Paper 259-28

An Introduction to Genomics and SAS® Scientific Discovery Solutions

Russ Wolfinger, SAS Institute Inc., Cary, NC
Kristen Quinlan, SAS Institute Inc., Cary, NC
Susan Flood, SAS Institute Inc., Cary, NC

ABSTRACT

The commoditization of genomics instrumentation has created a strong need for management and analysis of the resultant high volumes of complex data. This paper provides a brief description of the basic science underlying these data sources and an overview of the initial steps SAS has taken to address them in our new SAS Scientific Discovery Solutions product line. Three core SAS foundational elements--data warehousing, analytical servers, and JM--apply across the three principal foci of the central dogma of molecular biology: genetics, transcriptomics, and proteomics. The integration of these elements in a unified application provides you with unprecedented power to intelligently extract knowledge from your genomics data and make breakthrough discoveries. We highlight the technical aspects of this system using a microarray data example and then discuss several new development directions.

INTRODUCTION

The central dogma of molecular biology (Figure 1) provides the crucial scientific underpinnings of the content of this paper. Deoxyribonucleic acid (DNA), in the form of the classical double-helix built with the four amino acids adenine, cytosine, guanine, and thymine, exists and self-replicates in the cell nucleus. Genes, which are small segments of DNA, code for the transcription of thousands of different forms of ribonucleic acid (RNA). RNA molecules move across the nuclear membrane into the cell cytoplasm and serve as translation templates for tens of thousands of proteins. Gibson and Muse (2002) provide a more in-depth overview of the intricacies involved in this fascinating process.

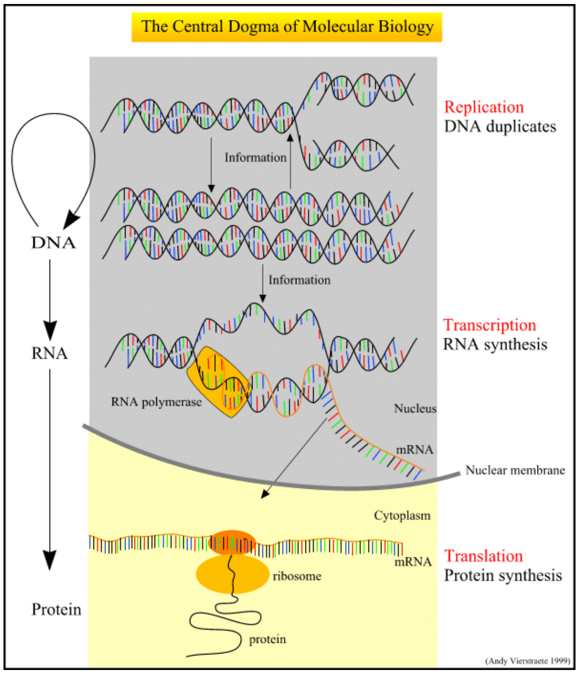


Figure 1: The Central Dogma of Molecular Biology

The central dogma has been well-known for decades, so why such a stir about it now? The answer is at the bottom of Figure 2. Breakthroughs in microtechnological instrumentation, such as DNA sequencers, microarrays, and mass spectrometers, have revolutionized laboratory practice over the past few years, transforming molecular biology from a data-poor to a data-rich enterprise. High-throughput sequencers create maps of the entire DNA blueprint of common organisms, including the 3 billion base pairs comprising the human genome. Microarrays provide a snapshot of the number of gene transcripts in a given biological sample, enabling gene expression profiling across thousands of genes simultaneously. Mass spectrometers quantitate thousands of protein levels from a sample using advanced laser and time-of-flight techniques. The data sets from these instruments, typically several megabytes from a single experiment, demand the use of advanced computer science and analytical techniques for intelligent interpretation. This exciting conflux makes SAS an excellent software platform from which to conduct genomics discovery.

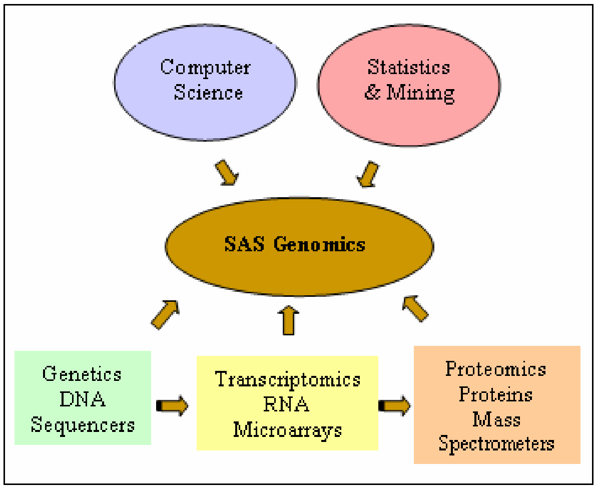
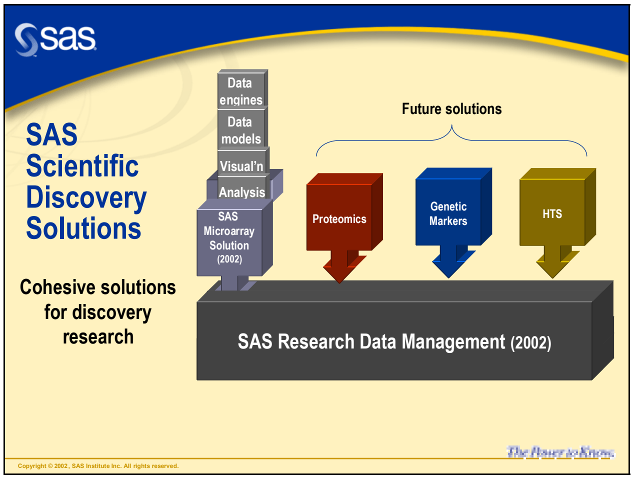


Figure 2: The central dogma and its attending technologies drive SAS Genomics.

[Note on terminology: We use "genomics" as a catch-all word to describe the aforementioned interplay between molecular biochemical data and analytical software. Another popular term is "bioinformatics," which is often used in the same context.]

What specific SAS technologies are applicable to genomics? Considering the breadth and depth of difficulties faced in this field, nearly all of them have bearing. It is difficult to know where to start. To provide appropriate focus, scope, and relevance for this paper, we concentrate on an example from the second stage of the central dogma, transcriptomics, and the key constituents of the initial release of the SAS Scientific Discovery Solutions (SDS) bundle. Figure 3 shows the foundation of this bundle as SAS Research Data Management (RDM), a Java-based graphical user interface to SAS/Warehouse Administrator that enables you to extract, transform, load, and manage genomics research data in a flexible and open fashion. The first vertical application of RDM is the SAS Microarray Solution (MAS), designed specifically for data on gene transcription. It includes input engines and analytical processes for various kinds of microarray data, and is also open and customizable. Examples and details about RDM and MAS follow in the next two sections. The final section covers several new applications and future development directions for SAS SDS.



sas

SAS Scientific Discovery Solutions

Cohesive solutions for discovery research

Data engines

Data models

Visual'n

Analysis

SAS Microarray Solution (2002)

Future solutions

Proteomics

Genetic Markers

HTS

SAS Research Data Management (2002)

Copyright © 2002, SAS Institute Inc. All rights reserved.

The Power to Know

Figure 3: RDM as a Foundation for SAS SDS

EXAMPLE: TRANSCRIPTION IN *DROSOPHILA*

Jin et al. (2001) describe results of a microarray experiment for assessing transcriptional changes in fruit flies. Twenty-four two-color arrays, each probing approximately 3,000 genes, reveal gene expression differences between two lines (Samarkand and Oregon), two genders (Female and Male), and two ages (1 week and 6 week). Figure 4 shows the basic steps for generating data from a single, generic, two-color array. Note that with this kind of array, two signals are generated for each gene on each array (one from the Cy3 fluorescent dye and one from the Cy5 fluorescent dye), so a split-plot arrangement is possible for the three experimental factors.

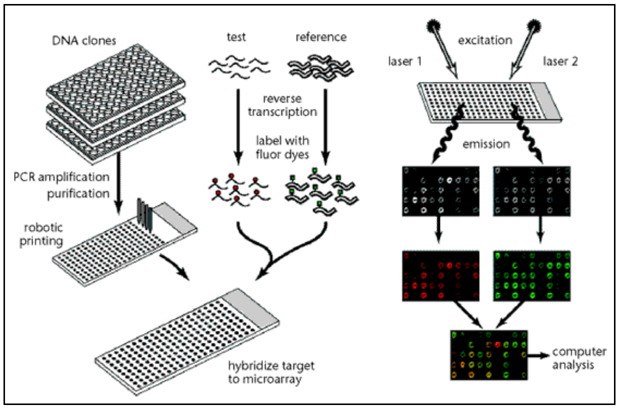


Figure 4: Data Generation Process for a Two-color Microarray

Since primary interest is in the aging effect, Jin et al. consider age (and dye, by necessity) to be the subplot factors, and line and gender to be the whole-plot factors. This experimental design is much more efficient than the popular reference sample design, in which a noninformative reference sample is always hybridized in either the Cy3 or Cy5 channel (Kerr and Churchill, 2001). In addition, this design enables simultaneous assessment of the main effects and interactions of the age, line, and gender effects on each gene, all adjusted for potential dye effects.

We use data for 100 genes from this experiment as a running example below. This small subset is one of two test examples that ship with the SAS Microarray Solution software. Refer to Hsieh et al. (2002) for details about the other example, from the Affymetrix Latin Square experiment, and to Chu et al. (2002a, 2002b) for issues surrounding Affymetrix data analysis in general.

SAS RESEARCH DATA MANAGEMENT

The vast amount of data generated by genomic experiments presents challenges in terms of managing and sharing data within and among organizations. Researchers need a direct, non-programming approach for acquiring data for analysis. Furthermore, numerical data are not the only information of interest. You may also want to view images, documents, and analysis results associated with the data.

SAS RDM is the data management core of the SAS Scientific Discovery Solutions. It is a Java client-server application that utilizes SAS warehousing technologies to provide a centralized repository for your organization's discovery research data, and other ancillary information.

DATA WAREHOUSING

The concept behind data warehousing involves extracting data and reorganizing that data in a form better suited to analysis and reporting. The key to data warehousing is metadata, or data about data. Examples of metadata include a data file name, location, storage format, and structure. A data warehouse captures metadata throughout the warehousing process: where the data came from, what transformations were performed, and information about the context of the data. Managing the metadata and controlling the transformation processes are the key benefits of data warehousing.

POOLED METADATA REPOSITORY

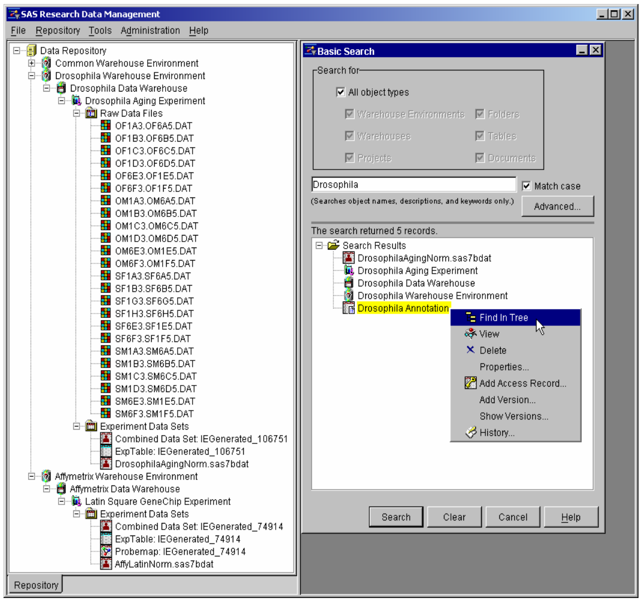
SAS RDM provides a platform for centralizing access and managing this discovery research data. At its core is the Pooled Metadata Repository (PMR), which provides a means of consolidating metadata from any number of data warehouses to create a single searchable repository. The PMR does not move the physical data, but rather gathers distributed metadata into one location

With the PMR, RDM extends the data warehousing concept to its scientific users by providing additional capabilities that enable compliance with FDA regulations, promote collaboration between departments/projects, organizations and/or locations, and allow the management of multiple object types used in discovery data. The RDM workspace surfaces the PMR to researchers in the form of searchable metadata. Figure 5 illustrates such a search on the data from the *Drosophila* example. You can search and download all information associated with a project or an experiment, including data, documents, and images.

In addition to the search and download capabilities, you can register new data to the PMR to share with colleagues via the Upload component. Upload functionality provides a way for you to add individual objects to the PMR one at a time. Figure 6 depicts upload of a PDF document related to the *Drosophila* experiment. Once you upload an object, it becomes available to anyone who has access to the PMR. You can add extended attributes to any object, which then become searchable metadata.

SECURITY MODEL

Among the regulatory compliance features is a security model that controls access to the system by requiring a user name and password. RDM controls access to warehouse objects as well as access to certain areas of the system at the user level or at a user-group level. Permissions range from no access to read-only access to full edit access. The security model also includes an audit trail of actions performed in the system (Figure 7) and versioning of all data loaded in the warehouse. The audit trail and version control simplify the ability to trace back the source of data and/or modifications made along the way.



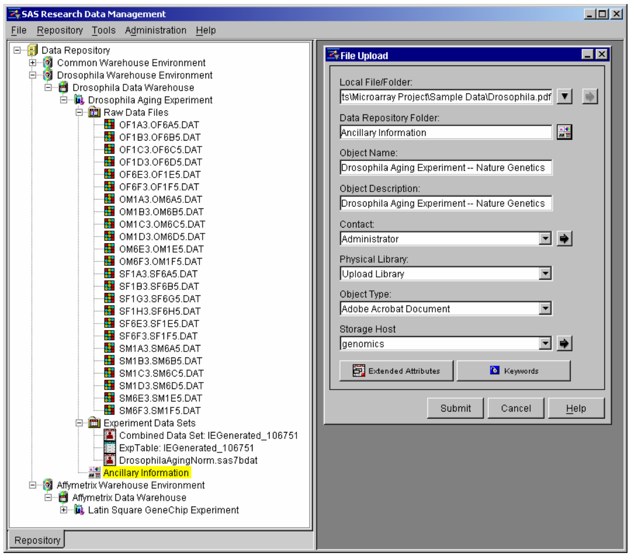
☑ All object types
☑ Warehouse Environments ☑ Folders
☑ Warehouses ☑ Tables
☑ Projects ☑ Documents

Drosophila ☑ Match case
(Searches object names, descriptions, and keywords only.)

The search returned 5 records.

Search Results
DrosophilaAgingNorm.sas7bdat
Drosophila Aging Experiment
Drosophila Data Warehouse
Drosophila Warehouse Environment
Drosophila Annotation

Figure 5: Perform a search on the Pooled Metadata Repository.

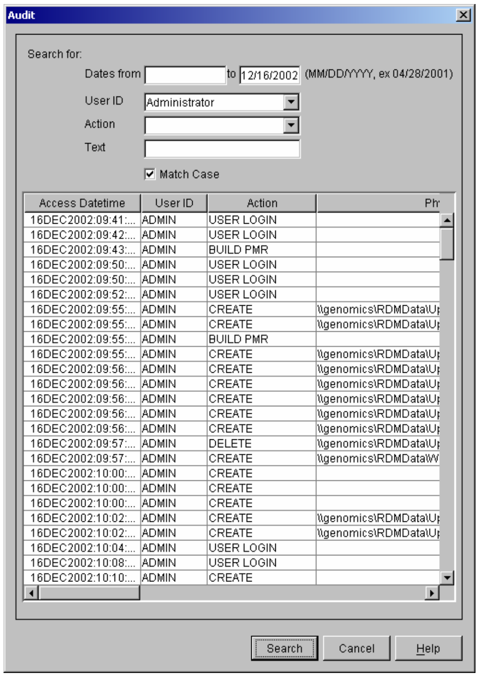


Local File/Folder: ts\Microarray Project\Sample Data\Drosophila.pdf
Data Repository Folder: Ancillary Information
Object Name: Drosophila Aging Experiment -- Nature Genetics
Object Description: Drosophila Aging Experiment -- Nature Genetics
Contact: Administrator
Physical Library: Upload Library
Object Type: Adobe Acrobat Document
Storage Host genomics

Figure 6: Upload a file to the Pooled Metadata Repository.

DATA MODELS

Built using add-ins to SAS/Warehouse Administrator, RDM enables you to implement virtually an unlimited number of data schemas. For microarray data such as those from our *Drosophila* example, the recently OMG-sanctioned MAGE-ML standard (http://www.mged.org) provides a useful foundation. You can use the SAS XML engine (Friebel, 2003) to import data from MAGE-ML. Systems engineers from SAS can assist you with this as a part of their warehousing pilot program.



Search for:
Dates from to 12/16/2002 (MM/DD/YYYY, ex 04/28/2001)
User ID Administrator
Action
Text
☑ Match Case

Access Datetime	User ID	Action	Ph
16DEC2002:09:41:...	ADMIN	USER LOGIN	
16DEC2002:09:42:...	ADMIN	USER LOGIN	
16DEC2002:09:43:...	ADMIN	BUILD PMR	
16DEC2002:09:50:...	ADMIN	USER LOGIN	
16DEC2002:09:50:...	ADMIN	USER LOGIN	
16DEC2002:09:52:...	ADMIN	USER LOGIN	
16DEC2002:09:55:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:55:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:55:...	ADMIN	BUILD PMR	
16DEC2002:09:55:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:56:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:56:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:56:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:56:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:56:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:09:57:...	ADMIN	DELETE	\\genomics\RDMData\Up
16DEC2002:09:57:...	ADMIN	CREATE	\\genomics\RDMData\W
16DEC2002:10:00:...	ADMIN	CREATE	
16DEC2002:10:00:...	ADMIN	CREATE	
16DEC2002:10:00:...	ADMIN	CREATE	
16DEC2002:10:02:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:10:02:...	ADMIN	CREATE	\\genomics\RDMData\Up
16DEC2002:10:04:...	ADMIN	USER LOGIN	
16DEC2002:10:08:...	ADMIN	USER LOGIN	
16DEC2002:10:10:...	ADMIN	CREATE	

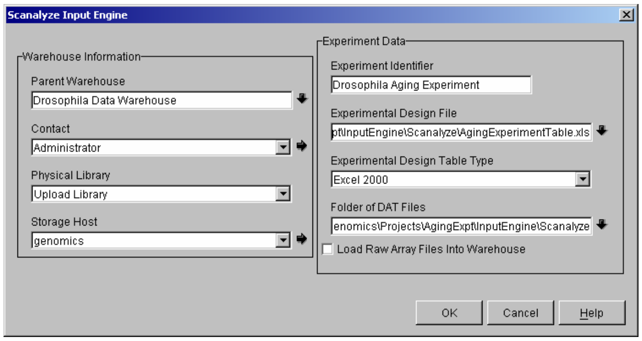
Figure 7: SAS Research Data Management Audit Trail

SAS MICROARRAY SOLUTION

SAS MAS is the first vertical deployment of the RDM platform, and integrates seamlessly with it to produce a whole product solution for managing, analyzing, and visualizing microarray data. The additional functionality of MAS includes input engines and analytical processes.

INPUT ENGINES

Input engines are software routines that pull data from instrumentation systems output and load the data into a SAS MAS warehouse. Input engines are specific to input data structures, and you can customize them to suit the vendors you select to perform your microarray experiments. Input engines enable you to directly import an entire collection of raw numerical array files with a few mouse clicks. Figure 8 shows the input engine for the *Drosophila* aging data, which were produced in the Gibson Lab (http://statgen.ncsu.edu/ggibson/) using the popular public domain image analysis software ScanAlyze (http://rana.lbl.gov/EisenSoftware.htm).



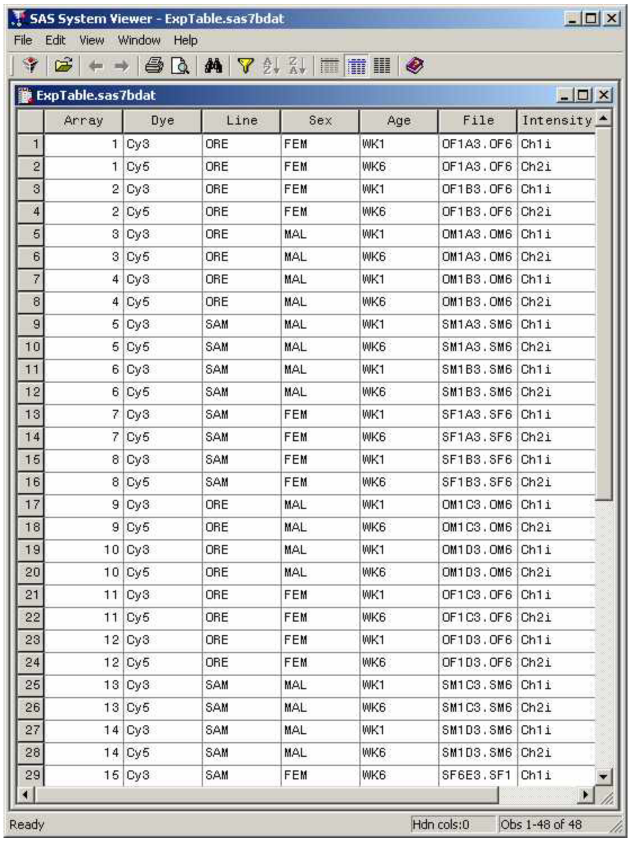
Parent Warehouse: Drosophila Data Warehouse
Contact: Administrator
Physical Library: Upload Library
Storage Host: genomics

Experiment Identifier: Drosophila Aging Experiment
Experimental Design File: pt\InputEngine\Scanalyze\AgingExperimentTable.xls
Experimental Design Table Type: Excel 2000
Folder of DAT Files: enomics\Projects\AgingExpt\InputEngine\Scanalyze
☐ Load Raw Array Files Into Warehouse

Figure 8: ScanAlyze Input Engine for *Drosophila* Aging Data

In addition to the raw data files, you must provide an experimental design file that shows how to map the experimental factors to the arrays. Figure 9 shows a portion of the experimental design table for the *Drosophila* experiment. Note there are two rows for every array corresponding to the Cy3 and

Cy5 channels. The three experimental factors are Line, Sex, and Age, and there are also variables indicating the name of the corresponding raw data file and the appropriate column within it. This kind of approach allows you to input designs of arbitrary complexity.


SAS System Viewer - ExpTable.sas7bdat
File Edit View Window Help
ExpTable.sas7bdat

	Array	Dye	Line	Sex	Age	File	Intensity
1	1	Cy3	ORE	FEM	WK1	OF1A3.OF6	Ch1i
2	1	Cy5	ORE	FEM	WK6	OF1A3.OF6	Ch2i
3	2	Cy3	ORE	FEM	WK1	OF1B3.OF6	Ch1i
4	2	Cy5	ORE	FEM	WK6	OF1B3.OF6	Ch2i
5	3	Cy3	ORE	MAL	WK1	OM1A3.OM6	Ch1i
6	3	Cy5	ORE	MAL	WK6	OM1A3.OM6	Ch2i
7	4	Cy3	ORE	MAL	WK1	OM1B3.OM6	Ch1i
8	4	Cy5	ORE	MAL	WK6	OM1B3.OM6	Ch2i
9	5	Cy3	SAM	MAL	WK1	SM1A3.SM6	Ch1i
10	5	Cy5	SAM	MAL	WK6	SM1A3.SM6	Ch2i
11	6	Cy3	SAM	MAL	WK1	SM1B3.SM6	Ch1i
12	6	Cy5	SAM	MAL	WK6	SM1B3.SM6	Ch2i
13	7	Cy3	SAM	FEM	WK1	SF1A3.SF6	Ch1i
14	7	Cy5	SAM	FEM	WK6	SF1A3.SF6	Ch2i
15	8	Cy3	SAM	FEM	WK1	SF1B3.SF6	Ch1i
16	8	Cy5	SAM	FEM	WK6	SF1B3.SF6	Ch2i
17	9	Cy3	ORE	MAL	WK1	OM1C3.OM6	Ch1i
18	9	Cy5	ORE	MAL	WK6	OM1C3.OM6	Ch2i
19	10	Cy3	ORE	MAL	WK1	OM1D3.OM6	Ch1i
20	10	Cy5	ORE	MAL	WK6	OM1D3.OM6	Ch2i
21	11	Cy3	ORE	FEM	WK1	OF1C3.OF6	Ch1i
22	11	Cy5	ORE	FEM	WK6	OF1C3.OF6	Ch2i
23	12	Cy3	ORE	FEM	WK1	OF1D3.OF6	Ch1i
24	12	Cy5	ORE	FEM	WK6	OF1D3.OF6	Ch2i
25	13	Cy3	SAM	MAL	WK1	SM1C3.SM6	Ch1i
26	13	Cy5	SAM	MAL	WK6	SM1C3.SM6	Ch2i
27	14	Cy3	SAM	MAL	WK1	SM1D3.SM6	Ch1i
28	14	Cy5	SAM	MAL	WK6	SM1D3.SM6	Ch2i
29	15	Cy3	SAM	FEM	WK6	SF6E3.SF1	Ch1i

Ready Hdn cols:0 Obs 1-48 of 48

Figure 9: Experimental Design Table for the *Drosophila* Example

A different input engine enables you to input annotation data for the genes on a particular chip, and you can use the resulting SAS dataset in any experiment for which it is appropriate.

ANALYTICAL PROCESSES

Analytical processes are special SAS macro programs that perform data manipulations and statistical calculations on the experimental data you have loaded into RDM. These processes employ the power of the SAS System to generate analysis data sets, listings, statistical results, and graphs. Analytical processes are reusable and flexible. They can range in functionality from very simple data displays to complex statistical modeling.

Scientists and statisticians alike can run analytical processes against any experiment dataset by simply providing appropriate values for input parameters. After you select an analytical process to run, a parameter input window requests information required for successful execution. The parameters values are specific to the dataset at hand, but you do not need to edit the code itself once you have written and loaded the analytical process into SAS MAS.

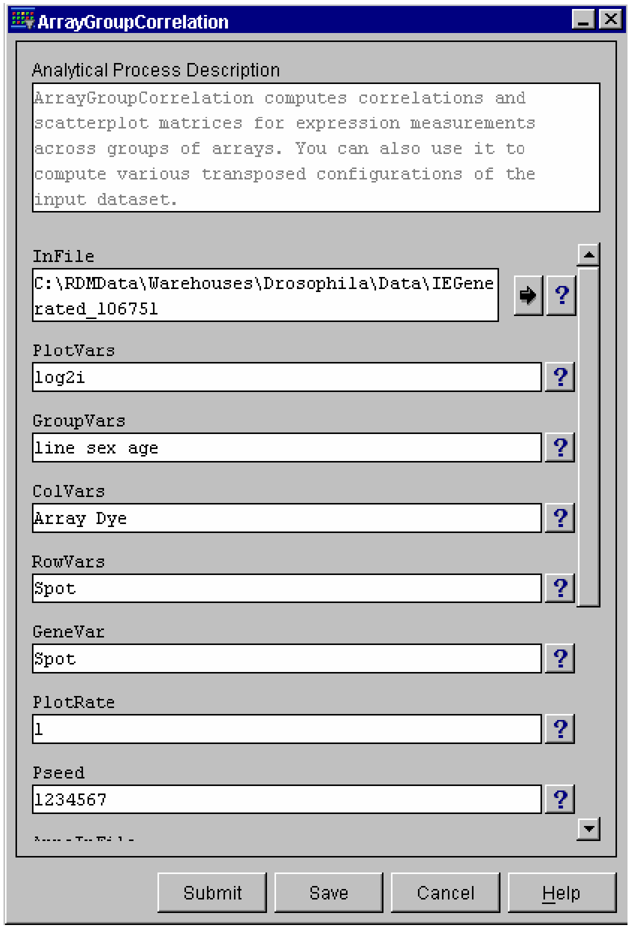
In its initial release, SAS Microarray Solution provides four analytical processes for use with your experiment data:

- **DataContents** displays the contents of a SAS dataset in HTML format.
- **ArrayGroupCorrelation** takes all of the array data from an experiment, divides it into groups selected by the user, and then performs a multivariate correlation analysis on each group.
- **MixedModelNormalization** normalizes microarray data by fitting a mixed linear model across all of the arrays in an experiment.
- **MixedModelAnalysis** provides a comprehensive look at results from fitting mixed models on a gene-by-gene basis.

ARRAY GROUP CORRELATION

Let's consider ArrayGroupCorrelation in more detail. This process accepts as input a SAS dataset in "tall skinny" format; that is, all of the raw array intensity measurements are stacked into one variable. (The default input engines build this kind of dataset for you.) It then enables you to group the observations into sets that will be plotted against each other in multivariate fashion.

Figure 10 displays the parameter input window for ArrayGroupCorrelation. MAS creates this window dynamically from specially configured SAS macro code. The *SAS Microarray Solution Analytical Process Programmer's Guide,* available only to SAS MAS customers, describes the requisite details. You must enter appropriate information in each field and then click Submit.



Analytical Process Description

ArrayGroupCorrelation 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 dataset.

InFile: C:\RDMData\Warehouses\Drosophila\Data\IEGenerated_106751

PlotVars: log2i

GroupVars: line sex age

ColVars: Array Dye

RowVars: Spot

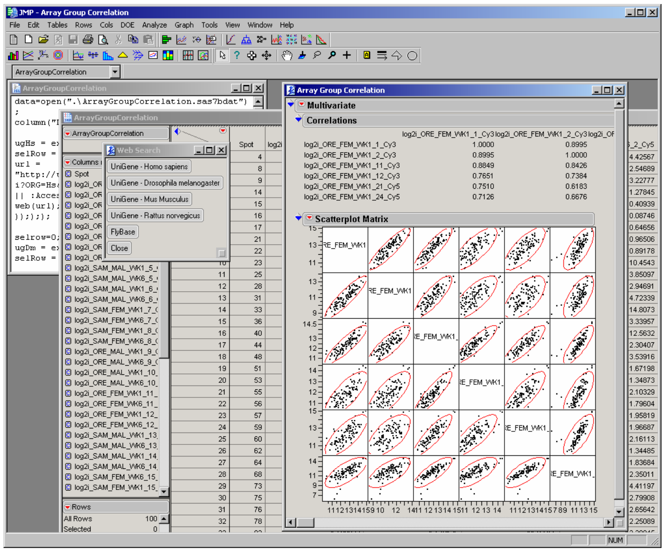
GeneVar: Spot

PlotRate: 1

Pseed: 1234567

Figure 10: Input Parameters for the ArrayGroupCorrelation Analytical Process

Upon submission, ArrayGroupCorrelation sends its macro code via SAS Integration Technologies to a preactivated SAS server, assigning each parameter value from the input window to its corresponding macro variable. The code performs some error checking and then calls Proc Transpose to put the data into the appropriate multivariate form. It then uses PUT statements to create a JMP Scripting Language (JSL) file. SAS MAS executes this file, producing output as in Figure 11. In addition to the standard Multivariate platform for each group, this particular script also creates a Web Search Dialog in JMP that allows you to automatically display the relevant UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) or FlyBase (http://flybase.bio.indiana.edu/) Web pages for all genes that are currently selected.


JMP - Array Group Correlation
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
ArrayGroupCorrelation
Web Search
UniGene - Homo sapiens
UniGene - Drosophila melanogaster
UniGene - Mus Musculus
UniGene - Rattus norvegicus
FlyBase
Close
Array Group Correlation
Multivariate
Correlations
Scatterplot Matrix

```
data=open(".\ArrayGroupCorrelation.sas7bdat")
```

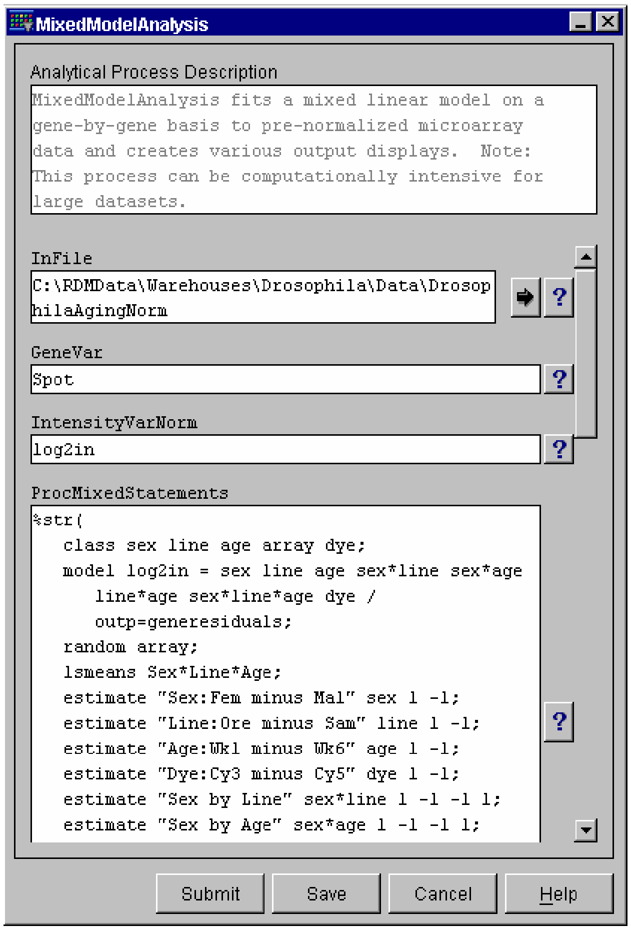
```
                          log2i_ORE_FEM_WK1_1_Cy3 log2i_ORE_FEM_WK1_2_Cy3
log2i_ORE_FEM_WK1_1_Cy3          1.0000                  0.8995
log2i_ORE_FEM_WK1_2_Cy3          0.8995                  1.0000
log2i_ORE_FEM_WK1_11_Cy3         0.8849                  0.8426
log2i_ORE_FEM_WK1_12_Cy3         0.7651                  0.7384
log2i_ORE_FEM_WK1_21_Cy5         0.7510                  0.6183
log2i_ORE_FEM_WK1_24_Cy5         0.7126                  0.6676
```

Figure 11: JMP results from ArrayGroupCorrelation

MIXED MODEL ANALYSIS

MixedModelAnalysis is a more complicated analytical process. It performs a high-level mixed model analysis of variance on pre-normalized array data. (MixedModelNormalization implements one basic kind of normalization that centers each channel to its geometric mean.) Figure 12 shows the top portion of the MixedModelAnalysis parameter input window.

The critical input parameter is a collection of Proc Mixed statements enclosed in the %str() macro to allow SAS punctuation. This code illustrates a full three-way factorial model adjusted by a main effect for dyes. The process fits this model to each gene separately using a BY statement with the GeneVar parameter value as the SAS variable. Array (equivalent to spot at the individual gene level) is considered to be a random effect, and accounts for the typically strong correlation observed between two measurements from the same spot. (Note that this model is on the normalized $\log_2$ intensities, not ratios.) The three-way interaction forms least-squares means, and a series of ESTIMATE statements test various one degree-of-freedom hypotheses of interest. Some references for this kind of approach are Deng et al. (2002), Gibson (2002), Jin et al. (2001), and Wolfinger et al. (2001).



Analytical Process Description

MixedModelAnalysis fits a mixed linear model on a gene-by-gene basis to pre-normalized microarray data and creates various output displays. Note: This process can be computationally intensive for large datasets.

InFile: C:\RDMData\Warehouses\Drosophila\Data\DrosophilaAgingNorm

GeneVar: Spot

IntensityVarNorm: log2in

ProcMixedStatements

```
%str(
   class sex line age array dye;
   model log2in = sex line age sex*line sex*age
      line*age sex*line*age dye /
      outp=generesiduals;
   random array;
   lsmeans Sex*Line*Age;
   estimate "Sex:Fem minus Mal" sex 1 -1;
   estimate "Line:Ore minus Sam" line 1 -1;
   estimate "Age:Wk1 minus Wk6" age 1 -1;
   estimate "Dye:Cy3 minus Cy5" dye 1 -1;
   estimate "Sex by Line" sex*line 1 -1 -1 1;
   estimate "Sex by Age" sex*age 1 -1 -1 1;
```

Figure 12: Parameters for MixedModelAnalysis Analytical Process

MixedModelAnalysis, like ArrayGroupCorrelation, executes on the SAS server and then creates a JSL file. Figure 13 shows some of the results for the *Drosophila* example. The upper-left window is a volcano plot, which graphs $\log_2$ fold change on the x-axis versus statistical significance, in terms of negative log *p*-value, on the y-axis. MixedModelAnalysis creates a separate volcano plot for each ESTIMATE statement you specify. The lower-left plot is a parallel coordinate plot of the lsmeans of all genes that pass a Bonferroni cutoff in at least one of the volcano plots. The bottom middle plot are the same lsmeans, but standardized to have mean zero and variance one. The unstandardized lsmeans are related to the x-axes of the volcano plots, whereas the standardized profiles are more closely related to the y-axes because the t-tests behind the negative log *p*-values are location-scale invariant. The upper-center graph plots the first two principal components of the standardized lsmeans of the significant genes.

The final display in Figure 13 is a two-way Wald hierarchical clustering analysis of the standardized lsmeans of the significant genes. Rows are genes and columns are lsmeans category. Although not legible, the left portion of this plot contains the annotation information for each gene. The colors in all of the displays are derived from this clustering analysis. The dynamic linking and interactivity of JMP make it ideal for intense exploration of these statistically curated results.

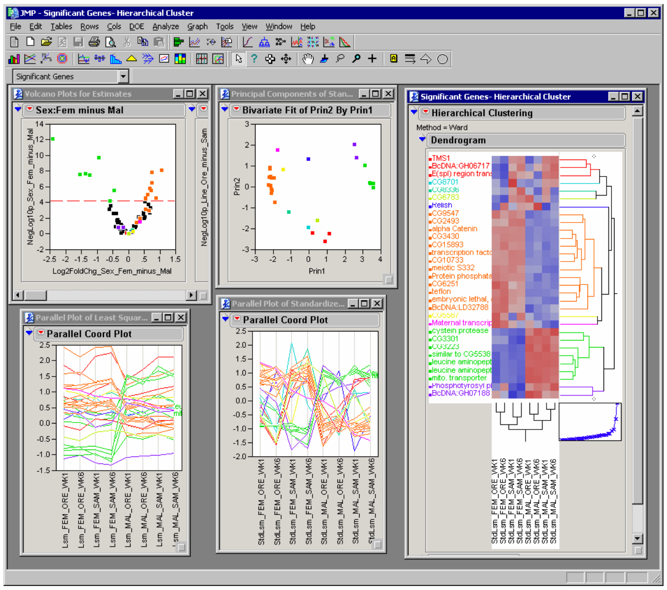

JMP - Significant Genes - Hierarchical Cluster
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Significant Genes
Volcano Plots for Estimates
Sex:Fem minus Mal
NegLog10p_Sex_Fem_minus_Mal
Log2FoldChg_Sex_Fem_minus_Mal
Principal Components of Stan...
Bivariate Fit of Prin2 By Prin1
Prin2
Prin1
Significant Genes - Hierarchical Cluster
Hierarchical Clustering
Method = Ward
Dendrogram
Parallel Plot of Least Squar...
Parallel Coord Plot
Parallel Plot of Standardize...
Parallel Coord Plot

Figure 13: Results of Mixed Model Analysis in JMP

The other two prewritten analytical processes, DataContents and MixedModelNormalization, do not generate output in JMP. Instead, they create HTML output via the Output Delivery System that you can view in a browser. Not only are these four processes useful for array analysis, but they provide you with templates for writing your own processes.

NEW DEVELOPMENT DIRECTIONS

TRANSCRIPTOMICS

The preceding examples represent a small fraction of the capabilities SAS offers you for analyzing microarray and related transcriptomic data. RDM and MAS provide a framework that enables you to tap into the full power of the SAS System in order to make the most of your array data. Consider the following capabilities, all of which you can utilize via MAS Analytical Processes:

- The SAS Data Step, SAS Macro Language, and JMP Scripting Language are very flexible and extensive environments for manipulating and processing array data.
- Proc Optex, Proc Factex, Proc Plan, the ADX Menu System, and the DOE features in JMP help you to create optimal designs.
- Over half of the 50+ procedures in SAS/Stat are applicable to array data. These include procedures for analysis of variance, clustering, density estimation, discriminant analysis, multidimensional scaling, multiple comparisons, nonparametric statistics, partial least squares, power and sample size, principal components, and smoothing. (Rodriguez 2003). New residual diagnostics will be available in Release 9.1 of Proc Mixed, enabling better quality control of your array data. Procedures from SAS/QC, SAS/OR, SAS/ETS, and SAS/IML (including IML Workshop) can also be useful. Scores of examples from other scientific disciplines are in the SAS Sample Library and at www.sas.com/techsup/download/stat .
- You can use RDM and MAS to preprocess and create analysis-ready data sets for Enterprise Miner. Its incredibly powerful collection of data mining methodologies and intuitive process-flow interface enable you to efficiently generate a wide range of cross-validated predictions. Furthermore, you can use Text Miner to process collections of scientific abstracts you create from important gene lists.
- SAS/MP-Connect enables parallelization of suitable methods across multiple CPUs. Refer to Doninger et al. (2003) for a recent microarray example.

BIOINFORMATICS

While SAS's core strengths remain in warehousing and enhanced analytics, we are developing new tools in association with RDM and MAS for such areas as motif detection, pathway overlays, and enhanced annotation from public databases.

PROTEOMICS

Proteomics is an area of active research and development. Neville et al. (2002) describe some initial analyses of mass spectrometry data using both a data mining approach and a more informal method based on Welch t-tests. For the latter, Figure 14 depicts overlaid blood serum spectra from a group of normal and diseased patients, along with a negative log10-pvalue spectrum indicating significant "hot spots" in the spectra above it. Significant group-wide differences could translate into critical diagnostic biomarkers after suitable validation.

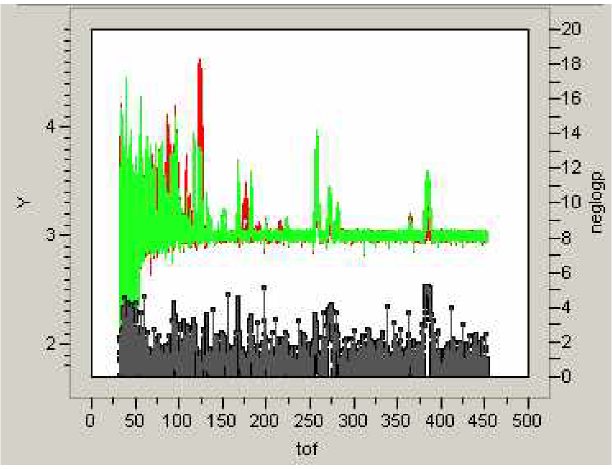


Figure 14: Log spectra (Y) of diseased (red) and normal (green) blood sera, along with negative log10 p-values (gray) testing for significant differences between the groups at each time-of-flight (tof) value.

GENETICS

This paper ends where it started, with a consideration of DNA. SAS/Genetics, introduced last year, is a standalone product containing procedures for the statistical processing of DNA marker data (Czika et al., 2001). New enhancements for Release 9.1 include the following:

- Two new procedures are available: HTSNP (for finding haplotype tagging single-nucleotide polymorphisms) and INBREED (copied from SAS/STAT).
- The GENOCOL and DELIMITER= options allow you to input marker data with single variables instead of two variables per maker.
- The ALLELE procedure now has the ALLELEMIN=, GENOMIN=, and HAPLOMIN= options for specifying minimum estimated frequencies.
- The CASECONTROL procedure has the new NULLSNPS= option for specifying SNPS for genomic control and the PERMS= option for computing permutation-based exact *p*-value approximations.
- The FAMILY procedure has a new "Family Summary" table valuable for checking for inconsistencies, a PERMS= option like that in CASECONTROL, and multiallelic SDT and

combined SDT/TDT options.

- The HAPLOTYPE procedure has a new stepwise EM algorithm.
- The PSMOOTH procedure has new TPM and TAU= options for implementing the truncated product method.

Details about these and other new features are available from SAS technical support. We are also working on an integrated genetic markers solution, similar in design to SAS Microarray, that incorporates macros derived from SAS/Genetics in place of the microarray best practices.

CONCLUSION

Rich scientific data from microtechnological instrumentation and SAS software are a formidable combination. Regardless of where along the central dogma your interests lie, SAS Scientific Discovery Solutions equip you with the power to intelligently manage and analyze your data with unparalleled proficiency.

SAS RDM combines the power of data warehousing with an easy-to-use interface for search and retrieval. Its warehousing technologies can manage many different types of data, regardless of their formats or type. With SAS RDM, you can register genetic, transcriptomic, protein and related data files to the warehouse and view them using their native applications. SAS RDM increases the productivity and effectiveness of your discovery organization by streamlining data management and increasing data accessibility.

SAS Microarray builds on the strong RDM foundation and enables you to tap into the power of the entire SAS System with its rich data manipulation and analytical capabilities. It is designed for both statisticians and scientists to work in a collaborative format. Statisticians can guide the scientist on experimental design and write appropriate analytical processes. They can then register these processes into the solution where scientists can easily access, use, and reuse them with no additional SAS programming. This enables both the scientist and the statistician to spend more time doing high knowledge tasks that require their expertise.

Of course software is no substitute for painstaking research, careful thought, and effective personal collaboration with colleagues. It can, however, significantly enrich these activities, and it has become indispensable as genomics information volume and density increase. We look forward to helping you overcome the inherent challenges and achieve incredible new insights in scientific discovery.

REFERENCES

Chu, T.-M., Weir, B., and Wolfinger, R.D. (2002a), A systematic statistical linear modeling approach to oligonucleotide array data analysis, *Mathematical Biosciences,* 176, 35-51.

Chu, T.-M., Weir, B., and Wolfinger, R.D. (2002b), Comparison of Li-Wong and mixed model approaches to oligonucleotide array data analysis, in review.

Czika, W., Yu, X., and Wolfinger, R.D. (2002), Genetic data analysis using SAS/Genetics, *SUGI 27 Proceedings*, SAS Institute, Inc., Cary, NC.

Deng, S., Chu, T.-M., and Wolfinger, R.D. (2002), Transcriptome variability in the normal mouse, manuscript to be published in the CAMDA proceedings, Duke University.

Doninger et al. (2003), Developing Client/Server Applications to Maximize V9 Parallel Capabilities*,* *SUGI 28 Proceedings*, SAS Institute, Inc., Cary, NC.

Friebel (2003), XML? We do that! *SUGI 28 Proceedings*, SAS Institute, Inc., Cary, NC.

Gibson, G. (2002), MMANMADA Tutorial, http://statgen.ncsu.edu/ggibson/Pubs.htm .

Gibson, G. and Muse, S. (2001), *A Primer of Genomic Science,* Sinhauer.

Hsieh, W.-P., Chu, T.-M., and Wolfinger, R.D. (2002), Who are those strangers in the Latin square? manuscript to be published in CAMDA proceedings, Duke University.

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.

Kerr and Churchill (2001), Experimental design for gene expression microarrays, *Biostatistics*, 2:183-201. http://www.jax.org/research/churchill/pubs/index.html

Neville, P., Tan, P.-Y., Mann, G., and Wolfinger, R.D. (2002), Generalizable mass spec mining and mapping, manuscript to be published in proceedings of First Annual Proteomics Conference, Duke University.

Rodriguez, R. (2003), SAS/STAT Version 9: Progressing into the Future, *SUGI 28 Proceedings*, SAS Institute, Inc., Cary, NC.

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.

ACKNOWLEDGMENTS

The SAS Genomics R&D Team: Carey Carpenter, Tzu-Ming Chu, Wendy Czika, Samuel Johnson, Geoffrey Mann, Chandrika Nayak, Kristen Quinlan, Russ Wolfinger, Xiang Yu, and Jun Zhang. RDM and MAS Testers: Shibing Deng and Kelly Graham. SAS/Genetics Testers: Gerardo Hurtado and Jack Berry.

It is infeasible to list here the scores of folks who have helped make SAS SDS and this paper a reality, although we would like to thank the following individuals for their singular contributions: Darrell Barton, Kurt Brumbaugh, Michael Campa, Bob Carpenter, Julia Carpenter, Rob Carscadden, Joe Carter; Gary Churchill, Virginia Clark, Glen Dalrymple, Cheryl Doninger, Alan Eaton, Rick Evans; Michael Fitzgerald, Tony Friebel, Jill Fitzgibbons, Greg Gibson, Russell, Gonsalves, Gerhard Held, Bob Hickey, Ralph Hollinshead, James Holman, Hans Meyer, Joe Mudd, Meltem Narter, Padraic Neville, Kay Obenshain, Ned Patz, Jon Pennino, Bob Rodriguez, John Sall, Pei-Yi Tan, Jill Tao, Randy Tobias, Len van Zyl, Bruce Weir, Fred Wright, and Liz Wolfinger.

CONTACT INFORMATION

We value and encourage your comments and questions. Please contact Susan Flood at:

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