, STDIZE
Affymetrix Tiling Bar Input Engine	IMPORT
Affymetrix Tiling CEL Input Engine	IMPORT
Agilent Input Engine	APPEND, IMPORT, SQL, STDIZE
Allele Specific Expression Filter	MULTTEST, REG,
ANOVA	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
ANOVA Normalization	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Append	APPEND
ArrayTrack Input Engine	APPEND, IMPORT, SQL, STDIZE
Batch Normalization	FASTCLUS, GCHART, GPLOT, MIXED, PRINCOMP
Batch Scoring	UNIVARIATE
Bin (Copy Number)	This process calls no additional SAS PROCs
Bin Intensities or Read Counts	This process calls no additional SAS PROCs
Bioconductor Expresso for Affymetrix	APPEND, IMPORT, SQL, STDIZE
Bivariate One-Way ANOVA	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Build Genotype Probability Data Set	IML, REG, SUMMARY
Calculate Square Root of Matrix	IML
Case-Control Association	ALLELE, APPEND, CASECONTROL, FORMAT, MULTTEST, PSMOOTH
Change Labels	RANK
Chromosome Color Theme	This process calls no additional SAS PROCs
Column Enrichment	APPEND, FREQ, GLMMOD, MULTTEST, RANK, TPSPLINE

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Combine Columns	This process calls no additional SAS PROCs
Combine Experiments	This process calls no additional SAS PROCs
Control Set Normalization	This process calls no additional SAS PROCs
Copy Number/LOH Control Set Adjustment	This process calls no additional SAS PROCs
Correlation and Grouped Scatterplots	SURVEYSELECT
Correlation and Principal Components	CORR, FACTOR, GCHART, GPLOT, MIXED, PRINCOMP
Create Annotation Analysis Group Variable	This process calls no additional SAS PROCs
Cross Correlation	APPEND, CORR
Data Filter	CORR, MIXED, MULTTEST
Data Standardize	STDIZE, UNIVARIATE
Difference Chooser	MIXED
Distribution Analysis	KDE
Discriminant Analysis	APPEND, CATALOG, CORR, DISCRIM, FASTCLUS, FREQ, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Distance Matrix	CLUSTER, DISTANCE, TREE
Distance Scoring	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Estimate Builder	MIXED
Experimental Design File Builder	OPTIONS
Export to GSEA Format	FREQ
Factor Analysis Normalization	FACTOR
Feature Flagger	FREQ, REPORT

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Filter Intensities	UNIVARIATE
Gene Model Summary	This process calls no additional SAS PROCs
Gene Set Enrichment	FREQ, GLMMOD, MULTTEST, RANK, SQL
GenePix Input Engine	APPEND, IMPORT, SQL, STDIZE
General Linear Model Selection	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GLIMMIX, GLMMOD, GLMSELECT, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Haplotype Estimation	ALLELE, APPEND, FORMAT, HAPLOTYPE, PSMOOTH
Haplotype Trend Regression	GLM, GLIMMIX, HAPLOTYPE, LOGISTIC, MULTTEST, PHREG, REG
HapMap Input Engine	IMPORT
Haseman-Elston Regression	ALLELE, APPEND, FORMAT, MIXED, PSMOOTH
htSNP Selection	ALLELE, APPEND, FORMAT, HTSNP, PSMOOTH
Illumina CN Input Engine	APPEND, IMPORT, SQL, STDIZE
Illumina Expression Input Engine	APPEND, IMPORT, SQL, STDIZE
Illumina miRNA Input Engine	APPEND, IMPORT, KDE, LOESS, REG, SQL, STDIZE, SURVEYSELECT, UNIVARIATE
Illumina SNP Input Engine	APPEND, IMPORT, SQL, STDIZE
IM and CIM Analysis	IML, REG, REPORT, SUMMARY
Import a Designed Experiment from Text, CSV or Excel Files	APPEND, IMPORT, SQL, STDIZE
Import Individual Text, CSV or Excel Files	APPEND, IMPORT, SQL, STDIZE
Imputed SNP (Tall Format) Input Engine	IMPORT
Imputed SNP-Trait Association	ALLELE, FORMAT, GLIMMIX, LOGISTIC, MIXED, PHREG, PSMOOTH, SURVEYFREQ
Imputed SNP (Wide Format) Input Engine	IMPORT
K Matrix Compression	CLUSTER, IML, MIXED, SQL, TREE

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
K-Means Clustering	FASTCLUS
K Nearest Neighbors	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
KEGG Pathway Color	UNIVARIATE
Kinship Matrix	IML, INBREED
LD Block Creation	ALLELE, IML
LD tagSNP Selection	ALLELE, APPEND, FORMAT, IML, PSMOOTH
Linkage Disequilibrium	ALLELE, APPEND, FORMAT, PSMOOTH
List Enrichment	FREQ, IMPORT, SQL
Loess Normalization	LOESS
Logistic Regression	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GENESELECT, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Malecot LD Map	ALLELE, APPEND, FORMAT, NLMIXED, PSMOOTH
MANOVA	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Marker Properties	ALLELE, FORMAT, PSMOOTH
Marker-Trait Association	ALLELE, APPEND, FORMAT, GLIMMIX, LOGISTIC, MIXED, PHREG, PSMOOTH
Merge	SQL
Merge Gene Sets	IMPORT, SQL
Missing Genotypes by Trait Summary	ALLELE, APPEND, CASECONTROL, FORMAT, PSMOOTH
Mixed Model Analysis	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Mixed Model Normalization	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Mixed Model Power	MIXED

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Multidimensional Scaling	APPEND, MDS
Multiple SNP-Trait Association	GLIMMIX, GLM, LOGISTIC, MIXED, PHREG, PRINCOMP, PSMOOTH
One-Way ANOVA	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
Parse a Column	This process is a .jsl script and calls no SAS PROCs
Partial Least Squares	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, PLS, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Partial Least Squares Normalization	GLMMOD, PLS
Partition (Copy Number )	FREQ, GENESELECT, SQL
Partition Trees	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GENESELECT, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
PCA for Population Stratification	ALLELE, APPEND, CORR, FORMAT, GLMMOD, IML, PLS, PRINCOMP, PSMOOTH, STDIZE, SUMMARY
Pleiotrophic Association	ALLELE, GLM, PSMOOTH
PLINK Input Engine	APPEND, IMPORT, SQL, STDIZE
P-Value Adjustment	APPEND, MIXED, MULTTEST, STDIZE, TPSPLINE, UNIVARIATE
P-Value Browser	APPEND, FORMAT, MULTTEST, PSMOOTH, TPSPLINE
P-Value Combination	This process calls no additional SAS PROCs
Pedigree Input Engine	APPEND, IMPORT, SQL, STDIZE
Phenotype Summary	FREQ
Population Genetic Distance	ALLELE, APPEND, FORMAT, PSMOOTH
Population Measures	ALLELE, APPEND, FORMAT, PSMOOTH
Principal Components Analysis	APPEND, GLMMOD, PLS, PRINCOMP, STDIZE, SUMMARY

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Principal Component Scoring	IML, PLS
Pseudolmage	APPEND, STDIZE, UNIVARIATE
Q-K Mixed Model	ALLELE, FORMAT, GLIMMIX, LOGISTIC, MIXED, PSMOOTH, SURVEYFREQ
QuantArray Input Engine	APPEND, IMPORT, SQL, STDIZE
Quantitative TDT	ALLELE, APPEND, FAMILY, FORMAT, GLM, IML, INBREED, MIXED, PSMOOTH, UNIVARIATE
Radial Basis Machine	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
Rank Rows	RANK
Ratio Analysis	LOESS
Recode Genotypes	ALLELE, APPEND, FORMAT, PSMOOTH
Recode Missing Genotypes	This process calls no additional SAS PROCs
Relationship Matrix	CORR, DISTANCE, IML, PRINCOMP
Reorder	APPEND
SAM Input Engine	IMPORT, APPEND, SQL
SNP Interaction Discovery	APPEND, CATALOG, CORR, DISCRIM, DISTANCE, FASTCLUS, FREQ, GENESELECT, GLIMMIX, GLMMOD, IML, LOGISTIC, MIXED, MULTTEST, NPAR1WAY, RANK, REG, ROBUSTREG, SQL, STDIZE, STEPDISC, TPSPLINE, TTEST, UNIVARIATE
SNP Power	IML
SNP-SNP Interactions	FORMAT, GLIMMIX, LOGISTIC, MIXED, PHREG, PSMOOTH, SQL
SNP-Trait Association	ALLELE, APPEND, FORMAT, GLIMMIX, LOGISTIC, MIXED, PHREG, PSMOOTH
ScanAlyze Input Engine	APPEND, IMPORT, SQL, STDIZE
Single Marker Analysis	IML, REG, REPORT, SUMMARY
Spectral 2D Peak Find	IML

## JMP Genomics APs Call SAS PROCs

Process	SAS PROCs Called
Spectral 3D Align	KDE
Spectral 3D Plot	SPECTRVIEW
Split Experiments	This process calls no additional SAS PROCs
Statistics for Columns	SUMMARY
Subset and Reorder Genetic Data	ALLELE, APPEND, FORMAT, PSMOOTH
Surface Summary	FORMAT, G3D, KDE, UNIVARIATE
Survey SNP-Trait Association	ALLELE, APPEND, FORMAT, PSMOOTH, SURVEYFREQ, SURVEYLOGIC, SURVEYREG
Survival Analysis	APPEND, MULTTEST, PHREG, TPSPLINE
Survival Predictive Modeling	PHREG, MULTTEST
TDT	ALLELE, APPEND, FAMILY, FORMAT, PSMOOTH
Track BAR Chart	This process calls no additional SAS PROCs
Track Gene GFF	This process calls no additional SAS PROCs
Track Gene Text	This process calls no additional SAS PROCs
Track Gene Web	This process is a .jsl script and calls no SAS PROCs
Track SNP Web	This process is a .jsl script and calls no SAS PROCs
Two-Way Plotter	GCHART, GPLOT, GREPLAY
Variance Components	ALLELE, APPEND, FORMAT, IML, INBREED, MIXED, PSMOOTH, UNIVARIATE
Verify Gender of Samples	This process calls no additional SAS PROCs

# Chapter 9

## Troubleshooting

---

This troubleshooting guide can help you diagnose and resolve problems that you may encounter when running JMP Genomics. If problems persist after troubleshooting, contact JMP Technical Support at [support@jmp.com](mailto:support@jmp.com). Select **Genomics > Documentation and Help > Contacting Technical Support**, in JMP, for additional information.

Process	Problem	Suggested Cause/Resolution
Installation of JMP Genomics	The message: <i>Existing Client Found</i> is displayed in the Install Shield Wizard window, indicating that a preexisting copy of SAS has been found on a network server.	A pre-existing copy of SAS has been found that is configured to run as a <i>thin</i> Client from a network server. JMP Genomics will only work with a personal copy of SAS loaded on the same Client machine and configured to work locally. Contact JMP Technical Support ( <a href="mailto:support@jmp.com">support@jmp.com</a> ) for instructions and assistance in resolving this problem. Select <b>Genomics &gt; Documentation and Help &gt; Contacting Technical Support</b> , in JMP, for additional information.

Process	Problem	Suggested Cause/Resolution
Opening JMP Genomics	A message appears stating that your license has expired.	<p>Your license to run JMP Genomics software has expired. You must obtain updated license files to renew your license. You will need two license files for JMP Genomics: one to update the JMP license and one to update the SAS license. Most likely, you will have received these files in an e-mail. If not, send an e-mail to <a href="mailto:support@jmp.com">support@jmp.com</a> to request updated license files. The license files must be saved to your local hard drive. To update the JMP license, complete the following steps:</p> <ol style="list-style-type: none"> <li>1 Start JMP.</li> <li>2 JMP will open a dialog asking if you want to update your license. Click <b>Yes</b> to open the license renewal tool.</li> <li>3 Click <b>Open License</b>.</li> <li>4 Browse to the folder in which you saved the license files and click <b>Open</b>.</li> </ol> <p>To update the SAS license, complete the following steps:</p> <ol style="list-style-type: none"> <li>1 Click <b>Start &gt; Programs &gt; SAS &gt; SAS 9.2 License Renewal &amp; Utilities &gt; Renew SAS Software</b> to open the renewal tool.</li> <li>2 Browse to the folder in which you saved the license files and click <b>OK</b>.</li> <li>3 When prompted click <b>Renew</b>.</li> <li>4 Click <b>OK</b> in the confirmation window.</li> </ol>
Moving data between JMP Genomics (32 bit) and JMP Genomics (64 bit)	Processes take significantly more time to run than previously observed.	Apply the hot fix found at <a href="http://ftp.sas.com/techsup/download/hotfix/HF2/B25.html#B25026">http://ftp.sas.com/techsup/download/hotfix/HF2/B25.html#B25026</a> .
Any process in using SAS DATA composed in Microsoft Word and pasted into JMP Genomics	A syntax error is displayed.	<p>Microsoft Word uses a <i>Smart Quote</i> feature that converts quotation marks from the format needed by SAS into a format more appropriate for English language text. SAS DATA step commands should either be composed in a simple text editor and then pasted into the JMP Genomics DATA Step input fields or directly in the DATA Step input fields. DATA Step commands should never be composed using Microsoft Word.</p>

Process	Problem	Suggested Cause/Resolution
<p>Any JMP Genomics process that uses one or more SAS programs</p> <p>Any JMP Genomics process that uses one or more SAS programs</p>	<p>A SAS log is displayed in your JMP Genomics session along with a message preceded by ERROR.</p>	<p>The generated SAS code might not complete successfully because of mis-specified parameters.</p> <p>Most of the error messages should be self-explanatory and provide some idea about what to do next. If not, examine the broader context provided by the SAS log to determine the problem. If this fails, consult and search the SAS documentation for the SAS code generating the error by clicking <b>Help &gt; SAS Documentation - Local</b> or <b>Help &gt; SAS Documentation - Web</b>.</p> <p>There is also the possibility of a bug in the SAS macro code. If you have found what appears to be a bug, please send the SAS log and explanation to support@jmp.com. Please describe your procedure in sufficient detail for us to reproduce the problem. If you are a SAS programmer, you might wish to view and even edit the original SAS code in the ProcessLibrary and/or MacroLib folders. Please also feel free to send suggested changes to the code to support@jmp.com.</p>
	<p>A WARNING dialog appears, telling you that SAS is connected and a process is already running.</p>	<p>JMP Genomics can only run one process at a time and does not queue jobs. Click <b>OK</b> in the dialog to wait, disregard the second Run, and let the current process continue running. Click <b>View Log</b> to view the current SAS log to get information about the current process. Click <b>Stop</b> to stop the current process. If the SAS process does not stop in a short period of time, it is okay to kill the sas.exe process directly from Windows Task Manager, and then click <b>Stop</b> again.</p>
	<p>A process runs longer than expected or produces no output.</p>	<p>In this situation, perform the following steps:</p> <ol style="list-style-type: none"> <li>1 Click <b>Run</b> again. A WARNING: SAS is Connected window should appear.</li> <li>2 Click <b>View Log</b>. If any SAS ERROR messages appear, click <b>Stop</b> to stop the current process. If not, proceed to the step below.</li> <li>3 View the SAS log that is displayed in the JMP Log window to see the most recently executed code. You can continue to click <b>View Log</b> as many times as you like to check the status of the SAS program. Alternatively, you can monitor generated file activity in the SAS working folder. To open this folder while the analysis is running, click <b>Open SAS Temporary Folder</b>. You should see various files being generated as the process runs. On Windows, press <b>F5</b> to refresh the folder while you are monitoring it.</li> </ol> <p>If these steps do not help, try running the process in the SAS 9.2 Display Manager as described below.</p>

Process	Problem	Suggested Cause/Resolution
Any JMP Ge-nomics process	Output of the process does not automatically open.	The output filename might contain the following characters: (), @, ^ and &, any place of output name, or contains [] at the beginning of the name, (such as [name], for example). If these characters are present, you can open the output by completing the following steps: 1 Navigate to the specified output folder. 2 Double-click on the <code>sasclean.jsl</code> script in the folder. All of the output should open.
Processes that perform repetitive computations	The SAS log is too long and is truncated.	In either of these cases, an alternative way to debug the process is to open the <code>.sas</code> file in the SAS 9.2 Display Manager (right-click and select <b>Open with SAS 9.2</b> ) and run it from there by pressing <b>F3</b> . The SAS Display Manager provides options for saving or deleting sections of long logs. On Windows operating systems, you can alternatively right-click on a <code>.sas</code> file and select <b>Submit to SAS 9.2</b> . SAS will then run in batch mode and produce <code>.log</code> and <code>.lst</code> files.
Processes that specify a lot of variables into one macro	The line length can become too long for SAS batch mode.	
Processes using wide data sets composed of long lists of variables	Numerous <b>ERROR</b> messages are generated in the SAS log.	The SAS Macro text expression limit of 65534 bytes might have been exceeded. Work-arounds for this situation include the following: 1 Recreate the data set or rename the variables to have the shortest possible names. 2 Modify the process specification to have list-style input for long lists of variables, such as <code>Col1-Col20000</code> . 3 Reduce the number of variables using K-Means Clustering, as follows. Transpose the data to tall form using <b>Transpose Rectangular</b> , run <b>K-Means Clustering</b> to generate a few thousand or less clusters, retain representatives from each cluster to use as the data, and then transpose back to wide form using <b>Transpose Rectangular</b> .

Process	Problem	Suggested Cause/Resolution
Opening a data file using the <b>File &gt; Open</b> command in any JMP Genomics process	The column names listed in the Available Variables box of a dialog appear different than the original column names in the data set.	SAS uses two ways to name a column: the variable label and the variable name. When a file is opened using the <b>File &gt; Open</b> command from the JMP menu, SAS variable labels will be displayed. These might differ from those displayed in the Available Variables list in the JMP Genomics process dialogs, which display SAS variable names for the available variables. To solve the problem, open the data file using the <b>Open</b> button on the process dialogs. This displays the table with names the same as those in the Available Variables lists. Alternatively, use the <b>File &gt; Open</b> command from the JMP menu and, in the Open Data File dialog, change File of type to SAS Data Sets and click the Use SAS Variable Names for Column Names check box.
Affymetrix Expression CHP Input Engine	An ERROR message, indicating that the specified EDF does not exist, is generated.	This message has been observed when an .xls file has been specified as the EDF if either the EDF is open in Excel or has been saved using a legacy (Excel 4.0 or older) format. To resolve this problem, first make sure the file is not open. Second, open the file in Excel and save it in the most recent format. Alternatively, open the file in JMP and save it as a SAS data set using the command <b>File &gt; Save As SAS Data Set</b> .
Affymetrix Expression CEL Input Engine	An ERROR message is displayed stating: The macro variable name is either all blank or missing.	Files selected on the Library Files tab have been mis-specified. Review your selections and correct as needed.
Agilent Import Engine	Running the process generates a long ERROR message along with a SAS Log and a SAS Message dialog indicating the successful generation of the SAS Data Set, EDDS and an Annotation data set.	The process has run successfully despite the appearance of the ERROR message. The likely cause of the ERROR message is the presence of non-numeric character strings in numerical columns. For example, Agilent places the string #IND, in empty numeric cells to indicate missing values. When SAS imports the data from these files, it reports an error and replaces the character string with a period (.). Open the resulting data sets to verify they are as you intended. If so, you can safely ignore the ERROR message and proceed with the data analysis.

Process	Problem	Suggested Cause/Resolution
Bioconductor Expresso for Affymetrix Import Engine	An ERROR message is generated when you try to choose an input data set using the Universal/Uniform Naming Convention (UNC).	The Bioconductor Expresso wrapper does not accept the Universal/Uniform Naming Convention (UNC) for describing the location of a volume, directory or file. The UNC format is (\\directory\subdirectory\file). To avoid using a UNC formatted path, do not begin navigating to the desired files/folders by clicking on the directories shown in the box on the left side of the <b>Open Data File</b> window, as this will format the resulting path in the UNC. Instead, begin navigating by clicking within the <b>Look in:</b> box at the top of the window. The format of the resulting path (C:\Directory\Subdirectory\file) is acceptable to the Bioconductor Expresso process.
Any input engine	An ERROR message is generated when you try to use an EDF generated by the EDF Builder and saved as a text file.	JMP's Text Data File default Import setting for the <b>End of Field</b> is set to <b>Tab and Comma</b> and the export settings preference for the <b>End of Field</b> is set to <b>Comma</b> . If the EDF is saved as a .txt file and the fields end with commas instead of tabs, the format of the EDF is not recognized by the input engines. JMP Genomics' default Import and Export should both be set to <b>Tab</b> .
	A long ERROR message beginning with ERROR: Limit set by ERRORS= option reached. Further errors of this type will not be printed.	This message is generated when the numeric column containing one or more character strings such as three hyphens (---) or N/A to represent missing data. In many cases, the resulting data set is successfully generated. However, the heading associated with the columns might not be positioned in the correct columns. You should verify that each column contains the appropriate heading.

Process	Problem	Suggested Cause/Resolution
Import Individual Txt, CSV or Excel Files Input Engine	Genotypes in the input files whose alleles are delimited using a "/" are misrepresented as date/time data.	<p>You should check the initial data format before importing the data. Select <b>File &gt; Open</b>. An Open Data File window opens. Select <b>Text Import Preview</b> from the Files of Type drop-down menu and click <b>Open</b>. Count the columns listing information and the columns listing SNP genotypes.</p> <p>Define information about the number of non-marker columns and the number of marker columns as follows:</p> <ul style="list-style-type: none"> <li>☞ Complete the <b>General</b> and <b>Option</b> tabs of the dialog, as normal.</li> <li>☞ Type <code>Column1-Column<del>xx</del> (SNP1-SNP<del>yy</del>) (\$)</code> in the <b>List of Variable Names and Types</b> field on the <b>Wide Data</b> tab of the <b>Import Individual Txt, CSV or Excel Files</b> dialog.</li> </ul> <p><i>Note:</i> <b>xx</b> indicates the right-most column containing non-marker information (assumes that marker information begins in columns <b>xx+1</b>); <b>yy</b> indicates the total number of columns containing marker information.</p> <ul style="list-style-type: none"> <li>☞ Run the process as normal.</li> </ul> <p><i>Note:</i> This fix assumes that your data is in the <i>wide</i> format.</p>
Any Import process	A SAS log is displayed in your JMP Genomics session along with a message preceded by <b>ERROR</b> or there are notes in the SAS log indicating invalid data for particular variables.	<p>When importing a file to a SAS data set, SAS determines the type of variable (character or numeric) based on the first N observations, where N is the value provided in the <b>Number of Rows to Scan</b> parameter on the <b>Options</b> tab of most of the <b>Import</b> processes. Sometimes, when a character value is present after the first N observations and the previous observations have all been numeric (so that the variable has already been defined as numeric), an error occurs when SAS attempts to read this character value.</p> <p>Try increasing the value for N in the <b>Options</b> tab until you no longer see these notes in the log.</p>
	The values in one or more columns are truncated.	<p>When importing a file to a SAS data set, SAS determines the length of variable (character or numeric) based on the first N observations, where N is the value provided in the <b>Number of Rows to Scan</b> parameter on the <b>Options</b> tab of most of the <b>Import</b> processes. Sometimes, when subsequent values are longer than those in the first N observations, SAS will truncate those values to the length determined for the N observations.</p> <p>Try increasing the value for N in the <b>Options</b> tab.</p>

Process	Problem	Suggested Cause/Resolution
Learning Curve Model Comparison AP	<p>No results displayed, no SAS error shown, but an error in the JMP log, such as:</p> <pre>Unable to open in ReadOnly mode. The system cannot find the file specified. in access or evaluation of 'Include', Include("C:\Program Files\SAS\JMP9\Genomics\WorkflowResults\cvlc_plot.js1")</pre>	<p>SAS Enterprise Miner, can cause PROC GENESELECT to end abnormally. Check to see whether you have SAS Enterprise Miner installed on your machine.</p> <p>To fix the problem, rename C:\Program Files\SAS\SAS 9.2\dmine\sasmsg\dmine.msg and replace it with C:\Program Files\SAS\SAS 9.2\genetics\sasmsg\dmine.</p>

Process	Problem	Suggested Cause/Resolution
KEGG Pathway Analysis	An ERROR message is generated stating: You selected to use proxy server to access web, but did not specify proxy server name or port number. Please run Configure Proxy Settings to set the value.	<p>The Proxy Server or Proxy Port number could have been incorrectly specified.</p> <p><i>Note:</i> A proxy server is a server that sits between a Web browser and an external server on the Web to filter requests, improve performance, and share connections. If your computer accesses the Internet through a proxy server, you must reset your proxy name and port number before you run the KEGG Pathway Analysis process.</p> <p>Run the <b>File &gt; Configure Genomics Settings</b> process as described in <a href="#">Configuring JMP Genomics Settings</a>. Make sure the correct name and number are entered in the dialog and click <b>Run</b> to configure your settings.</p>
	An ERROR message is generated stating: ERROR: KEGG throws RemoteException when searching pathways for hsa04520. Please refer to the Java log for further details.	<p>The KEGG API server is either down, very busy, or the connection to the KEGG API server is denied.</p> <p>Retry the process at another time.</p>
KEGG Pathway Analysis AP	An ERROR message is generated stating: ERROR: Failed to find genomics.config file.	<p>Check to see whether the genomics.config file is missing from the &lt;sasroot&gt;\sds\sasmisc directory (the default &lt;sasroot&gt; is C:\Program Files\SAS\SAS SASFoundation\9.2).</p> <p>If the config file is missing, reinstall your JMP Genomics. The install copies this configuration file to: C:\Program Files\SAS\SAS SASFoundation\9.2\sds\sasmisc.</p>

Process	Problem	Suggested Cause/Resolution
Any Workflow AP	Attempts to run a second AP or new Workflow fails. The JMP Log shows the following message: A second script is attempting to execute, possibly during a nested click event. It may be necessary to press Escape to terminate the previous script.	Press <b>Esc</b> to exit the JMP Script.
Saving a JMP table as an Excel spreadsheet	The following error message is surfaced: [Microsoft] [ODBC Excel Driver] Cannot modify the design of table 'Abrasion'. It is in a read-only database.	The ODBC setting must be changed using the following procedure: <ol style="list-style-type: none"> <li>1 Click <b>Start &gt; Settings &gt; Control Panel</b>.</li> <li>2 Double-click <b>Administrative Tools</b>.</li> <li>3 Double-click <b>Data Sources(ODBC)</b> to open the ODBC Data Source Administrator dialog.</li> <li>4 Select Excel Files in the Use Data Sources panel and click <b>Configure...</b> to open the ODBC Microsoft Excel Setup dialog.</li> <li>5 Click <b>Options</b> and uncheck the Read Only check box.</li> <li>6 Click <b>OK</b> on the open dialogs.</li> </ol>

# Chapter 10

## Glossary

---

### Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Accession Number</b>	Numerical identifier for a gene, marker, or gene product in a public data base
<b>Allele</b>	Any variant form of a nucleic acid sequence or marker at a particular locus
<b>Alpha</b>	Significance level. While <i>alpha</i> can be any value between 0 and 1, it is typically set at either 0.01, 0.05 or 0.10.
<b>Annotation Data Set</b>	Data set containing a variety of information on the identity, biological function, pathway association, etc., for the genes, gene fragments, gene products, probe sets, genetic markers, etc., under investigation.
<b>Association</b>	The statistically significant co-occurrence of two or more phenomena. In the context of genome-wide association studies (GWAS), mapping of a gene for a particular trait or disease is performed by detecting significant associations between the trait and marker genotype.
<b>AUC</b>	Area under the receiver operating curve (ROC) statistic that is used to assess the ability of a model to predict future responses. For most models, the greater the AUC, the better the model is for making accurate predictions.
<b>Bin</b>	A group of related or functionally similar observations that are considered as a unit for statistical analysis.
<b>Binary Variable</b>	A variable that contains two discrete values, for example, 0 and 1.
<b>Bivariate</b>	Involving two variables.
<b>Bootstrap</b>	The practice of using with-replacement empirical distribution of observations to estimate the statistical properties of the population from which the observations were made.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Box Plot</b>	Used to display the response distribution at different combinations of factor levels. Box plots can reveal differences in the response mean at different levels, suggesting main effects. Box plots can also reveal whether the response variation is homogenous across factor levels, an assumption made in analysis of variance.
<b>BY Group</b>	All of the observations with the same values for all BY variables.
<b>BY Variable</b>	An optional (in most APs) variable specification whose values define groups of observations, such as hour, month, or year. Specifying a BY variable allows you to animate an image so that you can see how response values change according to some grouping, like over time. Alternatively, BY variables can enable analyses to be performed separately on different groups as defined by a variable such as gender.
<b>Cell Plot</b>	A heat map or color map display, mapping colors to values of variables and displaying them in a rectangular grid.
<b>Censor Variable(s)</b>	These columns specify those observations for which data have been censored or truncated. For example, investigations of the effects of certain genes on life span may be terminated before all of the individuals have expired. The ultimate life spans for these individuals are unknown. All that can be said is that they exceed the period of the study. These data are considered <i>censored</i> .
<b>Centimorgan (cM)</b>	Also referred to as a <i>map unit</i> . A measure of genetic distance between loci; one centimorgan is equivalent to the physical separation between loci needed for a recombination frequency of 1%. Recombination between loci is affected by a variety of molecular and biochemical factors, in addition to physical distance, and the centimorgan should not be considered as representing a linear measurement.
<b>Character Variable</b>	A variable whose values can consist of alphabetic and special characters as well as numeric characters
<b>Check Box</b>	An item in a dialog or window that you can select without affecting any other items. You can deactivate a check box by selecting it again.
<b>Cholesky Decomposition</b>	A mathematical method for taking the square root of a symmetric matrix. The Cholesky root of a matrix $A$ is $L$ , where $A = LL'$ , and $L$ is a lower-triangular matrix. It is useful in QTL analysis because it allows modeling of a pedigree-induced covariance structure via a mixed model.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Class Variable</b>	The variable whose values define the groups for analysis. Class variables can have continuous values, but they typically have a few discrete values that define the classifications of the variable. Values may either be character or numeric.
<b>Clustering</b>	The process of dividing a data set into mutually exclusive groups such that the observations for each group are as close as possible to one another, and different groups are as far as possible from one another.
<b>Color Variable</b>	A variable whose values are used to specify how the graphical output of an analysis is to be colored
<b>Composite Interval Mapping (CIM)</b>	Method for mapping of a target QTL for a trait. It uses markers, located elsewhere in the genome, that have previously been shown to be associated with additional QTLs for that trait. Multiple analysis points across each inter-locus interval for the target QTL are assessed.
<b>Copy Number Variation (CNV)</b>	Any deletion, insertion, duplication or other variant in the DNA sequence of a genome that results in that sequence being present in greater or lesser numbers relative to those seen in a normal, reference genome. A CNV can result from relatively simple duplications, either tandem or inverted, or deletion of small or large blocks of sequence. They may also be more complex, involving gains or losses of homologous sequences at multiple sites in the genome. Disruption of contiguous gene sequences and altered gene dosages by CNVs have been shown to influence gene expression, increase phenotypic variation and cause disease.
<b>Covariance</b>	A measure of the relationship between two variables. It equals the correlation between the two variables times the square roots of their variances.
<b>Covariate</b>	An independent variable, not manipulated by the experimenter, that may influence the outcome of the experiment.
<b>Cross Validation</b>	A statistical method for evaluating how well a model will predict the outcome of additional experiments.
<b>Delimiter</b>	One or more characters that separate the designations for the different alleles in a genotype. JMP Genomics frequently uses a forward-slash (/) as a delimiter.
<b>Dependent Variable</b>	A variable whose value is determined by the value of another variable or by the values of a set of variables. This variable lists the responses you measure. In a two-dimensional plot, the dependent variable is usually plotted on the <i>y</i> (horizontal) axis.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Dialog</b>	An interactive window that allows you to set parameters for and run an analytical process.
<b>Distance Matrix</b>	A matrix of distances.
<b>Drill down</b>	To start at one level of a dimension hierarchy and to click through one or more lower levels until you reach the data that you are interested in.
<b>Eigenvalue</b>	A scalar value that determines by how much a corresponding eigenvector is scaled by the square matrix for which it is defined. In principal component analysis, the eigenvalues of the covariance or correlation matrix represent the variance of the components.
<b>Eigenvector</b>	For a given square matrix, a non-zero vector that changes length, but not direction, when multiplied by the matrix. The computation of principal components for a set of variables uses the eigenvectors of the variables' covariance or correlation matrix.
<b>ESTIMATE Statement</b>	A programming statement for certain SAS procedures (e.g. PROC MIXED) used to specify parameters used in both ANOVA and mixed model analyses to test an arbitrary set of linear hypotheses regarding the relative importance of different combinations of fixed effects parameters.
<b>Exon</b>	The portions of the coding region of a gene that make up the mature mRNA. Exons are separated in the genomic DNA by intervening sequences (introns) that are transcribed with the exons. Introns are excised and the exons are spliced together during post-transcriptional processing and mRNA maturation.
<b>Expression</b>	The process by which the information contained in a gene is used to make a functional RNA or protein
<b>Experimental Design Data Set (EDDS)</b>	<p>An EDDS is a SAS data set that provides information about the columns of a tall data set. It describes relevant experimental variables such as treatment conditions and covariates as well as a variable named <code>ColumnName</code>. Entries in the <code>ColumnName</code> column must exactly match the column names in the input tall data set. EDDSs have certain constraints that must be followed for the processes to run successfully.</p> <p>An EDDS is required by most processes using a tall input data set. Many of the input engines that generate a tall data set from raw data files also automatically generate the needed EDDS.</p>

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Experimental Design File (EDF)</b>	<p>A file which provides JMP Genomics with important information about how an experiment was carried out. It defines experimental variables such as treatment conditions and covariates and provides the basis for organizing and analyzing your data.</p> <p>An EDF is required by many of the input engines for the construction of a SAS data set from the raw data files. An EDF also serves as a precursor to the experimental design data set</p>
<b>Factor</b>	<p>Also referred to as an independent or predictor variable, a factor is a variable included in a model to account for variation in a response. Factors are the variables whose values (levels) you set to study their relationship to a response. You often experiment with many potentially influential factors at the same time.</p>
<b>False Discovery Rate</b>	<p>The expected percentage of a set of predictions that are assumed to be false. For example, if an analysis which predicts the association of 10 genes with a particular trait has a false discovery rate of 0.1, you can expect 9 of the predictions to be correct.</p>
<b>Field</b>	<p>A window area in which you can view, enter, or modify a value</p>
<b>Fixed Effects</b>	<p>The effects that drive the variation that you are interested in assessing that have a fixed number of well-defined levels. They can also include nuisance variables that you need to consider in your model. Fixed effects include factors such as experimental treatment, disease status, age/developmental status of the test organisms, gender. Variation due to fixed effects is the variation you are interested in estimating and must be kept in the analysis.</p>
<b>Genotype</b>	<p>The genetic complement an individual possesses at a particular locus.</p>
<b>Group Variable</b>	<p>A variable that is used for grouping results.</p>
<b>Haplotype</b>	<p>The combination of alleles that is transmitted from one generation to the next as a single unit on a chromosome.</p>
<b>Hardy-Weinberg Equilibrium</b>	<p>The state of a population in which gene frequencies and genotypic ratios remain constant from generation to generation.</p>
<b>Heat Map</b>	<p><i>See Cell Plot.</i></p>
<b>Hold-out Data</b>	<p>A portion of the data that is set aside during model development. Hold-out data can be used as test data to benchmark the fit and accuracy of the emerging predictive model.</p>

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Hoeffding Correlation (D)</b>	A nonparametric measure of association that detects general departures from independence. This statistic approximates a weighted sum over observations of <i>chi</i> -square statistics for two-by-two classification tables.
<b>Imputation</b>	The computation of replacement values for missing input values.
<b>Inbreeding Coefficient</b>	The probability that an individual's two alleles at any locus are identical by descent.
<b>Independent Variable</b>	This variable does not depend on the value of another variable; it represents the condition or parameter that is manipulated by the investigator. In a two-dimensional plot, the independent variable is usually plotted on the <i>x</i> (horizontal) axis.
<b>Index Variable</b>	One or more columns specifying how the observations are to be classified.
<b>Interval Mapping (IM)</b>	Method for mapping of a target QTL for a trait between two flanking markers.
<b>Jitter</b>	A random shifting of points by a slight amount along an axis so that more of those points can be effectively visualized in a graphical display.
<b>K_Rho</b>	The information for the LD measure <i>Rho</i> .
<b>Kendall Correlation</b>	A metric used to measure the degree of correspondence between two sets of rankings where the metrics used to assess each set of rankings are not equivalent.
<b>Kinship (Coancestry) Coefficient</b>	The probability that two alleles from a single locus in a randomly selected pair of individuals are identical by descent.
<b>K-Means</b>	A statistical method that creates optimally-separated groups of observations of a data using one of several methods. A set of points called cluster seeds is selected as a first guess of the means of the clusters. One cluster seed is selected for each of <i>k</i> clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters, and the process is repeated until no further changes occur in the clusters.
<b>Label Variable</b>	A column containing descriptive labels of up to 40 characters that can be printed in the output by certain procedures instead of, or in addition to, the variable name.
<b>Leaf</b>	In a tree map, a leaf is any segment that is not further segmented. The final leaves in a tree are called terminal nodes.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Level</b>	A successive hierarchical partition of data in a tree map. The first level represents the entire unpartitioned data set. The second level represents the first partition of the data into segments, and so on.
<b>Linkage</b>	The property by which alleles located on the same chromosome tend to be inherited together. The closer two loci are on the chromosome, the more tightly they tend to be linked.
<b>Linkage Disequilibrium (LD)</b>	<p data-bbox="460 513 1123 543">A measure of the association between two alleles at separate loci.</p> <hr/> <p data-bbox="460 574 1107 604"><i>Note:</i> A high LD does not imply that loci are physically linked.</p> <hr/> <p data-bbox="460 635 1231 725">An association (either positive or negative) between alleles can occur even if the loci are not located on the same chromosome, provided other factors affecting the population (directional selection, for example) are in effect.</p>
<b>Lod Score</b>	A logarithmic statistical estimate of the probability that two loci are located proximal to each other on a chromosome. A <i>Lod</i> score of 3, which indicates that two loci are 1000 times more likely than not to lie close to each other, is generally considered minimal for significance.
<b>Loess Normalization</b>	Method for eliminating non-biological bias and variation from microarray data by fitting a local regression curve to expression data. Method assumes that the majority of the genes in the study are not differentially affected by the experimental conditions.
<b>Loss of Heterozygosity (LOH)</b>	Results from a deletion (or other mutation) of the normal allele, at a locus heterozygous for the normal allele and a deleterious mutant allele. Produces a locus that is either homozygous or hemizygous for the deleterious allele.
<b>LSMeans</b>	Least squares means, which are estimates of means of classification effects that would be observed if the experimental design is balanced.
<b>Main Effect</b>	An effect measures the extent to which the response depends on the factors involved in the effect. A main effect is the change in the response due to a single factor. For two-level factors, the main effect is the difference between the mean response at the high level of a factor and the mean response at its low level.
<b>Marker Variable</b>	The column listing each individual's allele or genotype for the genetic markers used in an analysis. If marker variables are listed as alleles, there is a pair of marker variables for each marker.
<b>Mean</b>	Mathematical average for a collection of $n$ observations. It is calculated by dividing the sum of the observations by $n$ .

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Median</b>	In any set of $n$ observations arranged in order of magnitude, the median is represented by the observation positioned at $n/2$ .
<b>Menu Bar</b>	The primary list of items at the top of a window which represent the actions or classes of actions that can be executed. Selecting an item executes an action, opens a pull-down menu, or opens a dialog box that requests additional information.
<b>Missing Value</b>	A value in the SAS System indicating that no data is stored for the variable in the current observation. It is indicated by a single dot (•) for a numeric variable or a blank for a character variable.
<b>Mixed Model</b>	A statistical model containing both fixed and random effects.
<b>Model</b>	A formula or algorithm that computes output values from input values.
<b>Monophyletic</b>	Describes any group descended from a common ancestor that includes the ancestor and all of its descendents.
<b>Nominal Variable</b>	A variable that contains discrete values that do not have a logical order. Includes names and other verbal descriptions.
<b>Numeric Variable</b>	A variable that contains only numeric values and related symbols, such as decimal points, plus signs, and minus signs.
<b>Observation</b>	A row (horizontal component) in a SAS data set. Each observation contains one data value for each variable in the data set.
<b>Ordinal Variable</b>	A variable that contains discrete values that have a logical order. For example, a variable called <b>Rank</b> could have values such as 1, 2, 3, 4, and 5.
<b>Overfit</b>	To train a model to the random variation in the sample data. Overfit models contain too many parameters (weights), and they do not generalize well
<b>PCTL</b>	Percentile.
<b>Partial Least Squares (PLS) Coefficient</b>	A statistical technique that simultaneously partitions variability of both $X$ and $Y$ variables, somewhat similar to principal components.
<b>Pearson Correlation</b>	A statistical measure of the linear relationship between two variables. It is reported as some value between +1 and -1. A value of +1 indicates a perfect, direct, correlation between the variables; a value of -1 indicates a perfect, inverse, correlation between the variables.
<b>Pedigree</b>	The ancestral lineage of an individual or group of closely related individuals

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Phenotype</b>	The physical manifestation of a genotype.
<b>Population Stratification</b>	The phenomenon in which differences in allele frequencies between cases and controls in studies of genetic diseases that might have been ascribed to association of specific genes with disease are instead found to result from systematic differences in ancestry of the experimental groups.
<b>Power</b>	The probability of a statistical significance test allowing you to reject the null hypothesis when the alternative hypothesis is true.
<b>Predictor</b>	A function or variable used to estimate a response.
<b>Predictor Class Variable</b>	An independent variable used to predict a dependent variable. Its distinct levels correspond to different predictions, and are often modeled by constructing a set of 0-1 binary variables corresponding to each distinct level.
<b>Predictor Continuous Variable</b>	A numeric independent variable that predicts a dependent variable. Its predictions are computed directly as a function of the numeric variable, as in a linear regression.
<b>Predictive Model</b>	A statistical tool, made up of informative variables, that is used to forecast future behaviors or responses.
<b>Principal Components</b>	Linear combinations of all of the original variables that maximally explain variability. Useful when the analysis considers effects of many variables at once. By combining the variables into groups, you can reduce the total number in any one analysis.
<b>Probe Intensity</b>	Strength of the signal generated by hybridization of the target sequence to the specific probe set on the microarray.
<b>Probe Set</b>	All of the different probes specific for a transcript. As an internal control, microarrays typically include multiple target sequences taken from different regions of a transcript. Comparisons of intensities resulting from the hybridization of transcripts to each of these sequences allows for more effective evaluation of transcript levels. The collection of targets specific for a transcript are referred to as a probe set.
<b>PROC</b>	A SAS procedure; a group of SAS statements that call and execute a procedure, usually with a SAS data set as input
<b>Pull-down Menu</b>	The list of menu items or choices that appears when you choose an item from a menu bar or from another menu.
<b>p-Value</b>	The statistical probability that a statistic is as or more extreme than the observed value, assuming the null hypothesis is true. A smaller <i>p</i> -value allows you to more rigorously reject the null hypothesis.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Quantile</b>	Portions taken at regular intervals along a distribution that divide a data set into discrete subsets.
<b>Quantitative Trait Locus (QTL)</b>	A region of DNA that is associated with the strength of a particular phenotype. By themselves, QTLs do not determine whether or not a gene is expressed. Instead, each QTL interacts with other QTLs, located throughout the genome, to influence the relative level of expression.
<b>Random Effects</b>	The effects that cause extraneous variation in your results and have little to do with the questions being addresses. Random effects include factors such as physical differences between the arrays, or batch effects resulting from performing different parts of the experiments at different times, on different days, using different lots of reagents, etc. Variation resulting from random effects can confound your results and should be eliminated from your analysis.
<b>Random Number Seed</b>	The starting point for a random number generator. Unless a number is specified, an arbitrary value, such as the date/time of an event, is used.
<b>Residuals</b>	Values equal to the response values minus the predicted values.
<b>Rho</b>	<i>See</i> Spearman Correlation. Also, a measure of linkage disequilibrium.
<b>Root Mean Square Error (RMSE)</b>	A measure of the differences between the values predicted by a model or an estimator and the values actually observed. It is calculated by taking the square root of the mean square error value.
<b>SAS Data Set</b>	A file whose contents are in one of the native SAS file formats. SAS data sets contain data values in addition to descriptor information that is associated with the data.
<b>SAS Log</b>	A file that contains all the SAS statements you have submitted, messages about the execution of your analytical process and the SAS program running in the background, and in some cases, output from certain procedures. This file is generated and placed in the designated output folder.
<b>Single Nucleotide Polymorphism (SNP)</b>	DNA variant in which a sequence differs from the wild-type or other reference sequence by a single nucleotide.
<b>Smoothing Bandwidth</b>	A number determining the degree of smoothing for certain algorithms.
<b>Spearman Correlation</b>	Non-parametric method for examining whether two quantitative variables co-vary. Each pair of variables are converted to ranks and are linked with an “unseen” nominal variable.

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Standard Deviation</b>	A statistical measure of how “spread out” the data are. It is calculated by taking the positive square root of the sum of the squared deviations of each observation from the sample mean divided by $(n-1)$ .
<b>Standard Error</b>	The standard deviation of the sample mean. It is calculated by dividing the standard deviation by the square root of the sample size.
<b>Strata Variable</b>	A variable that partitions the data into blocks with similar characteristics.
<b>Tau Value</b>	A nonparametric measure of association based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently. Also, used for the truncated product $p$ -value adjustment method to indicate that there is at least one false null hypothesis among those with $p$ -values less than Tau when the null hypothesis is rejected.
<b>Test Data Set</b>	<i>See</i> Hold-out Data.
<b>t-statistic</b>	A measure of how extreme a statistical estimate is. It is calculated by subtracting a reference of hypothetical value from your estimate and then dividing the remainder by the standard error value for the experiment.
<b>t-test</b>	A test that assesses the statistical difference between the means of two different experimental groups.
<b>Transmission Disequilibrium Test (TDT)</b>	A family-based association test used to map binary traits.
<b>Training Data Set</b>	The portion of the initial data set that contains input values and target values that are used to develop a predictive model.
<b>Transcript Cluster</b>	A consensus sequence of nucleotide bases made up of the cluster of exons transcribed from a particular strand of a defined region of the genome.
<b>Transformation</b>	The process of applying a function to a variable in order to adjust the variable's range, variability, or both
<b>Tree</b>	The complete set of rules that are used to split data into a hierarchy of successive segments. A tree consists of branches and leaves, in which each set of leaves represents an optimal segmentation of the branches above them according to a statistical measure.
<b>Variable</b>	A column (vertical component) in a SAS data set. The data values for each variable describe a single characteristic for all observations

## Definitions/Explanations of Terms Used in JMP Genomics Documentation

Term	Definition/Explanation
<b>Variance</b>	A measure of deviation of a group of samples from the mean. It is calculated by squaring the standard deviation.
<b>Venn Diagram</b>	A graphical representation composed of two or more overlapping circles that shows all of the hypothetical relationships between two or more data sets.
<b>Volcano Plot</b>	A scatter-plot of the negative $\log_{10}$ -transformed $p$ -values derived from gene-specific $t$ test against the $\log_2$ -fold change in expression. Genes whose expression is decreased lie to the left of the mean; genes whose expression is increased lie to the right of the mean. Genes with statistically significant differential expression lie above a horizontal threshold. This plot provides an effective means for visualizing the direction, magnitude, and significance of changes in gene expression.
<b>Where Clause</b>	A SAS statement that allows you to filter a set of observations so that only the subset of data meeting the specific filtering criteria are considered in the analysis.
<b>Wizard</b>	An interactive utility program that consists of a series of dialog boxes, windows, or pages. You supply information in each dialog box, window, or page, and the wizard uses that information to perform a task.

# Chapter 11

## References

---

- Abecasis, G.R., W.O.C. Cookson, and L.R. Cardon. (2000). Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics* 8: 545-551.
- Agresti, A.(1990). Categorical Data Analysis. John Wiley & Sons, Inc. New York, NY.
- Allison, D.B. (1997). Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics* 66: 279-292.
- Allison, D.B., M. Heo, et al. (1999) Sibling based tests of linkage and association for quantitative traits. *American Journal of Human Genetics* 64: 1754-1764.
- Benjamini, Y. and Y. Hochberg. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57: 289 - 300.
- Blangero, J., J.T. Williams and L. Almasy. (2001). Variance component methods for detecting complex trait loci. in Genetic Dissection of Complex Traits, ed. D.C. Rao and M.A. Province, San Diego, CA: Academic Press, 151-181.
- Browning, B.L. and S.R. Browning. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *American Journal of Human Genetics* 84:210-223.
- Carlson, C.C., M.A. Eberle, et al. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* 74: 106-120.
- Cavalli-Sforza, L.L. and A.W.F. Edwards. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* 19: 233-257.
- Chu, T.-M., B. Weir, et al. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* 176: 35-51.
- Clayton, D. (2002). Choosing a Set of Haplotype Tagging SNPs from a Larger Set of Diallelic Loci. [<http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf>].
- Cockerham, C.C. (1969) Variance in gene frequencies. *Evolution* 23: 72-84.
- Cockerham, C.C. (1973) Analyses of gene frequencies. *Genetics* 74: 679-700.
- Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55: 997-1004.
- Dobbin, K. and R. Simon. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* 8(11): 1438-1445.
- Dudoit, S., Y. H. Yang, et al. (2002). Statistical methods for identifying genes with differential expression in replicate cDNA microarray experiments. *Statistica Sinica* 12: 111-140

- Edwards, L.J., K.E. Muller, and R.D. Wolfinger. (2008). An  $R^2$  statistic for fixed effects in the linear mixed model. *Statistics in Medicine* 27: 6137-6157.
- Efron, B., T. Hastie, *et al.* (2004). Least angle regression. *Annals of Statistics* 32(2): 407-499.
- Elston, R.C. and H.J. Cordell. (2001). Overview of model-free methods for linkage analysis. in Genetic Dissection of Complex Traits, ed. D.C. Rao and M.A. Province, San Diego, CA: Academic Press, 135-150.
- Excoffier, L. G. Laval, and S. Schneider (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1:47-50.
- Freidlin, B., G. Zheng, Z. Li, , and J. Gastwirth. (2002). Trend tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness. *Human Heredity* 53: 146-152.
- Gabriel, S.B., S.F. Schaffner, *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.
- Gauderman, W.J., C. Murcray, *et al.* (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genetic Epidemiology* 31: 383-395.
- Haseman, J.K. and R.C. Elston. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* 2: 3-19.
- Horvath, S., X. Xu, S.L. Lake, E.K. Silverman, S.T. Weiss and N.M. Laird. (2004). Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. *Genet. Epidemiol.* 26: 61-69.
- Hsieh, W. P., T.-M. Chu, *et al.* (2003). Who are those strangers in the Latin Square? in Methods of Microarray Data Analysis III. K. E. Johnson and S. M. Lin. Boston/New York/Dordrecht/London, Kluwer Academic Publishers: 247 pp.
- Huang, D.W., B.T. Sherman, and R.A. Lempicki. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.* 37:1-13.
- Idaghdour, Y., W Czika, *et al.* (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nature Genetics* 42: 62-67.
- Jin, W., R. M. Riley, *et al.* (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* 29: 389-395.
- Johnson, N.L and S. Kotz. (1970). Continuous Univariate Distributions-1. John Wiley & Sons: 300 pp.
- Johnson, W.E. *et al.* (2006). Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci.* 103: 12457-12462.
- Karolchik, D., R. Baertsch, *et al.* (2003) The UCSC Genome Browser Database. *Nucl. Acids Res.* 31: 51-54.
- Kent, W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.* 12: 996-1006.
- Kerr, M. K. and G. A. Churchill. (2001). Experimental design for gene expression microarrays. *Biostatistics* 2: 183-201.
- Kim, S.-Y. and D.J. Volsky(2005) PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics.* 6: 144.

- Kong, A., D.F. Gudbjartsson, *et al.* (2002). A high resolution recombination map of the human genome. *Nature Genetics* 31: 241 – 247.
- Kruglyak, L., M.J. Daly, M.P. Reeve-Daly and E.S. Lander. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* 58: 1347-1363.
- Lathrop, G.M., Lalouel, J.M. and White, R.L. (1986). Construction of human linkage maps: likelihood calculations for multilocus analysis. *Genet Epidemiol* 3: 39-52.
- Li, B., and S.M. Leal. (2008). Novel methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics* 83: 311-21.
- Li, Y. and G.R. Abecasis. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics* S79: 2290.
- Maddison W.P., D.L. Swofford, and D.R. Maddison. (1997). Nexus: An extendable file format for systematic information. *Syst. Biol.* 46:590-621
- Maniatis, N., A. Collins, *et al.* (2002). The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. (USA)* 99: 2228-2233.
- Marchini, J., B. Howie, S. Myers, G. McVean and P. Donnelly. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* 39 : 906-913
- Merchant, M. and S. R. Weinberger. (2000). Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* 21: 1164-1177.
- Monks, S.A. and N.L. Kaplan. (2000). Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *American Journal of Human Genetics* 66: 576-592.
- Mootha, V.K., C.M. Lindgren, *et al.* (2003). PGC-1 $\beta$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* 34: 267-273.
- Morris, A.P. and E. Zeggini. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* 34:188-193.
- Morton, N.E., W. Zhang, *et al.* (2001). The optimal measure of allelic association. *Proc. Natl. Acad. Sci. (USA)* 98: 5217–5221.
- Nei, m.. (1972). Genetic distance between populations. *American Naturalist* 106: 283-292.
- Nei, M., F. Tajima, and Y. Tatenno. (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *Journal of Molecular Evolution* 19:153-170.
- Oliehoek, P.A., J. J. Windig, J.A.M. van Arendonk, P. Bijma.(2006). Estimating Relatedness Between Individuals in General Populations with a Focus on Their Use in Conservation Programs. *Genetics* 173: 483-496.
- Price, A.L., N.J. Patterson, *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904-909.
- Purcell S, Neale B, Todd-Brown K, *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3):559-75
- Qu, Y., B.-L. Adam, *et al.* (2002). Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry* 48: 1835-1843.

- Redon, R., *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444: 444-454.
- Reynolds, J., B. S. Weir, and C. C. Cockerham. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767-779.
- Robinson, D., T.J. Pedersen, R.D. Wolfinger (2006) Chromosomal copy number analysis of Affymetrix® arrays with JMP® Genomics software. SAS Institute, Inc. Cary, NC.
- Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. Studies in Genetics VII. University of Texas Publications 7213, Austin.
- Rosenwald, A., G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, and L.M. Staudt. 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England J. Medicine* 346: 1937-1947.
- Slager, S.L., and D.J. Schaid. (2001). Case-control Studies of genetic markers: power and sample size approximations for Armitage's Test for Trend. *Hum. Hered.* 52: 149-153.
- Slifker, J.F. and S.S. Shapiro. (1980). The Johnson System: Selection and Parameter Estimation. *Technometrics* 22: 239-246.
- Smith, R., L.A. Owen, D.J. Trem, J.S. Wong, J.S. Whangbo, T.R. Golub, and S.L. Lessnick. (2006). Expression profiling of EWS/FLI identifies NKX2.2 as a critical target gene in Ewing's sarcoma. *Cancer Cell* 9:405-416.
- Stram, D.O., C.A. Haiman, *et al.* (2003). Choosing Haplotype-Tagging SNPs Based on Unphased Genotype Data Using a Preliminary Sample of Unrelated Subjects with an Example from the Multiethnic Cohort Study. *Human Heredity* 55: 27 - 36.
- Suarez, B.K. and S.E. Hodge. (1979), A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clinical Genetics* 15: 126-136.
- Subramanian, A., P. Tamayo, *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci (USA)* 102: 15545-15550.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58: 267-288.
- Tuzun, E., A.J. Sharp, *et al.* (2005) Fine-scale structural variation of the human genome. *Nature Genetics* 37: 727-732.
- Wall, J.D. and J.K. Pritchard (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *American Journal of Human Genetics* 73: 502-515.
- Wang, K. and D. Abbott (2008). A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology* 32: 108-118.
- Wang S., C.J. Basten, and Z.-B. Zeng (2007) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. (<http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>).
- Wang, T. and R.C. Elston. (2004). A modified revisited Haseman-Elston method to further improve power. *Human Heredity* 57: 109-116.
- Whittemore, A.S. and I-P. Tu. (1998). Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* 62: 1228-1242.
- Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76: 887-893.

- Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* 15: 323-354.
- Wu, M.C., P. Kraft, M.P. Epstein, *et al.* (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics* 86: 929-942.
- Wu, Z., R. LeBlanc, and R.A. Irizarry. (2004). Stochastic Models Based on Molecular Hybridization Theory for Short Oligonucleotide Microarrays Technical report. Johns Hopkins University. Dept. of Biostatistics Working Papers. ([www.bepress.com/jhubiostat/paper4/](http://www.bepress.com/jhubiostat/paper4/).)
- Varambally, S., J. Yu, *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* 8: 393-406.
- Venkatraman E.S. and A.B. Olshen. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23: 657-663.
- Yu, J., G. Pressoir, W.H. Briggs, *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38: 203-208.
- Zang, W., A. Collins, *et al.* (2002). Properties of linkage Disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. (USA)* 99: 17004 - 17007.
- Zapata, C., Alvarez, G., and C. Carollo (1997). Approximate variance of the standardized measure of gametic disequilibrium  $D'$ . *American Journal of Human Genetics* 61: 771-774.
- Zaykin, D.V., P.H. Westfall, *et al.* (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* 53: 79-91.
- Zhang, Z., E. Erzo, *et al.* (2010.) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42: 355-360.
- Zhao, K., M.J. Aranzana, *et al.* (2007). An *Arabidopsis* example of association mapping in structured samples. *PLOS Genetics* 3: 71-82.



# Index

## A

Accession Number [97](#)  
 Allele [97](#)  
 allele [18](#), [63](#), [93](#), [99](#), [101–103](#), [105](#)  
 Alpha [97](#)  
 alpha [72](#)  
 Annotation Data Set [97](#)  
 Annotation data set [91](#)  
 annotation data set [63](#), [69](#), [77](#)  
 Association [65](#), [77](#), [97](#)  
 association [3](#), [18](#), [65](#), [68](#), [72](#), [97](#), [101](#), [103](#), [109](#), [111](#)  
 AUC [97](#)

## B

Bin [64](#), [75](#), [97](#)  
 Binary Variable [97](#)  
 Bivariate [64](#), [97](#)  
 Bootstrap [97](#)  
 Box Plot [98](#)  
 Box plot [64](#)  
 BY Group [98](#)  
 BY Variable [98](#)  
 BY variable [98](#)

## C

Cell Plot [98](#), [101](#)  
 cell plot [72](#)  
 Censor Variable [98](#)  
 CentiMorgan [98](#)  
 Character Variable [98](#)  
 Cholesky Decomposition [98](#)  
 Class Variable [99](#), [105](#)  
 class variable [67](#), [71](#), [73](#)  
 Clustering [70](#), [90](#), [99](#)  
 cM [68](#), [98](#)  
 Color Variable [99](#)  
 Composite Interval Mapping [99](#)  
 Copy Number Variation [99](#)  
 Covariance [99](#)  
 Covariate [99](#)  
 covariate [33](#), [100](#)  
 Cross Validation [99](#)

**D**

Delimiter [99](#)  
 Dependent Variable [99](#)  
 Dialog [4](#), [7](#), [100](#)  
 dialog [5](#), [7-9](#)  
 Distance [100](#)  
 distance [18](#), [70-71](#), [98](#)  
 Drill down [100](#)

**E**

Eigenvalue [100](#)  
 Eigenvector [100](#)  
 ESTIMATE Statement [100](#)  
 estimate statement [67](#)  
 Exon [100](#)  
 exon [107](#)  
 Experimental Design Data Set [68](#), [100](#)  
 experimental design data set [66](#), [101](#)  
 Experimental Design File [9](#), [25](#), [31](#), [101](#)  
 experimental design file [3](#), [17](#)  
 Expression [20](#), [67](#), [91](#), [100](#)  
 expression [17](#), [19](#), [25-26](#), [103](#), [106](#), [108](#)

**F**

Factor [20](#), [68](#), [101](#)  
 factor [17](#), [20](#), [23-25](#), [28-30](#), [33](#), [35-36](#)  
 False Discovery Rate [101](#)  
 Field [101](#)  
 field [4](#), [9](#)  
 Fixed Effects [101](#)

**G**

Genotype [93](#), [101](#)  
 genotype [17-18](#), [68](#), [71-72](#), [77](#)  
 Group Variable [101](#)

**H**

Haplotype [68-69](#), [101](#)  
 haplotype [3](#)  
 Hardy-Weinberg Equilibrium [101](#)  
 Heat Map [101](#)  
 Hoeffding Correlation [102](#)  
 Hold-out Data [101](#)

**I**

Imputation [102](#)

Inbreeding Coefficient [102](#)  
 Independent Variable [102](#)  
 independent variable [99](#)  
 Index Variable [102](#)  
 Interval Mapping [102](#)

## J

Jitter [102](#)

## K

K\_Rho [102](#)  
 Kendall Correlation [102](#)  
 Kinship Coefficient [102](#)  
 K-Means [70](#), [83](#), [90](#), [102](#)

## L

Label Variable [102](#)  
 Leaf [102](#)  
 Level [103](#)  
 level [20–22](#), [25](#), [28–30](#), [33](#)  
 Linkage [103](#)  
 linkage [3](#), [78](#)  
 Linkage Disequilibrium [103](#)  
 LOD score [65](#)  
 Lod Score [103](#)  
 Loess Normalization [103](#)  
 Loss of Heterozygosity [103](#)  
 LSMeans [103](#)

## M

Main Effect [103](#)  
 main effect [98](#)  
 Marker Variable [103](#)  
 Mean [103](#)  
 mean [22](#), [68](#), [71](#), [98](#), [107–108](#)  
 Median [104](#)  
 median [71](#)  
 Missing Value [104](#)  
 missing value [91](#)  
 Mixed Model [104](#)  
 Model [104](#)  
 model [4](#), [23](#)  
 Monophyletic [104](#)

## N

Nominal Variable [104](#)

nominal variable [106](#)  
Numeric Variable [104](#)

## O

Observation [104](#)  
Ordinal Variable [104](#)  
Overfit [104](#)

## P

Partial Least Squares Coefficient [104](#)  
PCTL [104](#)  
Pearson Correlation [104](#)  
Pedigree [104](#)  
Phenotype [105](#)  
phenotype [17–18](#)  
Population Stratification [105](#)  
population stratification [73](#)  
Power [105](#)  
power [36, 72](#)  
Predictive Model [4, 105](#)  
predictive model [101, 107](#)  
Predictor [105](#)  
predictor [101](#)  
Principal Components [105](#)  
principal components [72](#)  
Probe Intensity [105](#)  
Probe Set [105](#)  
probe set [64](#)  
PROC [105](#)  
p-Value [105](#)  
p-value [108](#)

## Q

Quantile [106](#)  
Quantitative Trait Locus [106](#)

## R

random effects [36](#)  
Random Number Seed [106](#)  
Residuals [106](#)  
residuals [64](#)  
Rho [106](#)  
Root Mean Square Error [106](#)

## S

SAS Data Set [106](#)

SAS data set [7](#), [16](#), [100–101](#)  
SAS Log [106](#)  
SAS log [13](#), [89](#)  
Single Nucleotide Polymorphism [106](#)  
Smoothing Bandwidth [106](#)  
SNP [106](#)  
Spearman Correlation [106](#)  
Standard Deviation [107](#)  
Standard Error [107](#)  
standard error [36](#), [107](#)

## T

Test Data Set [107](#)  
Training Data Set [107](#)  
Transcript Cluster [107](#)  
Transformation [107](#)  
Transmission Disequilibrium Test [107](#)  
Tree [107](#)  
tree [102](#)  
t-statistic [107](#)  
t-test [107](#)

## V

Variable [107](#)  
variable [20](#), [36](#)  
Variance [108](#)  
variance [77](#)  
Venn Diagram [108](#)  
Volcano Plot [108](#)

## W-Z

Wizard [108](#)

