Paper 213-29

Progress in SAS® Scientific Discovery Solutions: Genetics, Microarrays, and Proteomics

Russ Wolfinger, SAS Institute Inc., Cary, NC
Wendy Czika, SAS Institute Inc., Cary, NC
Kristen Kleiss, SAS Institute Inc., Cary, NC

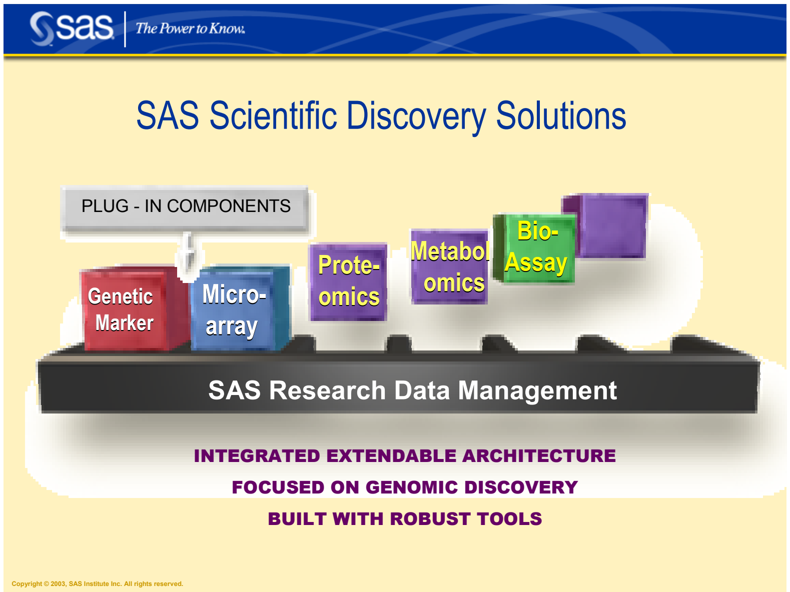
ABSTRACT

The SAS Genomics team is making steady progress in leveraging existing and new SAS technologies for more effective approaches to scientific discovery. This paper presents an overview of SAS/Genetics, SAS Research Data Management (RDM), and the SAS Microarray Solution, as well as introduces the two latest additions to the SAS Scientific Discovery Solutions line, the SAS Genetic Marker Solution, and the SAS Proteomics Solution. Release 1.3 of the RDM platform contains several major improvements, including an expanded architecture for supporting input engines and analytical processes, as well as advanced GUI support via XML. The SAS Genetic Marker Solution is built on the RDM platform and features analytical processes invoking SAS/Genetics procedures in the background. The SAS Proteomics Solution is also an add-on to the platform and focuses on the management and analysis of spectral data, enabling advanced data analyses and profiling of peptides or metabolites. This paper highlights the basic features of the new solutions and how they enable robust identification of discriminating characteristics that can potentially be linked to life-changing diagnostic and/or therapeutic biomarkers.

INTRODUCTION

The volume and complexity of data arising from the genomics revolution are driving increasing demand for statisticians and geneticists who can work with scientists and IT professionals to unravel the mysteries of molecular life. Appropriate software tools are indispensable in this endeavor, and SAS Scientific Discovery Solutions (SDS) provide a common framework for processing genomics data, thereby fostering efficient collaborations between these different groups of expertise.

SDS are all built upon SAS Research Data Management (RDM), a data repository and analysis management platform (Figure 1). RDM ties together several classical SAS technologies via a Java interface and enables non-statistical scientists to easily execute programs written by statisticians and SAS programmers, thereby fostering collaboration and increasing efficiency. Open, editable SAS code drives both input engines (IEs) for loading genomics data and analytical processes (APs) for analyzing the data. The latter typically create output in HTML or scripts for dynamic visualization and results exploration in JMP.


SAS Scientific Discovery Solutions

INTEGRATED EXTENDABLE ARCHITECTURE
FOCUSED ON GENOMIC DISCOVERY
BUILT WITH ROBUST TOOLS

Figure 1. SAS SDS Framework

This paper discusses three solutions built upon RDM. The SAS Microarray Solution is the most mature of the three, and has a rich collection of exploration, normalization, and statistical methods. The SAS Genetic Marker Solution targets genetic polymorphism data, and the APs invoke SAS/Genetics procedures in the background. The SAS Proteomics Solution focuses on the management and analysis of spectral data and enables profiling of peptides or metabolites. The RDM framework is also extensible to other domains and facilitates integration between them.

Before presenting new features of the Genetic Marker and Proteomics Solutions, the following three sections describe enhancements to SAS/Genetics, RDM, and the SAS Microarray Solution.

ENHANCEMENTS TO SAS/GENETICS

SAS/Genetics, released along with SAS 9.1, includes the following new features:

- PROC options for reading in different formats of marker data, such as columns containing genotypes
- PROC ALLELE options that can either subset ODS tables for allele, genotype, or haplotype frequencies or enable genotypes and haplotypes that are not observed to be displayed in their respective table
- NULLSNPS= option in PROC CASECONTROL to specify a set of SNPs other than those being analyzed for calculating the variance inflation factor for genomic control (Bacanu, Devlin, and Roeder 2000)
- Monte Carlo estimates of exact *p*-values for case-control and family-based trait-association tests
- Checks for genotype inconsistencies between parents and offspring in PROC FAMILY
- Stepwise EM algorithm (Clayton 2002) for estimating haplotype frequencies in PROC HAPLOTYPE
- ID statement in PROC HAPLOTYPE to include input data set variables in the output data set
- HTSNP procedure for identifying ‘haplotype tag’ SNPs (Johnson et al. 2001)
- INBREED procedure for calculating inbreeding or covariance coefficients within pedigrees
- Truncated product method (Zaykin et al. 2002) for adjusting *p*-values in PROC PSMOOTH

Additionally, an experimental version of the new PROC BTL is available to run on SAS Release 8.2. This procedure fits a multiple binary trait loci model to data from experimental populations by implementing mixed model analysis of variance and maximum likelihood methods (McIntyre, Coffman, and Doerge 2001; Simonsen 2004). To obtain this procedure, send an email to Christine.Duarte@sas.com.

ENHANCEMENTS TO THE RDM PLATFORM

Release 1.3 of RDM includes several major new enhancements. First, the core functionality of RDM has been expanded to provide full capabilities for IEs and APs. Recall RDM is required for all other SDS solutions and forms their foundational structure as well as a hierarchical data repository built using add-ins to SAS/Warehouse Administrator. The complete architecture for all of SDS is now included in RDM, and so the solutions simply consist of additional plug-and-play IEs and APs customized for the domain of interest (Figure 1). This architecture is three-tiered as depicted in Figure 2. The SAS Server components are from SAS Version 8.2.

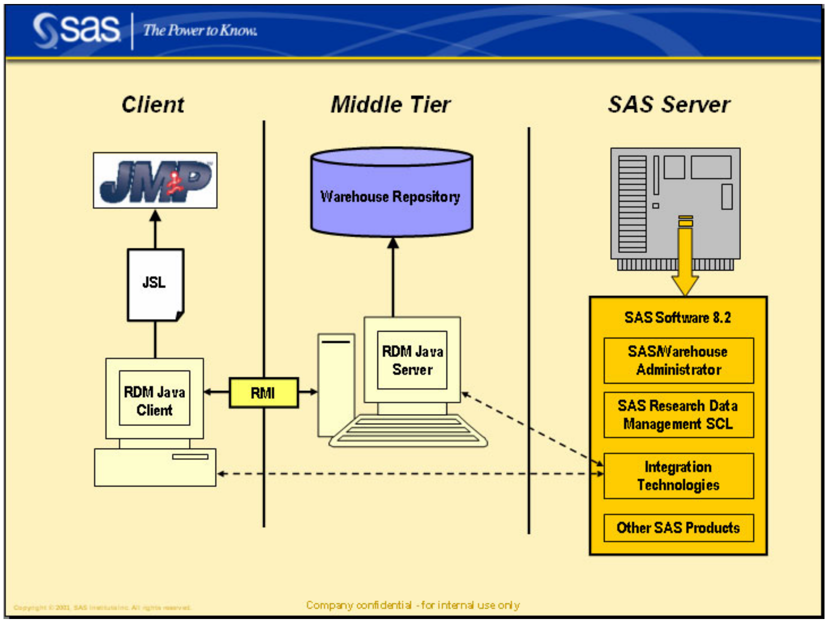
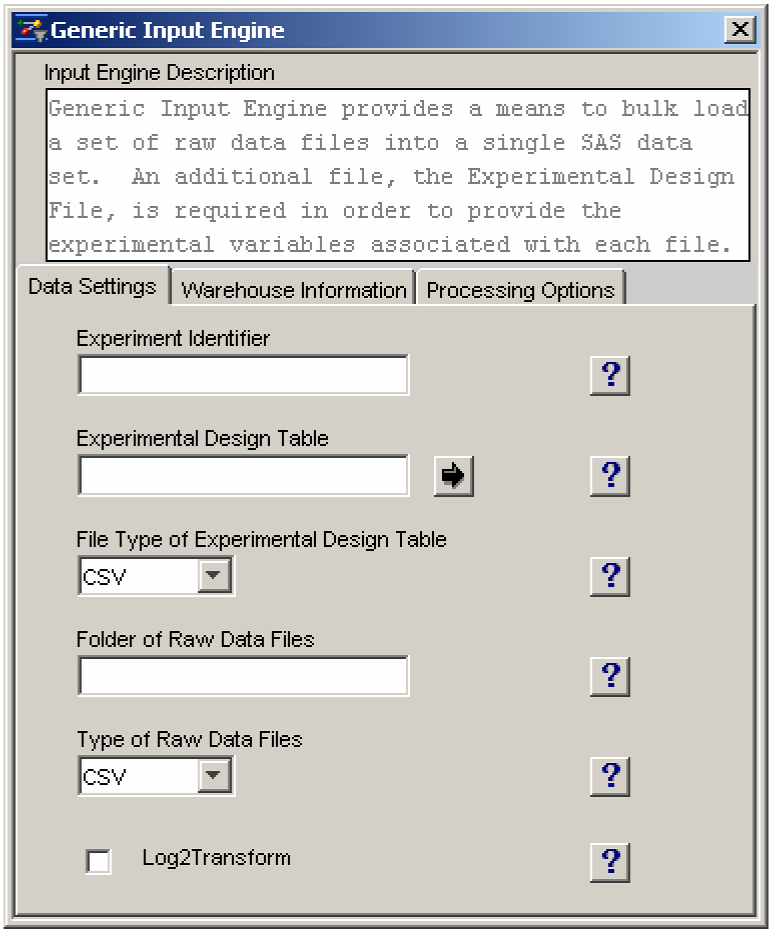


Figure 2. RDM Architecture

Second, the underlying processing of IEs and APs has been merged into a common framework, lending additional functionality to both. For example, you can now save settings for IEs for later use or modification. The Process Manager and Process Selector have expanded functionality, including a new Settings Manager for copying, editing, deleting, and setting security on your saved settings.

Third, RDM now comes with its own set of IEs and APs that are common to all domains. Figure 3 shows how the input screen for the Generic IE is flexibly designed for bulk loading sets of flat files, Excel spreadsheets, or SAS data sets along with their associated experimental design information.



Input Engine Description

Generic Input Engine provides a means to bulk load a set of raw data files into a single SAS data set. An additional file, the Experimental Design File, is required in order to provide the experimental variables associated with each file.

Experiment Identifier

Experimental Design Table

File Type of Experimental Design Table: CSV

Folder of Raw Data Files

Type of Raw Data Files: CSV

☐ Log2Transform

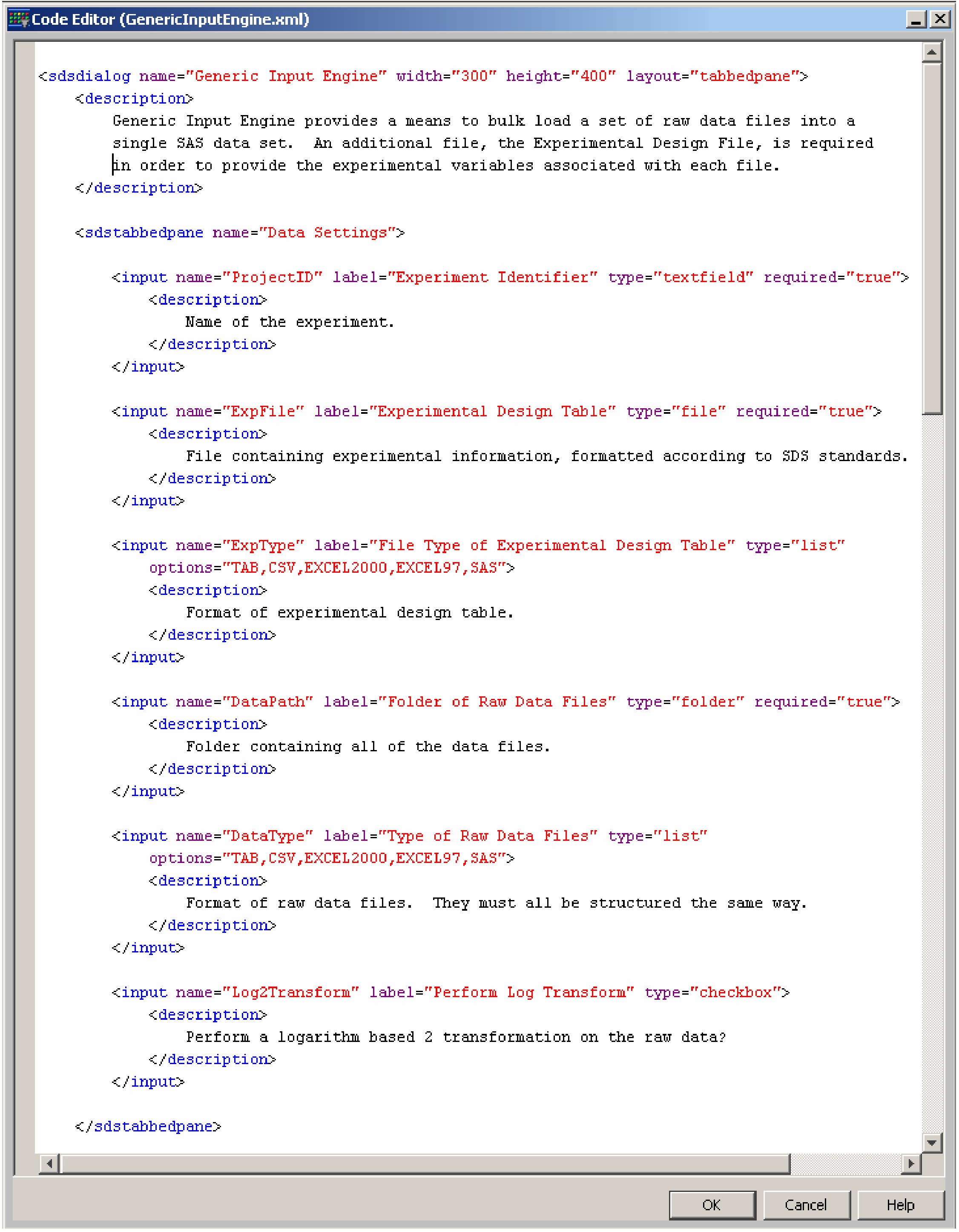
Figure 3. Generic Input Engine

The APs that come with RDM consist of the following useful data utilities:

- Data Contents
- Data Export
- Data Filter
- Data Merge
- Data Rank
- Data Sort
- Data Standardize
- Data Step
- Data Summary
- Data Transpose

Space prohibits detailed descriptions of each of these, but their titles typically correspond to the underlying SAS procedure that is invoked.

Fourth, a new XML interface is now available for implementing more user-friendly front-ends to IEs and APs. For example, Figure 4 shows the XML behind the display in Figure 3. As an IE or AP writer, you now have a flexible, powerful, and easy-to-use system for generating professional Java-Swing based graphical user interfaces (GUIs) without ever having to learn Java! You can provide drop-down lists and other value selection methods for your users, thus reducing input errors in the parameter fields. The previous text-based input GUIs are also still supported and the AP SAS code remains the same.



```
<sdsdialog name="Generic Input Engine" width="300" height="400" layout="tabbedpane">
    <description>
        Generic Input Engine provides a means to bulk load a set of raw data files into a
        single SAS data set.  An additional file, the Experimental Design File, is required
        in order to provide the experimental variables associated with each file.
    </description>

    <sdstabbedpane name="Data Settings">

        <input name="ProjectID" label="Experiment Identifier" type="textfield" required="true">
            <description>
                Name of the experiment.
            </description>
        </input>

        <input name="ExpFile" label="Experimental Design Table" type="file" required="true">
            <description>
                File containing experimental information, formatted according to SDS standards.
            </description>
        </input>

        <input name="ExpType" label="File Type of Experimental Design Table" type="list"
            options="TAB,CSV,EXCEL2000,EXCEL97,SAS">
            <description>
                Format of experimental design table.
            </description>
        </input>

        <input name="DataPath" label="Folder of Raw Data Files" type="folder" required="true">
            <description>
                Folder containing all of the data files.
            </description>
        </input>

        <input name="DataType" label="Type of Raw Data Files" type="list"
            options="TAB,CSV,EXCEL2000,EXCEL97,SAS">
            <description>
                Format of raw data files.  They must all be structured the same way.
            </description>
        </input>

        <input name="Log2Transform" label="Perform Log Transform" type="checkbox">
            <description>
                Perform a logarithm based 2 transformation on the raw data?
            </description>
        </input>

    </sdstabbedpane>
```

Figure 4. XML for Generic Input Engine

Fifth, IEs and APs in RDM and all of the solutions are now organized into convenient folders in the repository as shown in Figure 5. As repository objects, IEs and APs inherit all RDM functionality available in the repository, including security, versioning, auditing, and search capabilities.

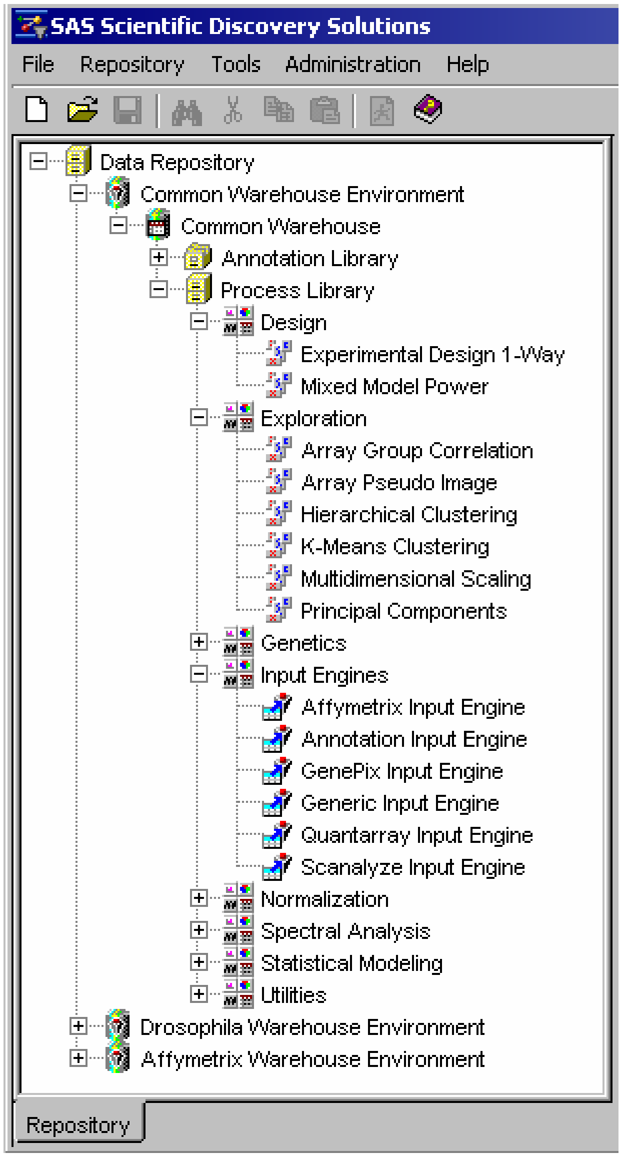


Figure 5. SAS SDS Repository

Finally, a host of minor improvements have been made to the numerous functionalities of RDM, including the following:

- User preferences for customizing common activities in RDM and setting default values
- Ability to save the last used settings and selections for reuse in various activities
- Progress indicators for all running processes, including IEs, APs, uploads, and building the repository
- Additional options for building the repository with improved efficiency
- More flexibility in assigning security records to the repository objects and other components in the solutions
- Expanded audit trail functionality to enable administrators to choose which activities to audit
- A more simplified installation and configuration for all solutions

ENHANCEMENTS TO THE SAS MICROARRAY SOLUTION

SAS Microarray 1.3 is also much more powerful than preceding releases and takes full advantage of all of the aforementioned enhancements to RDM. For example, all of the IEs and APs have been upgraded with the new XML-based GUI layer.

SAS Microarray 1.3 offers IEs for the instrumentation or scanners listed in the middle of Figure 5 as well as a Generic IE and one for loading Annotation files. An expanded API for writing your own IEs is also available. Connection to your internal database is accomplished via consultation with SAS systems engineers and field staff using SAS/ACCESS products and Proc SQL. MAGE-ML is implementable via the SAS XML engine.

SAS Microarray APs are now categorized as follows, and those followed by "(new)" are new for Release 1.3. Note that the underlying functionality of the new APs has existed in previous releases, but front-ends and macros for them are now officially supported within the system.

DESIGN

Experimental Design One-Way randomly assigns levels of a single treatment to arrays. You specify the number of treatment levels, the number of arrays, and the number of dyes, and the process tells you how to hybridize the arrays in an optimal fashion.

Mixed Model Power (new) computes the statistical power of a set of hypothesis tests arising from a mixed linear model; refer to Wolfinger et al (2001) for an example. You specify the experimental design, PROC MIXED statements, variance components, and ranges of values for alpha and effect sizes, and the process outputs a table of power values calculated using a noncentral *t*-distribution. Figure 6 displays some example output, showing how power increases with effect size and alpha value.

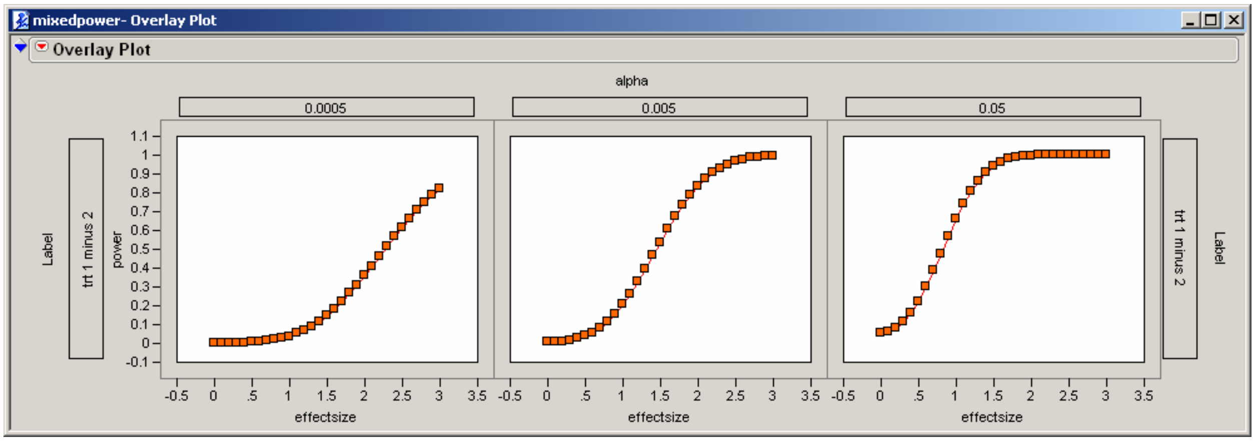


Figure 6. Output from Mixed Model Power

EXPLORATION

Array Group Correlation computes correlations and scatterplot matrices for expression measurements across groups of arrays. You can also use it to compute various transposed configurations of the input data set.

Array Pseudo Image generates interactive pseudo image plots of array data as shown in Figure 7. You can examine such plots for abnormal patterns or spatial trends, and zoom in on interesting areas.

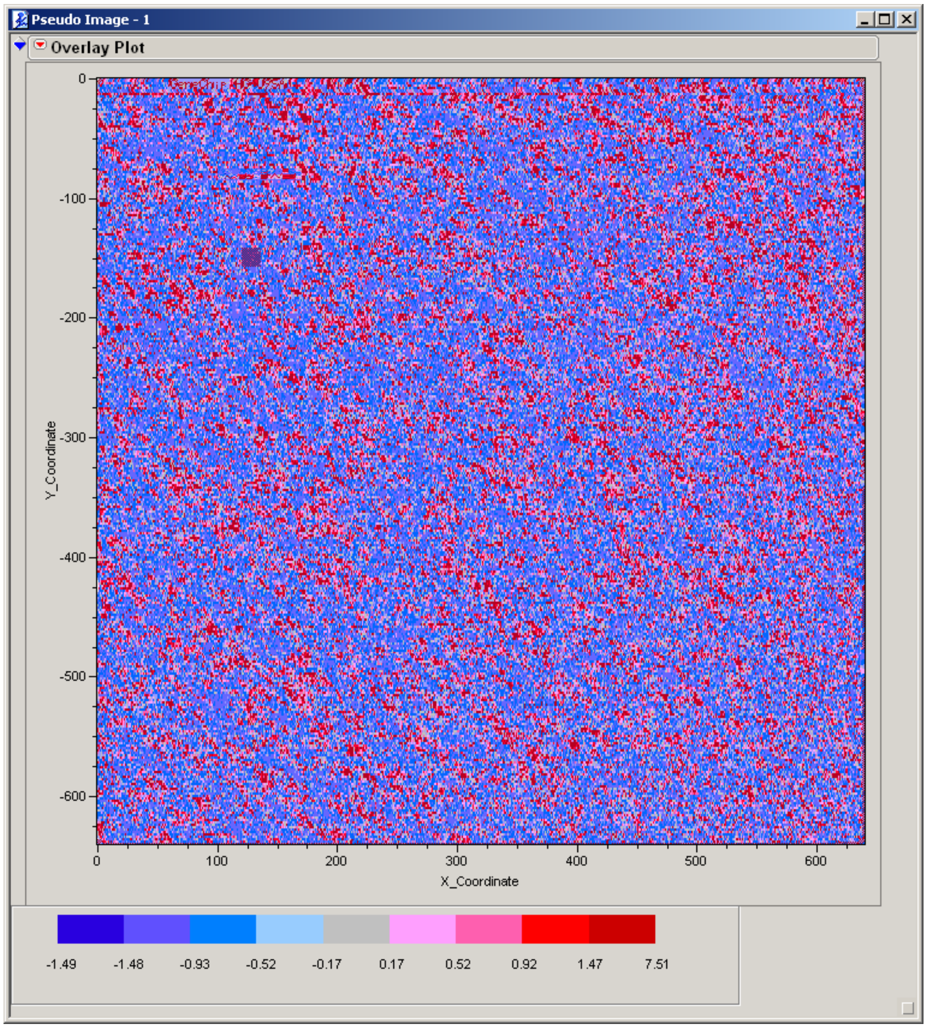


Figure 7. Pseudo Image of an Affymetrix GeneChip

Hierarchical Clustering (new) creates a cell plot (heat map) and dendrogram organization of observations using one of the following five distance methods: Average Linkage, Centroid, Complete Linkage, Single Linkage, or Ward's. Variables may also be clustered to form a two-way clustering arrangement.

K-Means Clustering (new) forms optimally separated clusters of observations using the classical *k*-means criterion. Separation rules may additionally be based upon Normal Mixtures or Self-Organizing Maps. Figure 8 displays parallel coordinate plots of an example with six k-means clusters computed from the standardized least squares means for 100 sample genes from the *Drosophila* aging data (Jin et al, 2001).

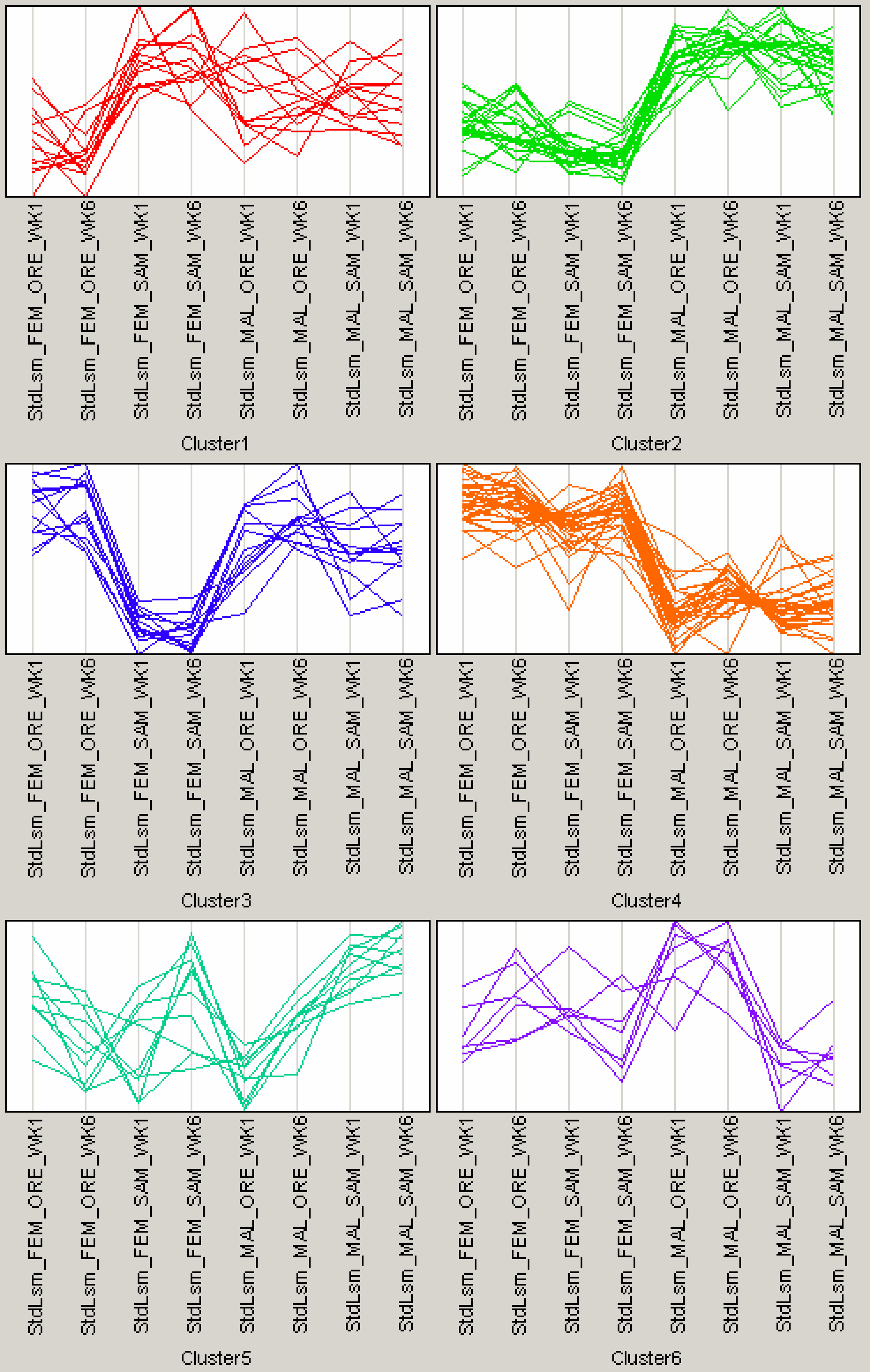


Figure 8. Example Results of K-Means Clustering

Multidimensional Scaling (new) provides an optimal 2- or 3-D representation of points for which distances have been computed in higher dimensions.

Principal Components (new) computes linear projections that maximally separate the data in terms of variance. This AP uses Proc PLS to compute the components, which has the advantage of using a fast iterative algorithm that accommodates missing values.

NORMALIZATION

Data Standardize (new) standardizes values of numeric variables using a variety of different methods, including Std, Mean, Median, Range, IQR, and MAD from PROC STDIZE in SAS/STAT.

Mixed Model Normalization normalizes microarray data by fitting a mixed linear model across all of the arrays in an experiment, and is a natural precursor to a gene-by-gene mixed model analysis of variance. It creates an output SAS data set that includes the input data along with a new, normalized variable for the normalized intensities.

Quantile Normalization (new) normalizes data by aligning their quantiles (Irizarry et al, 2003).

STATISTICAL MODELING

Discriminant Analysis (new) is a classical statistical method for predicting a classification variable from a set of continuous responses. It can be regarded as inverse prediction from a multivariate analysis of variance (MANOVA).

Mixed Model Analysis fits a mixed linear model on a gene-by-gene basis to prenormalized microarray data and creates various output displays (Chu et al, 2002).

Partial Least Squares (new) is a method for simultaneously modeling variability in both dependent variables and predictor variables. Like principal components analysis, the method works by extracting successive linear combinations of the predictor variables, but the linear combinations are chosen to jointly maximize variance explained in both the Xs and Ys in the model.

Figure 9 displays a scatterplot matrix of three PLS components computed using a two-level categorical variable as the Y variable. The points are colored red and green according to the two levels, and the first component alone does a good job at separating them.

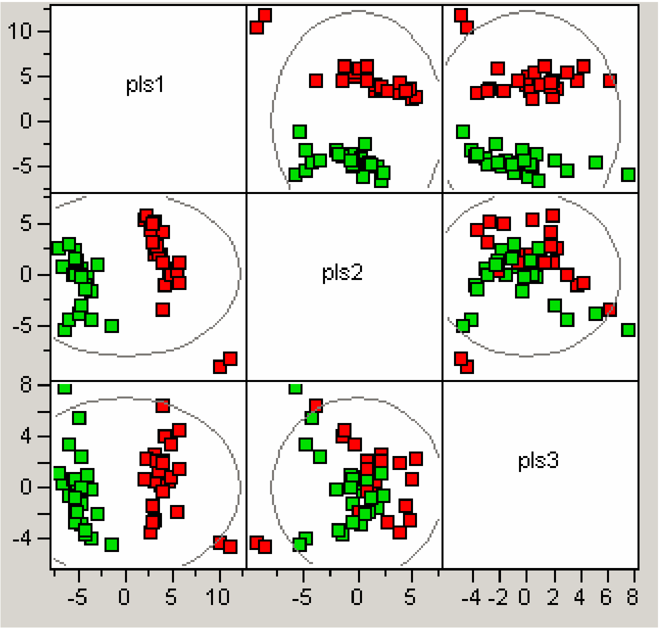
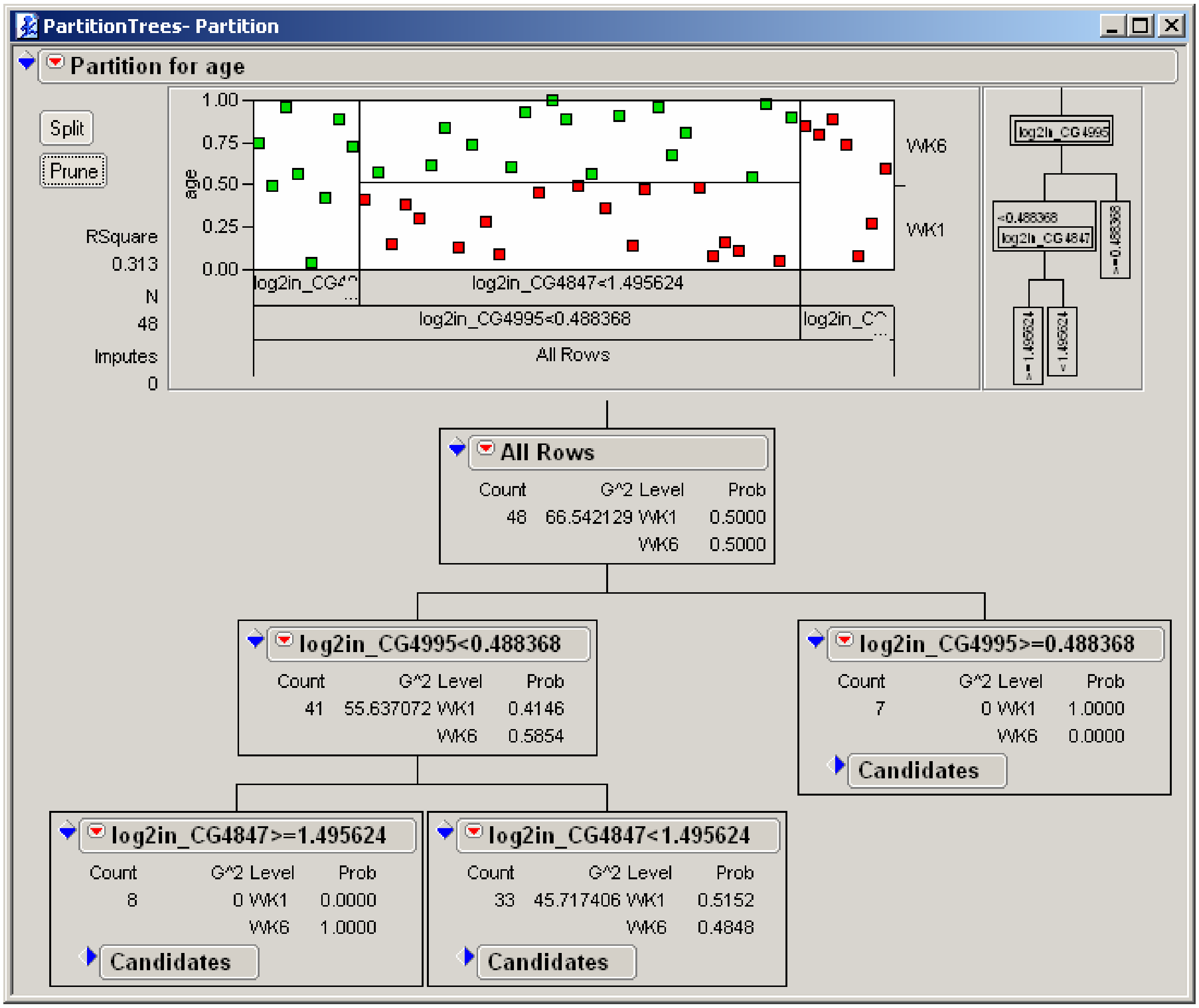


Figure 9. Plots of PLS Components, Colored by a Classification Variable

Partition Trees (new) recursively partitions data according to optimal splitting relationships created between dependent and predictor variables. It enables you to interactively create simple tree-based rules for predicting the dependent variable.

Figure 10 is also based upon a subset of the *Drosophila* data of Jin et al (2001). This analysis predicts age (with two levels: 1 week and 6 weeks) using gene expression measurements as predictors. The gene CG4995 is selected as the first splitting variable, and then CG4847 as the second to partition the group with values of CG4995 below 0.488368. The R-Square for the depicted model is only 0.313, and so additional splits will likely be necessary on the bottom middle leaf in order to obtain better predictions.



```
RSquare
  0.313
      N
     48
Imputes
      0
```

```
All Rows
Count    G^2 Level        Prob
   48    66.542129 WK1    0.5000
                   WK6    0.5000
```

```
log2in_CG4995<0.488368
Count    G^2 Level        Prob
   41    55.637072 WK1    0.4146
                   WK6    0.5854
```

```
log2in_CG4995>=0.488368
Count    G^2 Level        Prob
    7            0 WK1    1.0000
                   WK6    0.0000
```

```
log2in_CG4847>=1.495624
Count    G^2 Level        Prob
    8            0 WK1    0.0000
                   WK6    1.0000
```

```
log2in_CG4847<1.495624
Count    G^2 Level        Prob
   33    45.717406 WK1    0.5152
                   WK6    0.4848
```

Figure 10. Example Output from Partition Trees

ANNOTATION

In addition to the preceding APs, SAS Microarray 1.3 includes enhancements to annotating genes and results. The **Create Annotation Engine** provides an easy and automatic way to generate and update gene annotation information against various biomedical databases. Information in these databases is growing exponentially and updated frequently, and the Create Annotation Engine enables you to obtain the latest gene information whenever you desire.

The generated gene annotation contains information such as Genbank accession number, UniGene cluster title and ID, gene symbol, LocusLink ID, OMIM ID, GI number, protein sequence ID, EST's clone ID, Library ID, associated GO terms, and KEGG pathway map ID. Various approaches are used to access these databases:

- A local version of the UniGene database is automatically uploaded into the SDS system periodically. Accessing a local image of the UniGene database has much better performance than direct access to the UniGene web interface.
- A local version of the MySQL GO database is accessed through its SQL interface.
- The KEGG database is accessed through its Web Service interface.

To run the annotation engine, you must create an input file containing a gene identifier variable such as GenBank accession or UniGene cluster ID as well as a key variable such as AffyID or Spot ID for merging with microarray data sets or result data sets after analyses. Table 1 provides a short example of such a table; its format can be Comma Separated, Microsoft Excel, SAS Data Set, or Tab Delimited.

Table 1. Example of an Input File for Creating or Updating Annotation

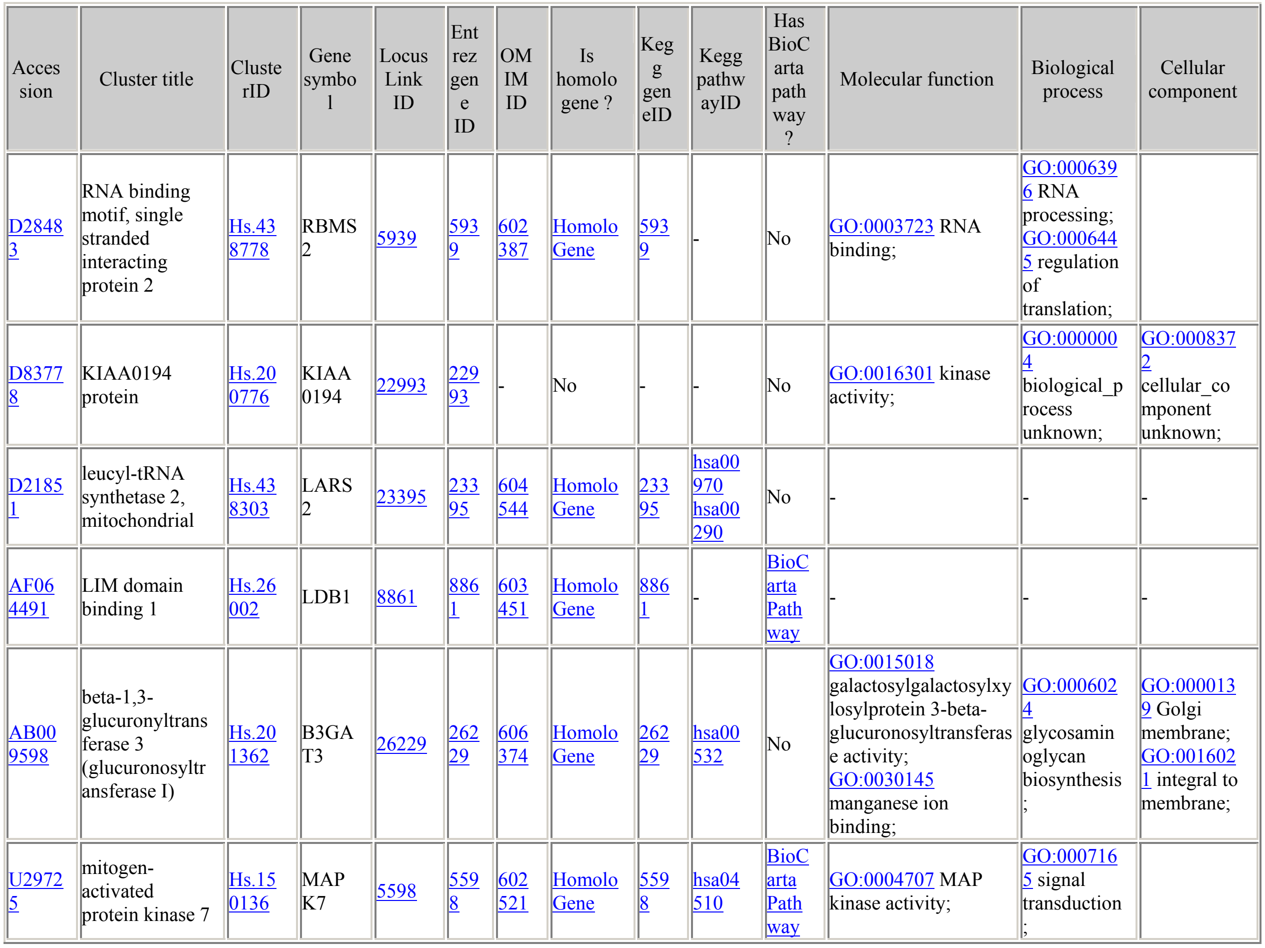
AffyID_REF	GenBank Accession
AFFX-BioB-5_at	J04423
AFFX-HUMRGE/M10098_3_at	M10098
AFFX-HSAC07/X00351_M_at	X00351
1729_at	L41690
266_s_at	L33930
32042_at	S72904
32541_at	S46622
34471_at	M36769
202435_s_at	NM_000104.2
202438_x_at	BF346014
202439_s_at	NM_000202.2

The annotation output files are stored in RDM in both SAS data set and HTML formats. The HTML file provides convenient clickable links to the public databases, including ones to Genbank, UniGene, LocusLink, Entrez Gene, OMIM, HomoloGene, Gene Ontology, and the KEGG Pathway Database. Table 2 shows a few rows from an example-generated annotation file in HTML format.

In addition to annotation creation and loading functions, the Web Search Dialog surfaced in JMP features new links to many of the aforementioned databases, as well as an automated bulk upload of a list of gene IDs to the curated database Onto-Express (http://vortex.cs.wayne.edu/projects.htm).

Finally, overlaying expression results on pathway diagrams can be very useful for scientific discovery. For this functionality the input file contains two variables; one is the KEGG ID and the other is the gene's expression result such as *p*-value or fold change. The output is a list of pathway diagrams on which the genes of interested are colored according their expression result.

Table 2: Subset of Generated SAS Annotation information in HTML Format



Accession	Cluster title	ClusterID	Gene symbol	LocusLink ID	Entrez gene ID	OMIM ID	Is homologene ?	Kegg geneID	Kegg pathwayID	Has BioCarta pathway ?	Molecular function	Biological process	Cellular component
D28483	RNA binding motif, single stranded interacting protein 2	Hs.438778	RBMS2	5939	5939	602387	HomoloGene	5939	-	No	GO:0003723 RNA binding;	GO:0006396 RNA processing; GO:0006445 regulation of translation;	
D83778	KIAA0194 protein	Hs.200776	KIAA0194	22993	22993	-	No	-	-	No	GO:0016301 kinase activity;	GO:0000004 biological_process unknown;	GO:0008372 cellular_component unknown;
D21851	leucyl-tRNA synthetase 2, mitochondrial	Hs.438303	LARS2	23395	23395	604544	HomoloGene	23395	hsa00970 hsa00290	No	-	-	-
AF064491	LIM domain binding 1	Hs.26002	LDB1	8861	8861	603451	HomoloGene	8861	-	BioCarta Pathway	-	-	-
AB009598	beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)	Hs.201362	B3GAT3	26229	26229	606374	HomoloGene	26229	hsa00532	No	GO:0015018 galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase activity; GO:0030145 manganese ion binding;	GO:0006024 glycosaminoglycan biosynthesis;	GO:0000139 Golgi membrane; GO:0016021 integral to membrane;
U29725	mitogen-activated protein kinase 7	Hs.150136	MAPK7	5598	5598	602521	HomoloGene	5598	hsa04510	BioCarta Pathway	GO:0004707 MAP kinase activity;	GO:0007165 signal transduction;	

THE SAS GENETIC MARKER SOLUTION

The SAS Genetic Marker Solution offers IEs for importing and APs for analyzing genetic marker data, which contain unrelated individuals or nuclear families sampled from a population and their genotypes at a set of genetic markers. There are Genetic Marker IEs specifically designed for data in the Arlequin or NEXUS format, while data in more conventional formats such as the LINKAGE PEDFILE can be accommodated by a Generic IE. The APs utilize procedures in SAS/Genetics and SAS/STAT for analyses that are critical in determining the genetic basis of a specific disease or trait being studied. The following APs are provided by the Genetic Marker Solution:

Case-Control Association performs significance testing for association between single-marker genotypes or alleles and a binary trait. The *p*-values from these tests can be adjusted and corrected for multiple testing, then plotted along the marker map. Figure 11 displays the negative log *p*-values from the three available case-control tests with a reference line at the requested alpha level so that markers with significant *p*-values (those above the reference line) can be easily identified. *p*-values from other Genetic Marker APs discussed below that calculate single-marker association test statistics can be adjusted then visualized in the same manner

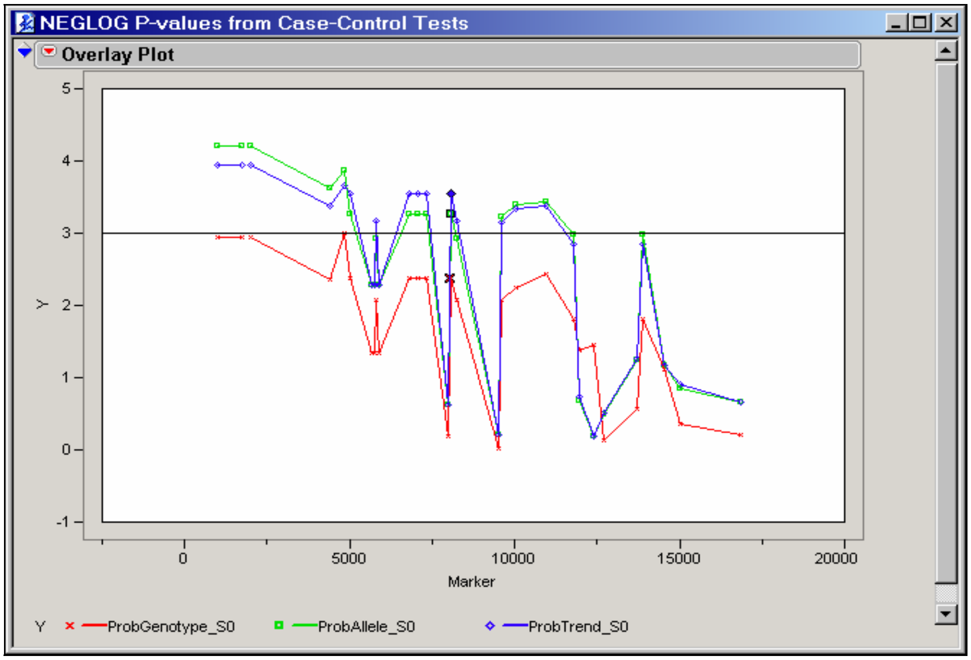


Figure 11. Plot of *p*-values from Case-Control Association AP

Haplotype Estimation calculates maximum likelihood estimates of haplotype frequencies using the expectation-maximization (EM) algorithm. Tests for multi-locus linkage disequilibrium (LD) and associations between a binary trait and haplotypes can also be performed.

Haplotype Trend Regression uses the individuals' haplotype probabilities from PROC HAPLOTYPE as inputs to a regression to model the effect of haplotypes on a trait.

Linkage Disequilibrium reports test statistics and various measures of LD between pairs of markers. These statistics are represented graphically (see Figure 12): the colors of symbols on the scatter plot correspond to various *p*-value ranges for the LD test statistic, a contour plot illustrates changing ranges of values of an LD measure across the marker map, and charts display estimates of two-locus haplotype frequencies.

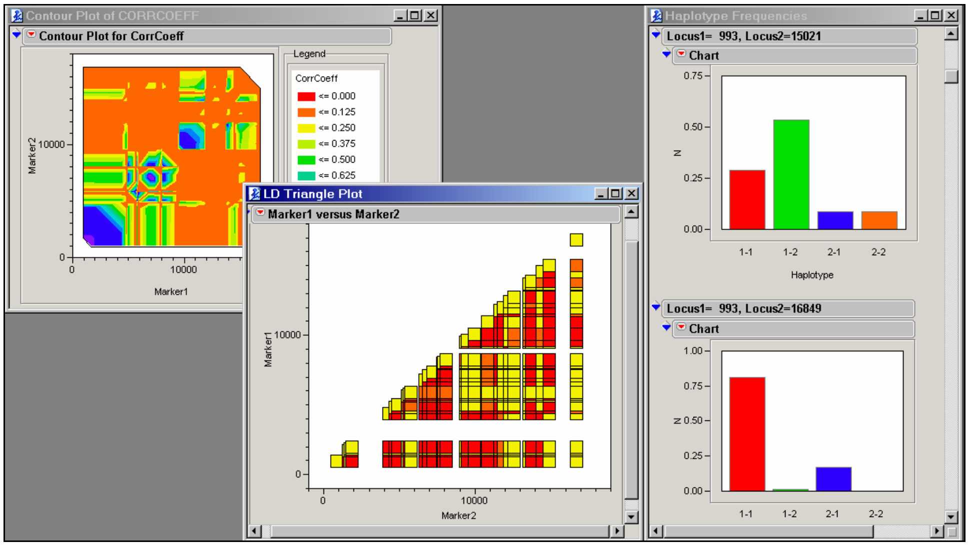


Figure 12. Visual Representations of Statistics from Linkage Disequilibrium AP

Marker Properties provides summary statistics of markers, including measures of marker informativeness, test statistics for Hardy-Weinberg equilibrium, and allele and genotype frequencies. In addition to reporting these in tabular format, this AP creates frequency charts and a cell plot representing individuals' genotypes as shown in Figure 13.

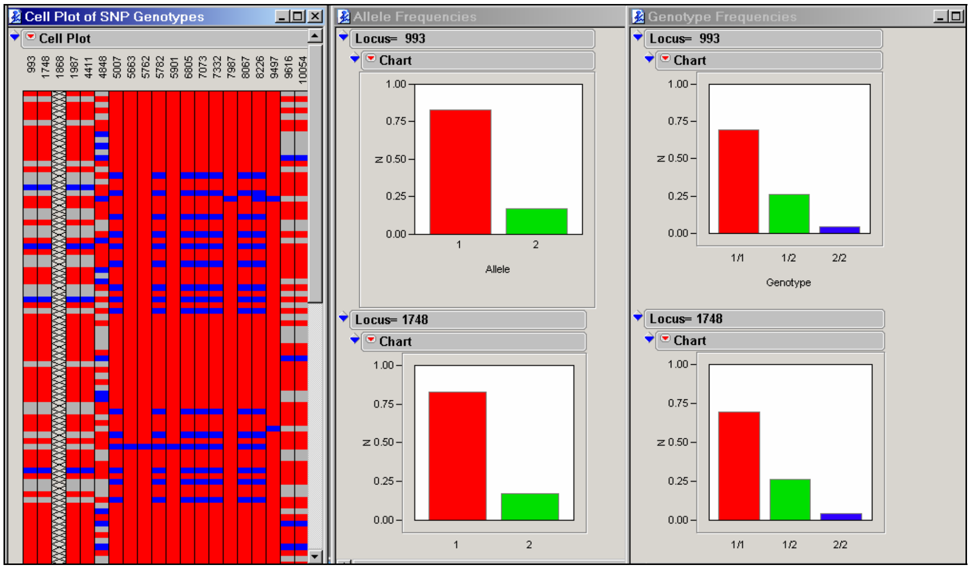


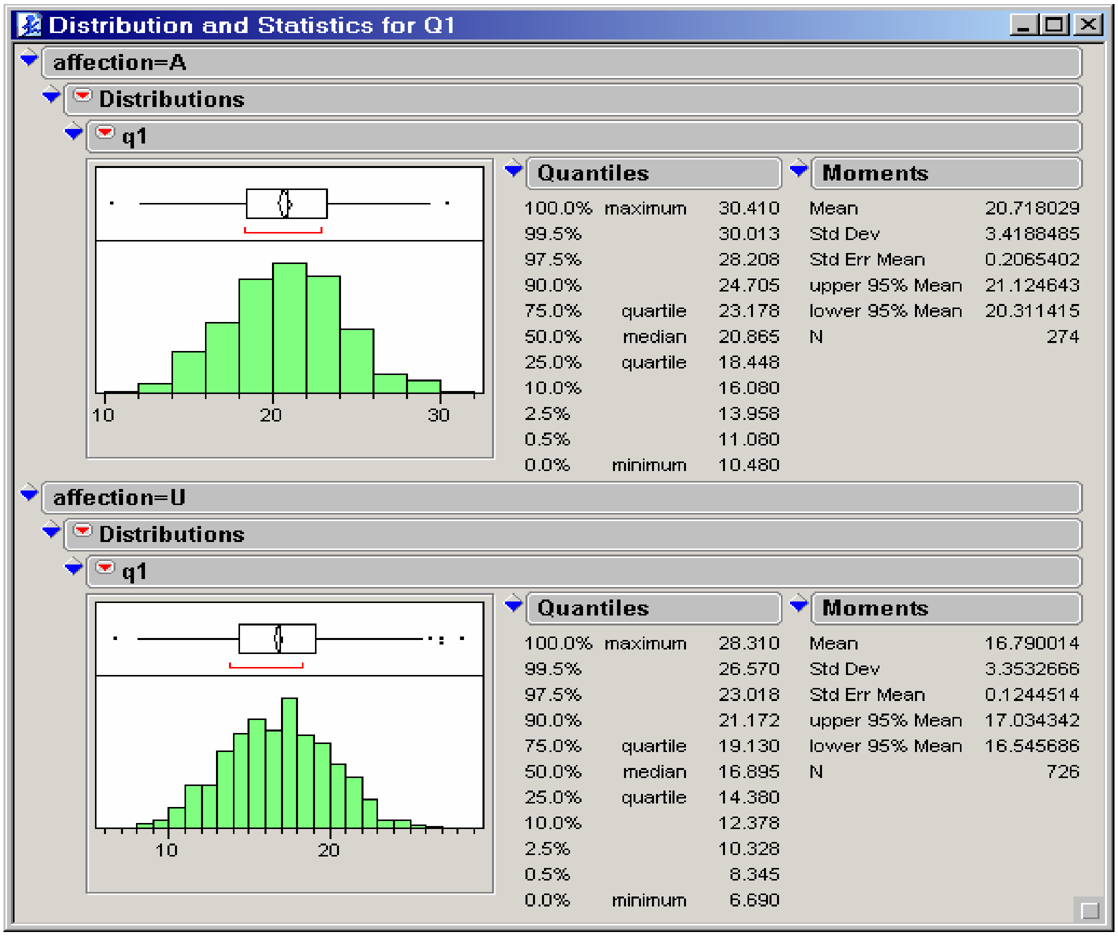
Figure 13. Cell Plot and Frequency Charts from Marker Properties AP

Phenotype Summary calculates and graphs frequencies for discrete phenotypic variables or distributional statistics for continuous variables, as shown in Figure 14.

Quantitative TDT provides tests for linkage and association between individual markers and a quantitative trait using samples of nuclear families.

Quantitative Trait Association performs significance testing for association between single-marker genotypes or alleles and a quantitative trait, and also fits a regression model using multiple markers that are added to or remain in the model according to the selection method applied.

TDT provides tests for linkage and association between individual markers and a binary trait using samples of nuclear families.



affection=A

Quantiles

100.0%	maximum	30.410
99.5%		30.013
97.5%		28.208
90.0%		24.705
75.0%	quartile	23.178
50.0%	median	20.865
25.0%	quartile	18.448
10.0%		16.080
2.5%		13.958
0.5%		11.080
0.0%	minimum	10.480

Moments

Mean	20.718029
Std Dev	3.4188485
Std Err Mean	0.2065402
upper 95% Mean	21.124643
lower 95% Mean	20.311415
N	274

affection=U

Quantiles

100.0%	maximum	28.310
99.5%		26.570
97.5%		23.018
90.0%		21.172
75.0%	quartile	19.130
50.0%	median	16.895
25.0%	quartile	14.380
10.0%		12.378
2.5%		10.328
0.5%		8.345
0.0%	minimum	6.690

Moments

Mean	16.790014
Std Dev	3.3532666
Std Err Mean	0.1244514
upper 95% Mean	17.034342
lower 95% Mean	16.545686
N	726

Figure 14. Distribution Plot and statistics of Q1 by Affection from Phenotype Summary AP

THE SAS PROTEOMICS SOLUTION

The SAS Proteomics Solution is still under development and is primarily designed to handle spectral data from a variety of instruments including mass spectrometers (MS) and nuclear magnetic resonance (NMR) machines. It includes several of the APs that are also a part of the SAS Microarray Solution, including Experimental Design One-Way, Mixed Model Power, Array Group Correlation, Multidimensional Scaling, Principal Components, Mixed Model Normalization, Discriminant Analysis, Partial Least Squares, Partition Trees, and Mixed Model Analysis. SAS Proteomics also features the following APs customized for spectral data:

Spectral Bin reduces the total number of index values by binning them and averaging the spectra within bins.

Spectral Detrend removes a baseline trend from each member of a group of spectra in order to make them statistically comparable.

Spectral Peak Find finds and quantifies peaks in a group of spectra.

Spectral Plot creates plots of spectra. Three plots are currently supported: a cell plot (heat map) of the data, an overlay plot (line plot) of the data, and an overlay plot of the means of two groups with confidence bands.

Figure 15 shows the last of these three plots before and after running Spectral Bin and Spectral Detrend. The plots overlay the means of two groups of spectra (cancer and normal, colored red and green, respectively) along with confidence bands computed from a t-distribution. The black spectrum along the bottom (indexed on the right vertical axis) displays the negative log 10 p-value of the t-test between the two groups at each m/z (mass over charge) value. Refer to Hawkins et al. (2003) for some additional plots and details on alternative statistical methods for spectral data.

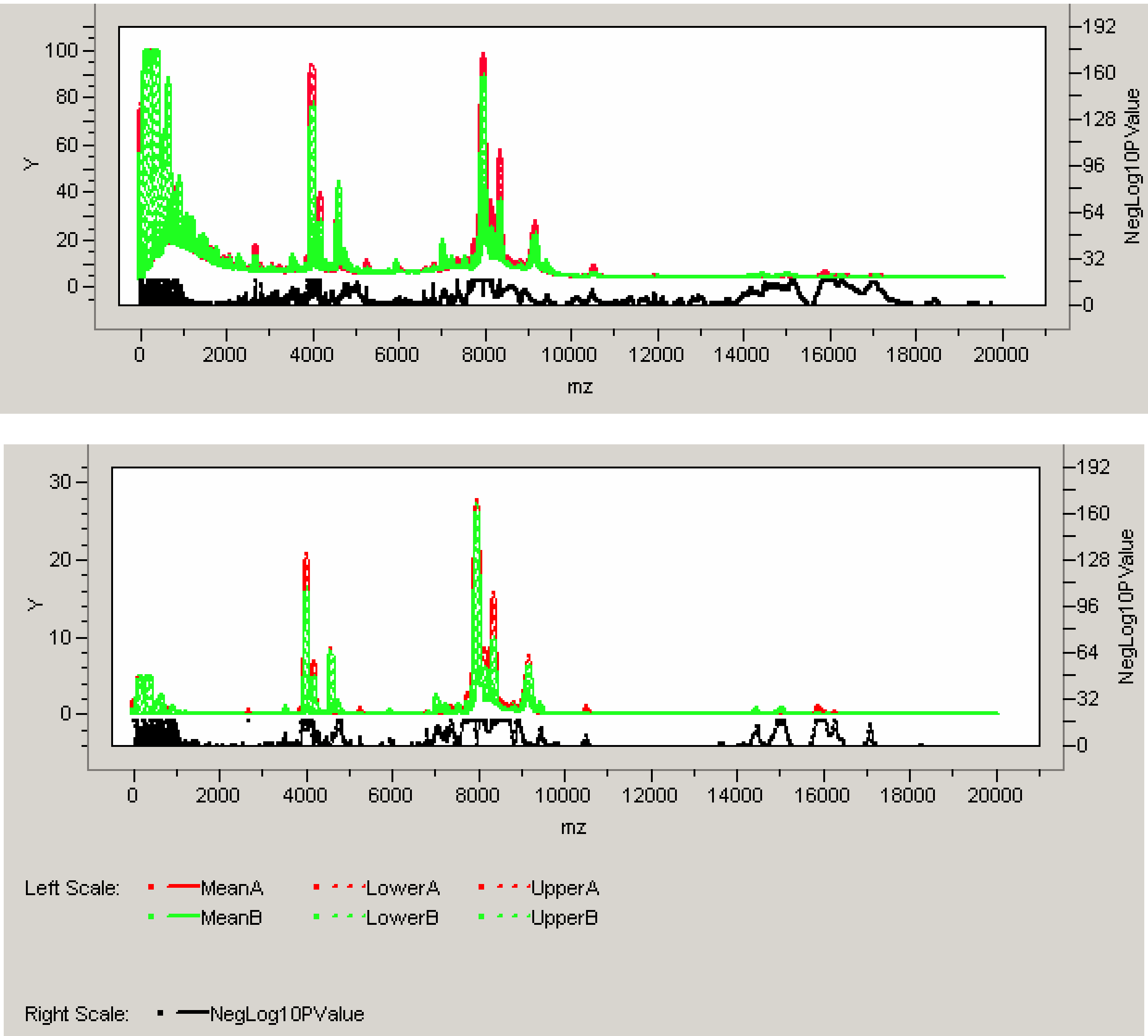


Figure 15. Output from Spectral Plot AP, the top graph is from the original data, and the bottom graph is after pre-processing the data with Spectral Bin and Spectral Detrend

CONCLUSION

SAS SDS is evolving into a uniquely powerful software system for scientific discovery. The rich new features in SAS/Genetics, RDM, and the SAS Microarray Solution, along with the new SAS Genetic Marker and SAS Proteomics solutions, combine to provide you with unprecedented capabilities to successfully surf the ever growing waves of genomics data. More information about the solutions (which come bundled with many classical SAS products) is available at http://www.sas.com/industry/pharma/sds/index.html. SAS looks forward to working together with you to enhance your research collaborations.

REFERENCES

Bacanu, S-A., Devlin, B., and Roeder, K. (2000), "The Power of Genomic Control," *American Journal of Human Genetics,* 66, 1933 -1944.

Chu, T.-M., Weir, B., and Wolfinger, R.D. (2002), “A systematic statistical linear modeling approach to oligonucleotide array data analysis,” *Mathematical Biosciences,* 176, 35-51, http://statgen.ncsu.edu/ggibson/Pubs.htm.

Clayton, D. (2002), "SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs," http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt.

Hawkins, D.M., Wolfinger, R.D., Liu, L., and Young, S.S. (2003), “Exploring Blood Spectra for Signs of Ovarian Cancer,” *Chance,* 16(4), 19-23.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003). “Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.” *Biostatistics*. Vol. 4, Number 2: 249-264.

Jin, W., Riley, R., Wolfinger, R.D., White, K.P, Passador-Gurgel, G. and Gibson G. (2001) Contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*, *Nature Genetics,* 29:389-395.

Johnson, G.C.L. et al. (2001), "Haplotype Tagging for the Identification of Common Disease Genes," *Nature Genetics*, 29, 233 - 237.

McIntyre, L., Coffman, C., and Doerge, R. (2001), “Detection and Location of Single Binary Trait Loci in Experimental Populations,” *Genetical Research*, 78, 79 - 92.

Neville, P., Tan, P.-Y., Mann, G., and Wolfinger, R.D. (2003) Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum, *Proteomics,* 3(9), 1710-1715.

Simonsen, K. L. (2004), “A Probability Model for the Inheritance of Binary Traits,” Purdue University Statistics Department Technical Report Series #tr03-04.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001), “Assessing gene significance from cDNA microarray data via mixed models,” *Journal of Computational Biology,* 8, 625-637, http://statgen.ncsu.edu/ggibson/Pubs.htm.

Wolfinger, R.D., Quinlan, K., and Flood, S. (2003) An Introduction to genomics and SAS Scientific Discovery Solutions, SUGI 28 Proceedings, SAS Institute, Inc, Cary, NC.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., and Weir, B.S. (2002), "Truncated Product Method for Combining P-values," *Genetic Epidemiology*, 22, 170 - 185.

ACKNOWLEDGMENTS

It is infeasible to list here the scores of individuals who have helped with continued progress on SAS SDS, although we would like to thank the following for their singular contributions: **SAS Genomics R&D:** Carey Carpenter, Tzu-Ming Chu, Christine Duarte, Samuel Johnson, Geoffrey Mann, Chandrika Nayak, Xiang Yu, Jun Zhang; **Testing:** Jack Berry, Shibing Deng, Kelly Graham, Gerardo Hurtado, Susan Shao; **R&D:** Leonardo Auslender, Julia Carpenter, James Cox, Cheryl Doninger, David Duling, Alan Eaton, Drew Foglia, James Holman, Bob Huemmer, Joe Mudd, Padraic Neville, Oliver Schabenberger, Kevin Scott, Maura Stokes, Bob Rodriguez, Pei-Yi Tan, Mike Thorland, Randy Tobias, Peter Villiers; **Marketing:** Susan Flood, Ed Helton; **JMP:** Brian Corcoran, Bob Hickey, Kathy Jablonski, Brad Jones, Kristin Nauta, Richard Potter, John Sall, Toshi Tsuboi; **Education:** Rick Cornell, Herb Kirk, Bob Lucas, Linda Miller, Kari Richardson, Lori Rotheberg, Jill Tao; **Publications:** Jill Fitzgibbons; **Editing:** Virginia Clark; **Technical Support**: Rob Agnelli, Kathleen Kiernan, Phil Gibbs, Susan Powers,Eddie Routten, Toby Trott; **Public Sector**: John Bander, Darrell Barton, Christine Carrmichael, Dara Hammond, Courtney Henson, Sharon Lang, Sally Muller, Kay Obenshain, Mel Proctor; **Life Sciences Organization:** Tammy Baird, Eugenia Bastos, Joe Carter, Ron Coleman, Pete Curtice, Glen Dalrymple, Dave Danyluk, Dwight Fowler, Harry Hart, Chris Glass, Michael Kilhullen, Pat McDonald, Bob McMahon, Tiffany Miller, Bob Trepper, Jeff Wang, Pete Whyte; **EMEA:** Ad de Groot, Gerhard Held, Hans Meyer, Karen Miller; **External Collaborators:** Sam Adams, Kurt Brumbaugh, Shannon Burns, Pierre Bushel, Gary Churchill, Cynthia Coffman, Meg Ehm, Greg Gibson, Wen-Ping Hsieh, Bob Kelly, Stan Martin, Lauren McIntyre, Jerry Nehls, Dhalia Nielsen, Rick Paules, Bruce Weir, Liz Wolfinger, Len van Zyl, plus loads of thanks to all of our SAS/Genetics and SDS customers!

CONTACT INFORMATION

We value and encourage your comments and questions. For questions regarding SAS Scientific Discovery Solutions, please contact

Susan Flood
SAS Institute Inc.
SAS Campus Drive
Cary, North Carolina 27513
Phone: (919) 677-8000
Email: Susan.Flood@sas.com
Web: www.sas.com