JMP GENOMICS: AN OVERVIEW

Sunil Archak
NBPGR, New Delhi-110 012
sarchak@nbpgr.ernet.in

1. Introduction

The challenge and responsibility of agricultural scientists is to provide food, fiber and fuel for the ever-increasing populace. Deciphering genome sequences of model plants and crop plants will help agricultural biotechnologists to understand the plant genome architecture and subsequently to identify and employ economically important genes.

The complete genome sequence of model plant Arabidopsis thaliana ignited plant genomics research. Today, hundreds of plant genomes are being sequenced (fully or partially) resulting in an ocean of sequence data. It is a serious challenge for biologists, statisticians and computer scientists alike to interpret the biological meaning of a combination of only four letters. Further, ESTs (Expressed Sequence Tags) provide a snapshot of whole genome with a small percentage of cost of whole genome sequencing. The next generation sequencing output is also a great resource for plant genomics. These technologies will generate enormous amount of sequencing data, which will be beneficial to understand the basic physiology of plants and functions of thousands of genes. Research involving signal transduction (abiotic stress response in plants); non coding RNAs (miRNAs and siRNAs); molecular markers, and tools of transcriptomics, proteomics and metabolomics; and systems biology approach (to study protein-protein interactome in plants) make use of microarray technology apart from sequencing. Hence, the types and the amount of data will always be on the exponential path.

It is here that bioinformatics and computational biology will be a great help in predicting functions of plant genes. It is also realized how in silico methods along with experimental approaches like two hybrid systems, affinity purification and mass spectrometry, biomolecular fluorescence complementation etc. could be used to understand the proteome of plants. As a result, computational tools, methods of statistical analysis and the software packages become very important and felicity in their usage turns out to be extremely significant.

One of such myriad softwares is JMP Genomics. JMP Genomics is a statistical discovery software tool from SAS and JMP. Major objective in using JMP Genomics is to uncover meaningful patterns in high-throughput genetics, expression, copy number and proteomics data. For a biologist, dynamically interactive graphics make it easy to explore data relationships using a comprehensive set of traditional and advanced statistical algorithms. For a statistician, connecting the data, script and results in a single loop for reiterative operations and extrapolations.

JMP Genomics software brings high-powered, sophisticated genomic data exploration and analysis to the desktop. However, the operations are done on SAS for heavy-duty processing of genomics data sets.By dynamically linking advanced statistics with graphics to provide a complete and comprehensive picture of research results, JMP Genomics helps biologists, biostatisticians and statistical geneticists understand data generated from genetics, expression, exon, copy number and proteomics studies. Users enjoy a menu-driven system that simplifies the workflow throughout the research process and interactive data visualization capabilities that let them see and explore their data from every angle, then easily share findings with colleagues and publish in reputed journals.

At present, we use JMP Genomics 4.1, based on functionality from the SAS 9.2 and JMP 8 platforms. In the course of this training program, we shall not be able to master the JMP Genomics. The objective of the session is to:

1. Know about JMP Genomics
2. Know available menus
3. Understand the available options
4. Run selected options
5. Find help

Definitions of terminologies often mentioned during the discussion

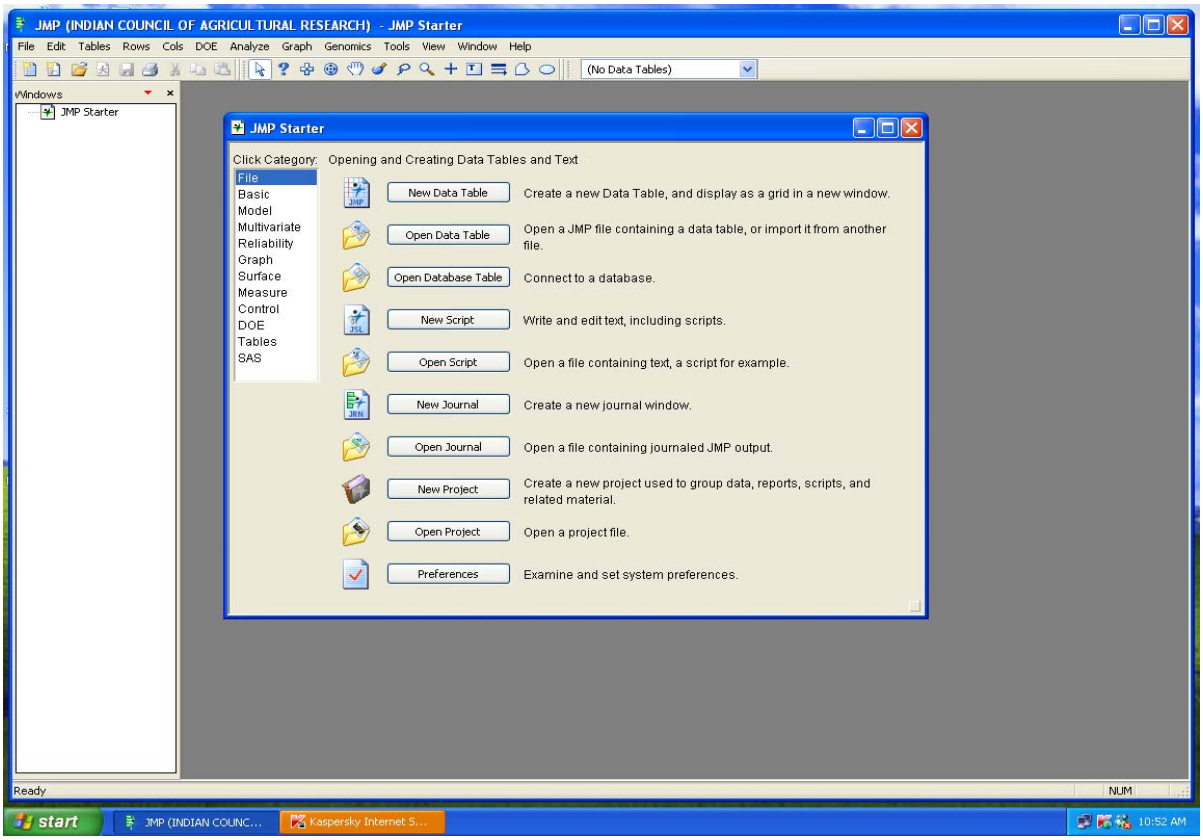
Genetics	The branch of biology that studies heredity and variation in organisms
Genetic Analysis	The methods as well as results of systematic study of correlation between traits and genes, the implications of such linkage and their applications
Gene	DNA sequence that conveys the information required to make a molecule, usually a protein or RNA. Gene possesses the complete information to define the composition of the product as well as the signals that control when and where the RNA/protein should be available
Population Genetic Analysis	The process of making inferences about the evolutionary and demographic history of a gene (or organism) on the basis of data on genetic variation in a species
Genome	The entirety of an organism's hereditary information in the form of total nuclear DNA sequence typically expressed in number of base-pairs; contained in the haploid set of chromosomes
Genotype	Genetic constitution of an individual/lineage/organism particularly expressed as the specific allele makeup with reference to a specific trait under consideration; also a group of organisms sharing a specific genetic constitution
Haplotype	A contraction of the term haploid genotype; haplotype is a combination of alleles at multiple loci/genes/markers/polymorphisms that are transmitted together
Bioinformatics	The discipline dealing with the application of statistics and computer science to the field of molecular biology entailing the creation and advancement of databases, algorithms, computational and statistical techniques and theoretical principles to solve structural and practical problems arising from the management and analysis of biological data
Computational Biology	The actual process of analyzing and interpreting data is referred to as computational biology
Biotechnology	A synthetic discipline amalgamating the pure biological sciences (genetics, microbiology, animal cell culture, molecular biology, biochemistry, embryology, cell biology) as well as chemical engineering, bioprocess engineering, information technology etc. to develop bio-products for the welfare of mankind
Genetic Engineering	The direct human manipulation of an organism's genetic material to influence gene expression in a way that does not occur under natural conditions

Genomics/ Transcriptomics/ Proteomics	The branch of genetics that studies organisms in terms of their whole genome (full DNA sequences)/total RNA species/total proteins encoded
Marker	A gene or a segment of DNA, having polymorphic genetic property, with or without an identifiable location on a chromosome, and whose inheritance can be followed
Association genetics	A research field dedicated to the identification of correlations between phenotypic traits and genetic markers with the aim to identify and locate the underlying genes
Mapping	Determination of the order and relative positions of genes/markers on the chromosome; distances are measured in linkage units (centimorgans, cM)

A glossary is available at http://www.nbpgr.ernet.in/repository/glossary.htm.

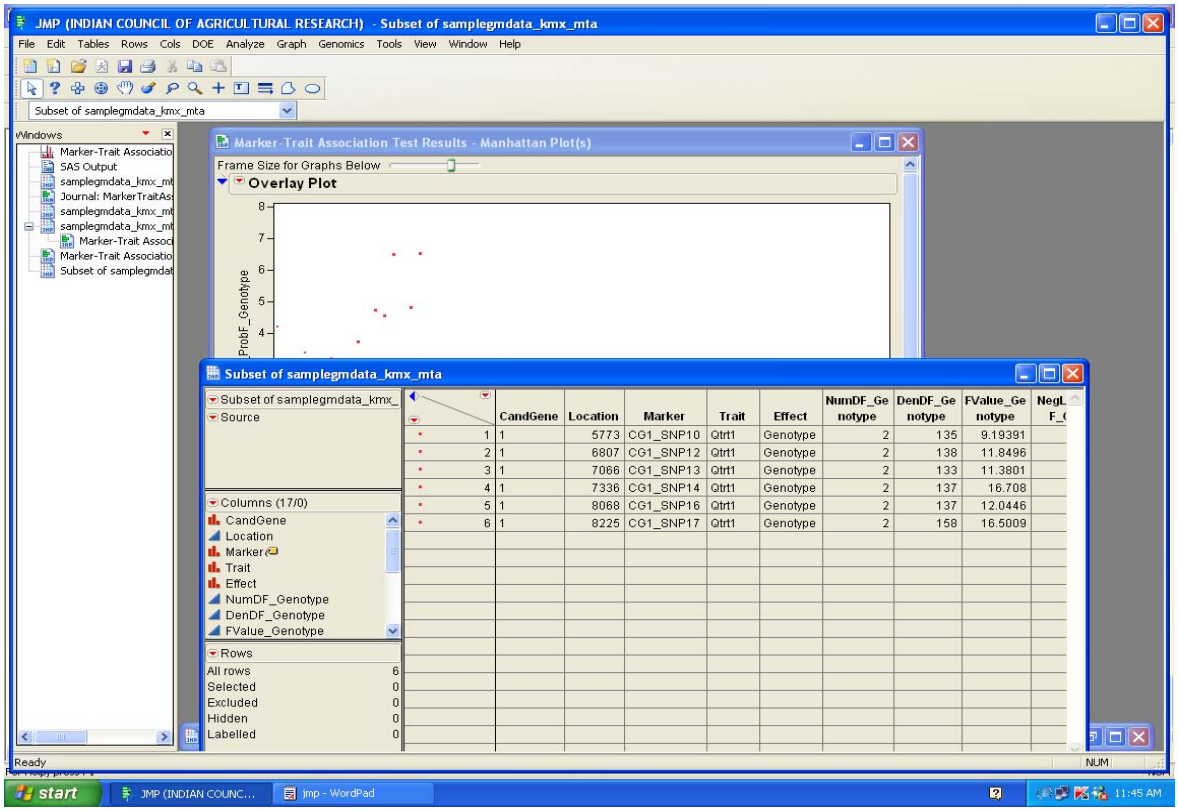
2. Let's JMP !!

1. Before beginning to use JMP Genomics, "**Set genomic preferences**"
2. Every time we start JMP Genomics, we see a "**Tip of the Day**" dialog box, giving useful and unique feature. Tips can also be found in the HELP menu.
3. Every time we start JMP Genomics, we also see the "**JMP starter**" dialog box, giving opportunity to start an operation by selecting one of the 10 options. During the training however, we shall not be using the "**JMP starter**" but directly use drop-down menus.



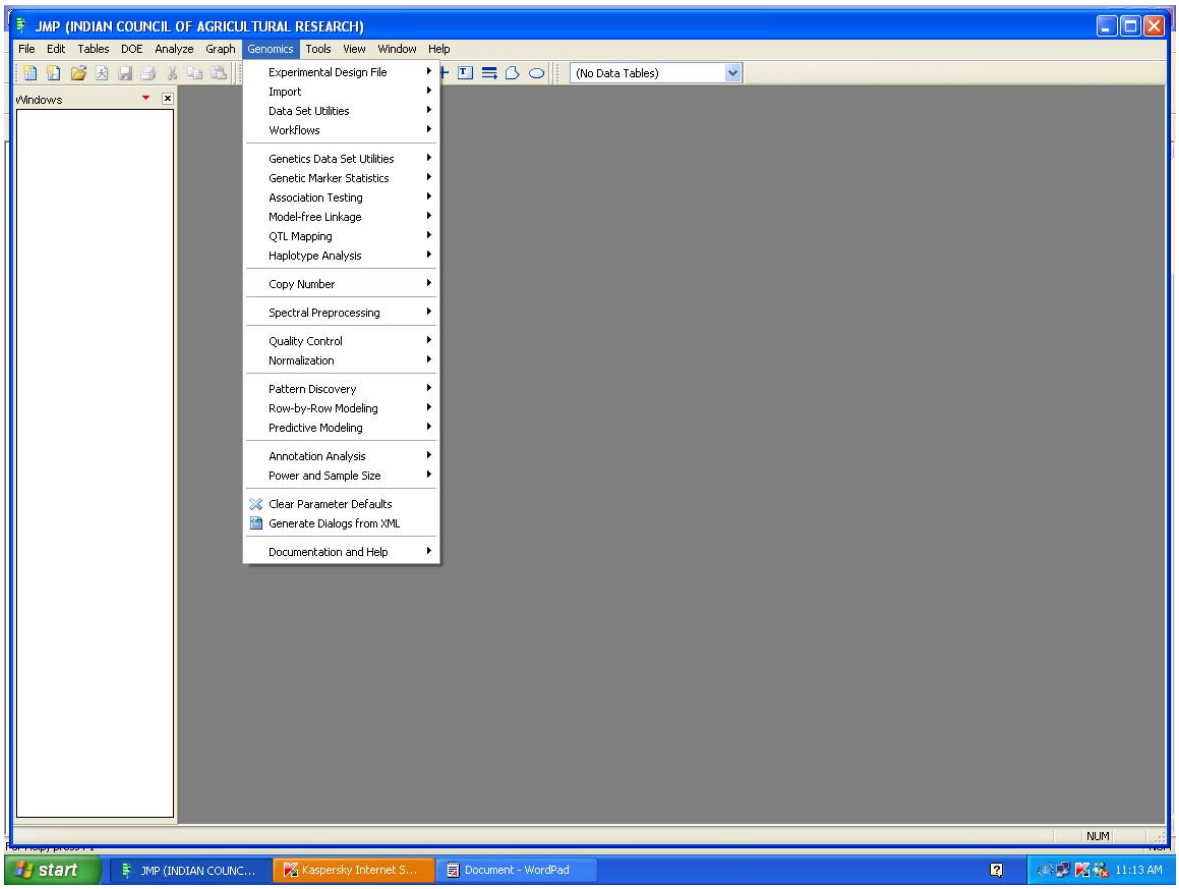
There are 11 drop down menus (13 on opening) that include **File**, **Edit**, **Tables**, **Design of Experiment (DoE)**, **Analyze**, **Graph**, **Genomics**, **Tools**, **View**, **Window** and **Help**. Following section gives a glance at these items.

1. **File** menu items of interest are **Open** (to open an item that is already worked upon), Import (to get a non-JMP data file to begin working), **Print setup** (to arrange to print a table, window, graphics, script etc.), Save as **SAS data set** (especially for the imported items and large data sets), **Save session script** (to save the complete operations as JMP script in case operations are to repeated with ease; however this takes greater memory to finish operation).
2. **Edit** menu items of interest are **Run/Stop script** (one of the many ways to run a function, others include clicking on **Run** button in any dialog window and using **Ctrl+R**); **Submit to SAS** (to submit specific statistical operations to run on the background server; menu driven operations, however, take care of this automatically).
3. **Table** menu provides excellent options for editing an opened table. Item of interest is the **Subset** (to tabulate selected individual observations of significance from



4. **Design of Experiment (DoE)** deals with setting up an input data file before even running an experiment. Optimum use of the software lies in minimizing the error probability which is possible with employing available robust statistical tools. This option needs cross-talk between biologist and statistician. Not dealt with during the training.
5. **Analyze** provides an option to exercise independent bits of statistical tests on selected data sets and sub-sets to generate powerful graphical results that are amenable to further analysis. **Distribution** is the oft used option.
6. **Graph** provides an option to develop independent and alternative bits of visualizations (without actually running the processes of JMP Genomics) of selected data sets and sub-sets that are amenable to further analysis. **Overlay plot** to generate scatter plots along chromosomes and **Cell plot** to generate heat maps are often used.

7. **Genomics** is essentially an **add-on** to JMP software. It has **eight part** drop down menu (explained below from **a** to **i**).



(a) Input facility utilities contains **Experimental design file** (to obtain the pre-made design, say from data source like Affymetrix, or to create one from a flat file); **Import** (an option to import specialized data files of known standards OR common csv/txt/xls files); **Data set utilities** (to preview data input files that may contain more than a million columns hence should not be opened in JMP); **Workflows** (an extremely important and useful option to explore for beginners; allows the implementation of an entire repertoire of suitable analytical procedures available in JMP, under a given type of analysis viz. Genetic analysis or expression analysis etc.).

(b) Genetic analysis and copy number part primarily contains **Genetics dataset utilities** (has six options, to associate correctly multiple input files viz. data file and annotation file); **Genetic marker statistics** (has seven processes, that facilitate summarize and display information and statistics on alleles, genotypes, and phenotypes and their linkage patterns); **Association testing** (provides nine processes to map binary or quantitative traits by association with genetic marker data in both stratified and non-stratified populations, Principal components analysis is one important utility); **Model-Free Linkage** (has three processes to perform various affected sib-pair chi-square tests of linkage for a dichotomous trait, tests for linkage between genetic markers and a quantitative or binary trait locus, and use pedigree and IBD data to test for linkage to locate quantitative trait loci); **QTL Mapping** (has three processes to quickly scan the whole genome for evidence of QTL signals, build a genotype probability data set that includes probabilities of each possible QTL genotype in each location for all individuals, and can be used to generate a single Interval Mapping model to locate quantitative traits across the genome; **Haplotype**

Analysis (has three processes, Haplotype Estimation, Haplotype Trend Regression, and htSNP Selection); **Copy Number** submenu gathers together those quality control, data manipulation and modeling processes that are particularly useful for copy number analysis.

(c) Spectral Preprocessing submenu provides a set of **proteomics** processes useful for analyzing two dimensional and three-dimensional spectra.

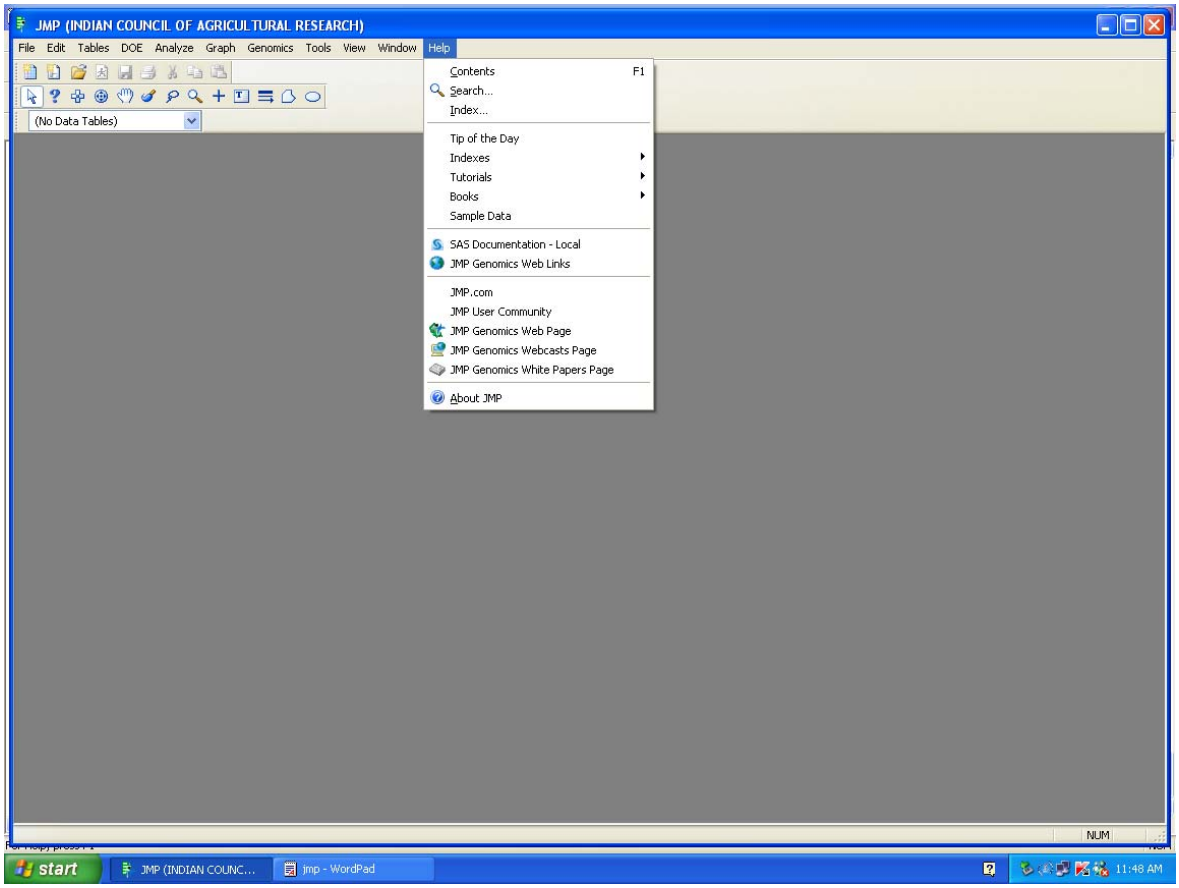
(d) Expression analysis part provides processes including **Quality Control** provides ten processes that assess the quality of raw data; **Normalization** provides nine processes that normalize data; **Pattern Discovery** provides eight processes that investigate the nature, magnitude and causality of the relationships between the observations in data; **Row-by-Row Modeling** provides eleven processes that fit statistical models to the rows of a tall data set and compare aggregated results.

(e) Statistical analysis part deals with **Power and sample size** and **Predictive modeling** using data from genetic markers, microarrays, or proteomics as predictor variables based on processes viz. Discriminant Analysis, Distance Scoring, General Linear Model Selection, K Nearest Neighbors, Logistic Regression, Partial Least Squares, Partition Trees, Radial Basis Machine, and Survival Predictive Modeling.

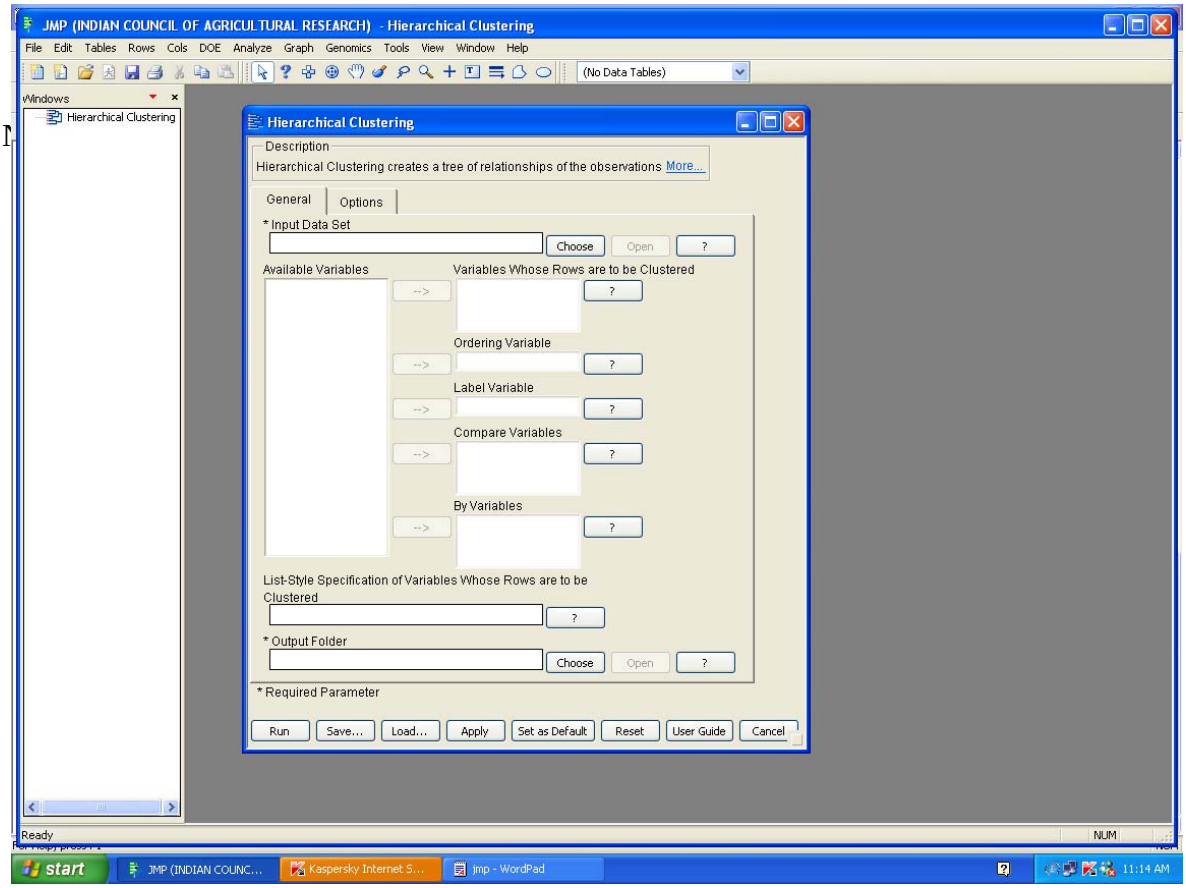
(f) **Annotation Analysis** submenu provides a set of bioinformatic tools that can help incorporate, in addition to statistical results, biological meaning (Probe or Probe Set ID, GenBank Accession Number, UniGene Cluster ID, Gene ID, Description, Chromosomal location, Ensembl ID, Swiss-Prot ID, EC number, OMIM ID, dbSNP ID, RefSeq Accession, Gene Ontology ID, Genomic location/coordinate)

(g) Additional tools include **Clear parameter defaults** and **Generate dialogs from XML**

(h) Help is available under **Documentation and Help** submenu (in addition to Help menu)

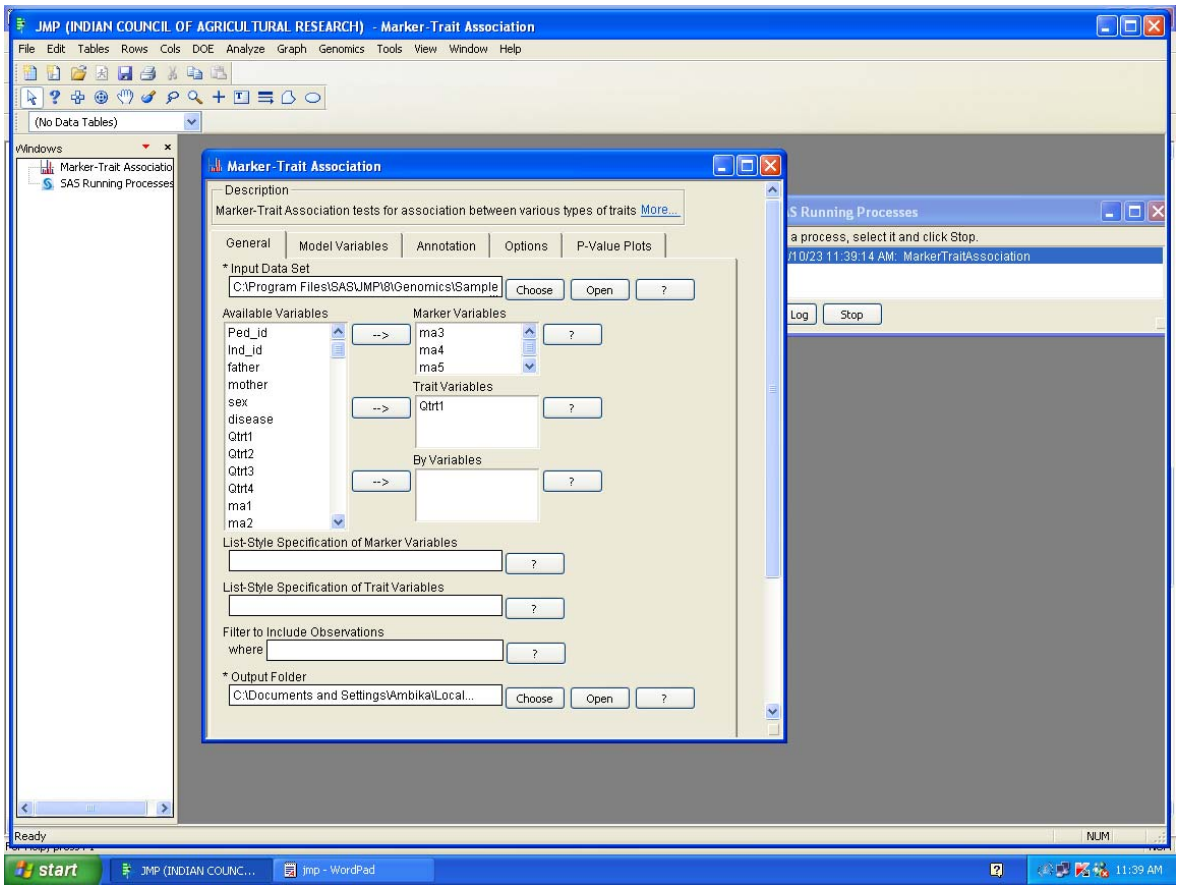


8. **Tools** provide direct editing tools akin to any image processing interface.
9. **View** allows selection of windows and sidebar for optional display.
10. **Window** menu is to select a specific result after the analyses is completed and toggle among multiple windows; "**close all**" closes all the windows to facilitate fresh beginning.
11. **Help** is organized into **Contents** (contents, search and index); **Tip of the day**; **Indexes** (statistics, JSL functions, object scripting, display box); **Tutorials** (nine dialog box based sets); **Books** (seven JMP guides, one itemized JMP Genomics guide in portable document format); **Sample data** (to learn/teach basic JMP tools); **Links** (external webpages). Screen-shots of JMP Genomics to familiarize with the package



Example of an input dialog box

Every input dialog box has at the top a **description** of the function. One to many tabs are provided to input different information, for instance above box has two tabs, **General** and **Options**. Clicking on **Choose** button opens a file selection window. In case of requirement of multiple input files, each tab may require to select files that are corresponding and properly formatted. Alternatively, one can input the whole set of related files by using **Load** button and selecting from the available set of input files. **Open** button allows a look at the data table that may help decide on variable classes. **Question mark** muttons provide instant and brief help. Successful run depends upon the correct format of the file, but correct answer depends upon proper choice of variables from the available variables. It is equally important to choose a unique output folder to save all the output files viz. output tables, graphs, scripts, journal etc. **Reset** button is for fresh beginning whereas **Cancel** button is for abandoning the attempt. One can access relevant manual content by clicking on **User Guide**. It is advised not to fiddle with **Set as Default** button until attaining expertise. Clicking on Run button will execute the selected analysis.



Example of running a process

Input file chosen, marker variables and trait variables selected, output folder chosen. Note that there are **FIVE** input tabs. **Annotation** is a common tab to provide information about the markers; **Options** for additional statistical strengths and **P-value plots** for significance tests. The analysis is being executed. Note a side bar listing all the open windows. Invoking an analysis starts a relevant SAS process in the background with a provision to view the **Log** or to **Stop** the process.

Notes:

……

……

……

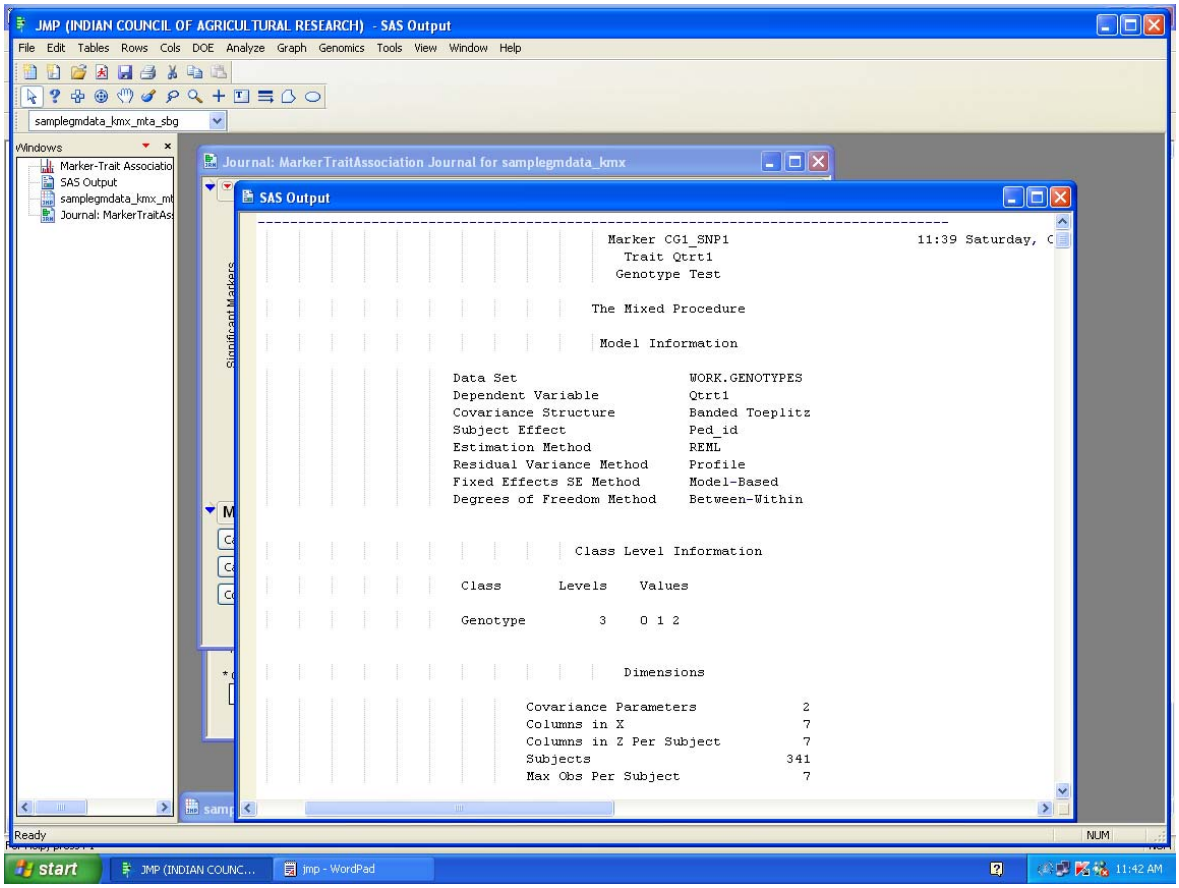
……

……

……

……

……



Example of a primary output box

A typical SAS output window opens up to indicate the completion of the process, with details of the completed process. Note a side bar listing all the open windows. Typically they include the input window, data table window, SAS output window and a Journal window.

Notes:

……..

……..

……..

……..

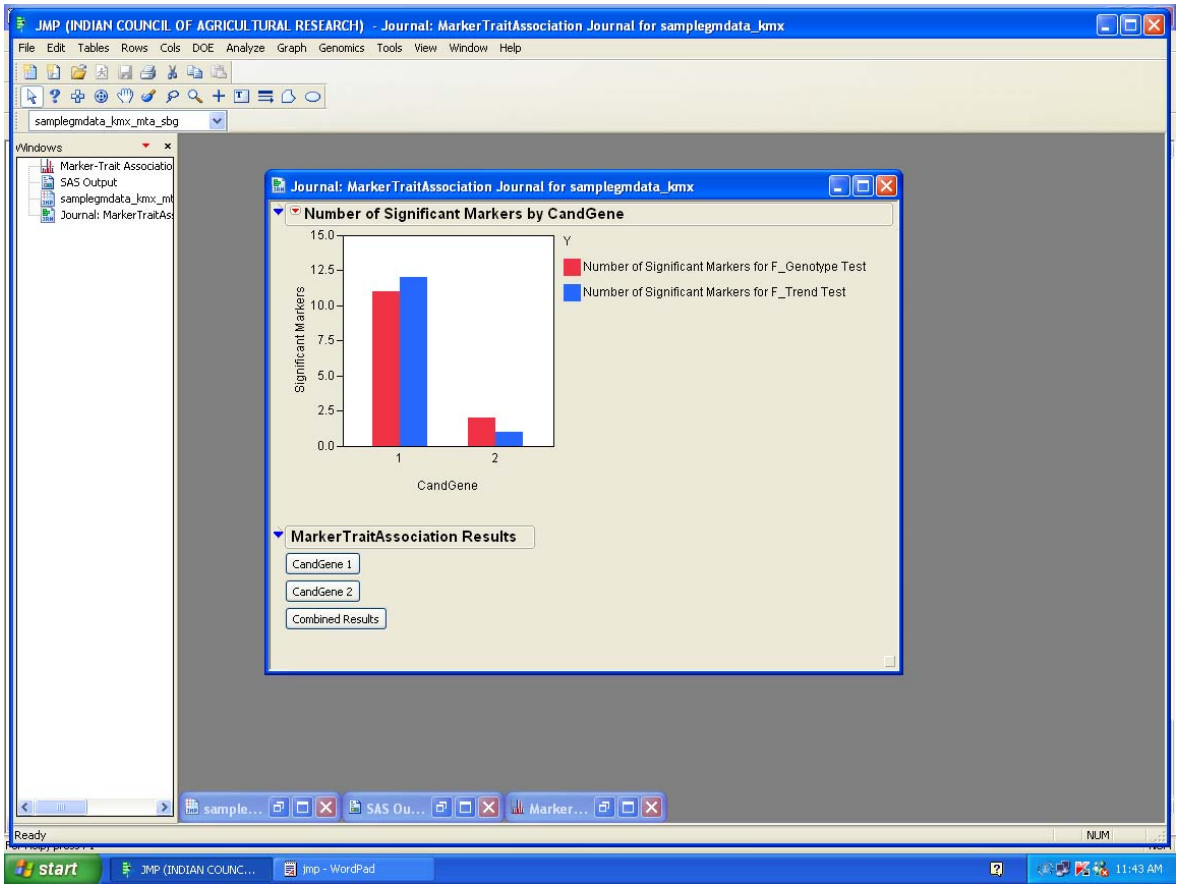
……..

……..

……..

……..

……..



Example of an output journal

JMP output is in the form of a Journal. Multiple outputs are presented in a collapsed format with blue diamonds to expand each one of them. In the above example, both the results are expanded with the second one having further options of display. Note that out of the four result windows, three are minimized that are also listed in the sidebar and anyone can be selected for view.

Notes:

……..

……..

……..

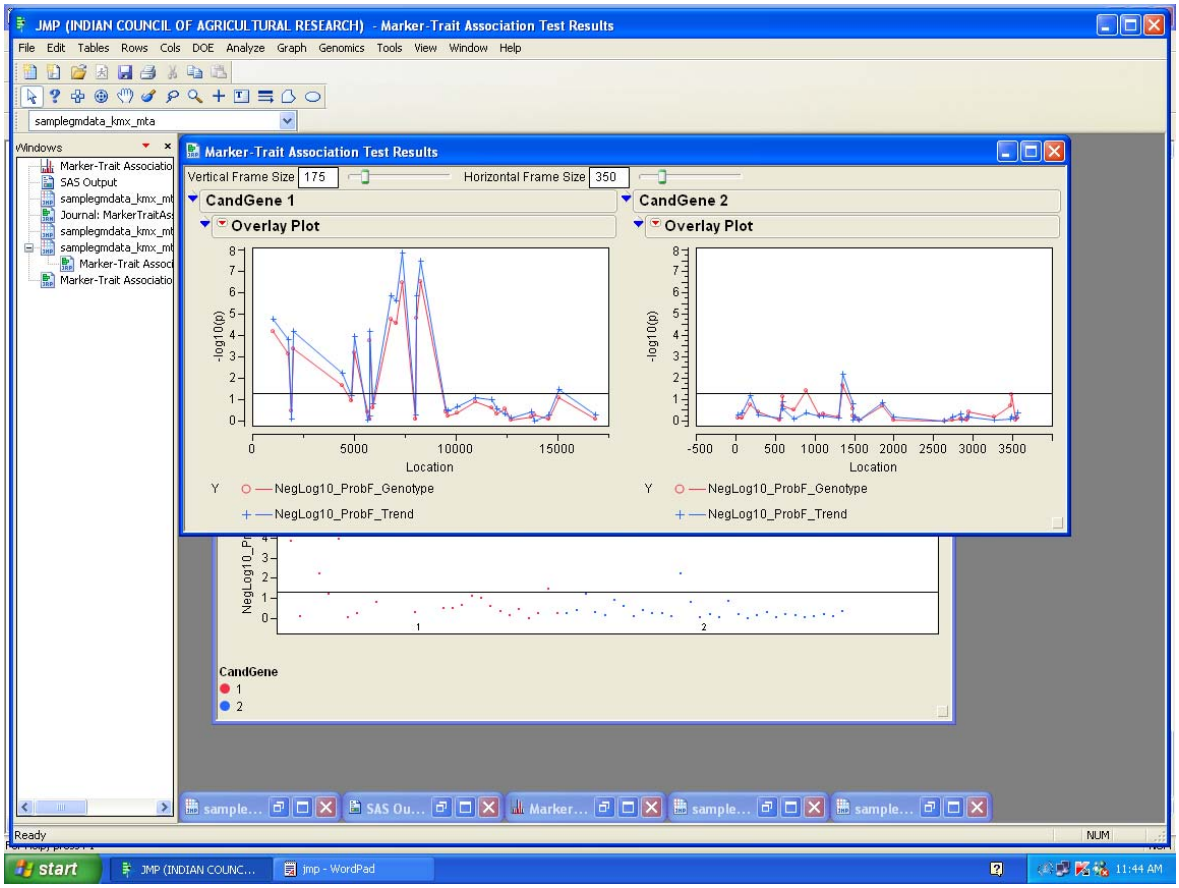
……..

……..

……..

……..

……..



Example of graphical output

Graphical output is equipped with editing tools (size, color, background, statistical fortifications etc.) as well as facility to expand analytical repertoire (by drop down menu using red diamond). The graphics can be saved for documentation and publication. Note that each additional analytical window gets listed in the sidebar for reference

Notes:

……..

……..

……..

……..

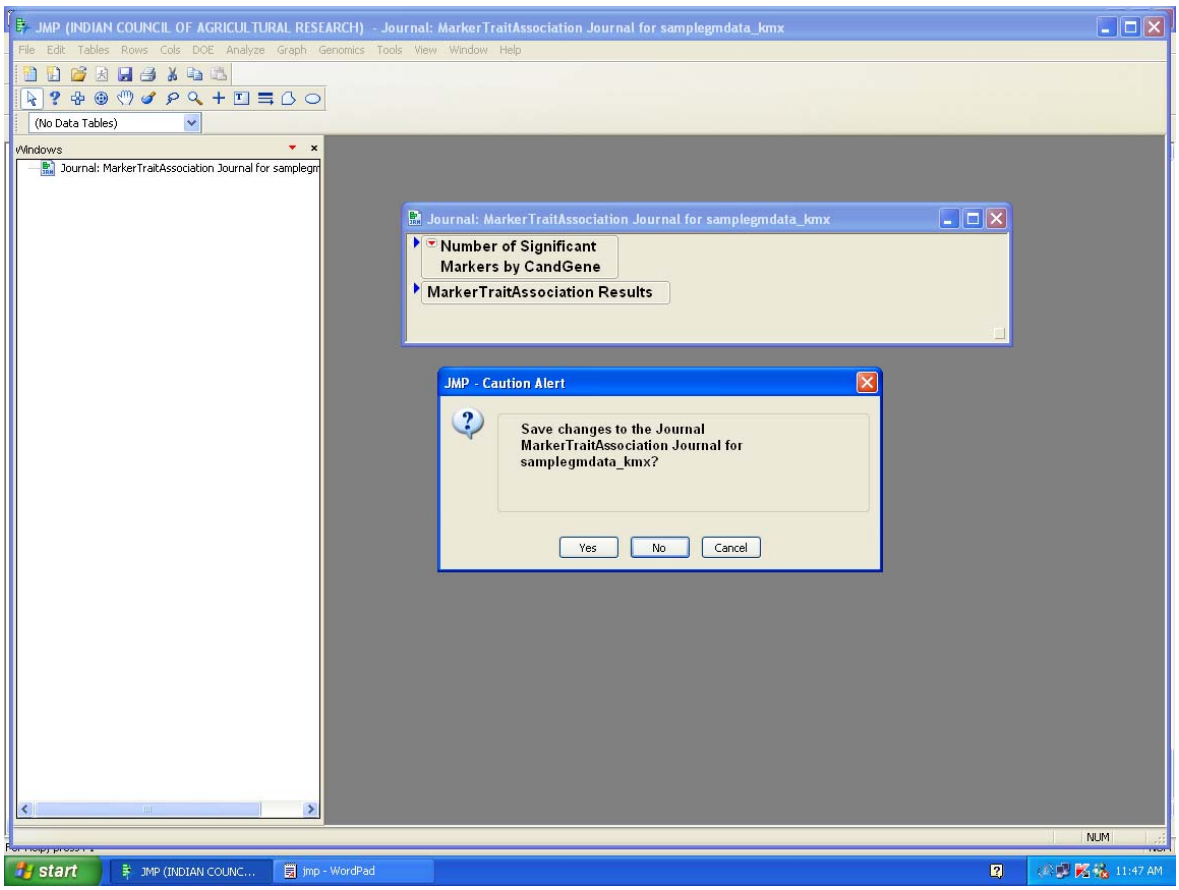
……..

……..

……..

……..

……..



Example of a caution alert dialog box

Attempt to end the analysis is responded to with a query to save the process result. If saved, the journal gets located in the output folder. Note that each and every result including graphical output gets saved in the journal along with internal links for later calling.

Notes:

……..

……..

……..

……..

……..

……..

……..

……..

……..

……..

……..

3. Hands on experience

Best way to begin is to run the basic workflows using the data sets supplied with the software. Following workflow processes shall be done by participants:

Workflow	Objective
Basic genetic workflow	To understand all the genetic analytical procedures available by default
Basic expression workflow	To understand all the expression analytical procedures available by default

If time permits, then following individual analytical processes shall be demonstrated/done hands on:

Individual analysis	Objective
QTL analysis	To learn how the software allows a quick scan of the whole genome for evidence of QTL signals, build a genotype probability data set that includes probabilities of each possible QTL genotype in each location for all individuals
Chromosome Color Theme	The process creates a settings (.sas) file from a text file that defines a Chromosome Color Theme that can be used for display of the Chromosome Color Plot

Notes:

……..

……..

……..

……..

……..

……..

……..

……..

……..

……..