This is to certify that the
dissertation entitled

DESIGN, MANAGEMENT, AND QUALITY CONTROL OF TOXICOGENOMIC EXPERIMENTS

presented by

Lyle David Burgoon

has been accepted towards fulfillment
of the requirements for the

Doctoral degree in Pharmacology & Toxicology

Major Professor's Signature

6/21/05

Date

DESIGN, MANAGEMENT, AND QUALITY CONTROL OF TOXICOGENOMIC EXPERIMENTS

By

Lyle David Burgoon

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Pharmacology and Toxicology

2005

ABSTRACT

DESIGN, MANAGEMENT, AND QUALITY CONTROL OF TOXICOGENOMIC EXPERIMENTS

By

Lyle David Burgoon

High throughput "omic" technologies, such as the cDNA microarray, have the potential of increasing mechanistic understanding of the biological underpinnings related to a biological outcome, enhancing safety assessments during the development of a new chemical entities, identification of new druggable targets, selection of patient candidates for therapeutic treatment, and for monitoring exposures to hazardous chemicals through biomarkers. However, for these potentials to be realized, investigators must ensure their experiments are properly designed with respect to their intended purpose, the data is appropriately managed to decrease human error, and prevent loss of data, and that the data are of sufficient quality to ensure the results are appropriate. To address these needs, the dbZach System, a database and associated computational applications, has been developed to manage data derived from toxicogenomic and pharmacogenomics experiments. Using historical data within the laboratory, a quality control protocol was developed, consisting of three different divisions. The first division uses a trained support vector machine (SVM), a statistical learning theory method, for identifying high and low quality arrays based on global intensity characteristics. The second division uses a semiparametric normalization method for identifying misaligned subgrids on the microarray, to ensure proper feature alignment and quantification. The third division utilizes boxplots to identify arrays with incongruent distributions, and line plots to identify trends with regards to the number of identified and saturated features. Using

data within dbZach, three temporal experimental designs were compared: the independent reference, loop, and modified loop designs. By comparing the results from these experiments based on the amount of experimental error, identifying temporal confounds, and analyzing differences in the temporal clustering relationships, the modified loop design was judged the most appropriate design. However, when economic considerations are made, the loop design may be preferred when used with a larger number of biological replicates.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank all of the members of the Zacharewski laboratory, especially Dr. Tim Zacharewski for the opportunity to work within the laboratory, and his guidance throughout the execution of these projects; my guidance committee for their patience, assistance, and knowledge; the funding agencies that have made this work possible, including NIEHS and EPA; our collaborators on these and other projects within the laboratory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Efforts to understand the biological effects of chemical exposures, both therapeutic and toxic, singularly and in mixtures, are beginning to take a global, or systems level scope, ushering in the disciplines of systems biology and systems toxicology. The systems approach attempts to identify critical nodes in biological information networks (e.g., a ligand binds to a receptor, eliciting changes in gene expression, and perturbations in protein expression and activation) by comparing information flows between normal and chemically exposed or diseased samples (Hood and Perlmutter, 2004). Systems toxicology is a toxicology specific rendering of the systems biology idea; integrating data from toxicogenomic studies with more traditional toxicology measures (e.g., clinical chemistry and histopathology) (Waters and Fostel, 2004). Toxicogenomics is the combined study of the genomic, the proteomic (i.e., all of the proteins expressed within a system) (Wetmore and Merrick, 2004), and the metabolomic (i.e., the full complement of metabolites within a system) (Nicholson and Wilson, 2003) response to chemical exposures and environmental stressors (Nuwaysir, *et al.*, 1999; Waters and Fostel, 2004; Waters, *et al.*, 2003b).

"Omic" technologies are believed to be necessary to engender systems toxicology as they are global technologies – measuring the expression of thousands to near complete sets of biological molecules of interest. By measuring large segments, if not the entire complement, of molecules within a class (genome, proteome, metabonome), it may be possible to study many of the interconnected biological networks, and the flow of information/data from source to outcome; the essence of the systems biology (Ideker, *et al.*, 2001) and systems toxicology (Waters and Fostel, 2004) frameworks.

Potential of Genomics in Pharmacology & Toxicology

Microarrays provide the ability to simultaneously monitor the expression of thousands of genes during disease progression or onset and in response to chemical or therapeutic exposures. These abilities have fostered the growth and maturation of the fields of pharmacogenomics (Weinshilboum and Wang, 2004) and toxicogenomics, and may serve to revitalize the pharmaceutical industry (Lindsay, 2003).

Pharmacogenomics

Although pharmacogenomics is typically defined in the translational sense, as a broadening of pharmacogenetics, where patient populations are screened to identify sensitive and resistant groups that may not respond predictably to a therapeutic treatment, based on global gene expression (Weinshilboum and Wang, 2004), pharmacogenomics can also be thought of as a means to identify novel drug targets (Lindsay, 2003; Ross, *et al.*, 2004).

For example, gene expression studies have been used to classify cancer biopsies and tumor samples (Glas, *et al.*, 2005; Golub, *et al.*, 1999; Kristensen, *et al.*, 2005; Mischel, *et al.*, 2004; Selvanayagam, *et al.*, 2004; Yeang, *et al.*, 2001) to improve clinical outcomes by identifying the correct tumor class and treatment regimen. The ultimate goal of these investigations is to create new diagnostic tests for the assessment of tumors, and tailor treatment programs specific to the patient (i.e., personalized medicine) to reduce toxicity and increase the therapeutic potential (Ross, *et al.*, 2004). However, molecular profiles from these same experiments can be used to identify new drug targets specific to the particular class of cancer under investigation. The idea is that since the

tumor samples can be differentiated from normal tissues based on gene expression profiles, there must be drug targets embedded within these profiles that can be exploited (Lindsay, 2003).

Computational Toxicology

Computational toxicology is the application of computer science, mathematics, statistics and information technology to the field of toxicology. Computational toxicology encompasses physiologically based pharmacokinetic modeling, dose-response modeling, and analysis of data from toxicogenomic studies.

Toxicology data can be thought of as existing within a source-to-outcome, or source-to-disease, continuum (Kavlock, *et al.*, 2003; Waters and Fostel, 2004). Several forces act upon this continuum to modulate the net outcome, including the pharmacokinetics and efficacy of the compound, the cellular and systems response to the exposure, and interactions between these levels. For example, injury may commence at an early time point, parallel to the expression of genes that encode drug efflux and metabolism proteins. The activation of these compensatory mechanisms may lead to increased excretion of the chemical.

By combining the source-to-outcome toxicology data together, in a database system, a new kind of data mining activity can emerge: toxicological intelligence gathering. Similar to business intelligence, where businesses integrate data from across the business spectrum, combining customer information with product flows and other business indicators, to generate patterns predictive of the business process; toxicologic intelligence gathering integrates data from across the toxicology spectrum, combining different types of data from chemical exposures, to identify patterns in gene, protein and

metabolite expression, pathology, and gross observations to be predictive of a toxicological process and mechanism of action.

Toxicologic intelligence itself provides the infrastructure to perform further computational toxicology experiments. These other computational toxicology efforts include development of algorithms for data normalization and analysis, pattern recognition, and correlation across experiments and experimental types (e.g., correlation of gene expression and metabolite expression data).

Engendering the Toxicogenomics Paradigm

Toxicogenomic studies generate a wealth of disparate data. Consider a complete toxicogenomic investigation for a new chemical entity will include data from genomic, proteomic, and metabonomic experiments, histopathological analysis, clinical chemistry, and gross observations. Although each set of data individually may be useful for understanding the biological effects following exposure, their integration would be more useful for the development of mechanistic understanding and systems toxicology models. For example, phenotypic anchoring of uterine gene expression changes to histological changes yield a better mechanistic understanding than any of these pieces of data alone (Moggs, *et al.*, 2004; Paules, 2003).

Toxicogenomics offers a wealth of potential with respect to safety assessment and mechanistic investigations, such as making drug development more efficient by identifying toxic drugs earlier in the development process (Ulrich and Friend, 2002), and facilitating mechanistic research (Boverhof, *et al.*, 2004; Luyendyk, *et al.*, 2004; Nuwaysir, *et al.*, 1999; Waters and Fostel, 2004). However, to fully realize its potential, several key ingredients must be present: 1) a data management solution, 2) appropriate

experimental designs, 3) high quality data, and 4) multivariate data analysis methods. Each of these will be reviewed in more detail below.

Toxicogenomics Data Management

Two methods for toxicogenomics data management currently exist: 1) use of flat files (i.e., spreadsheets, tab-delimited text files, etc), and 2) use of databases. Whereas a series of flat files may work for smaller projects, they discourage comparisons across studies, and become error-prone when performing complex data filtering tasks. Databases, however, are conceptually a series of flat files which facilitate cross-study comparisons. By serving as a central storage point, databases also support software development, preventing changes to the software to accommodate different flat file formats.

Databases are conceptually made up of two parts (from lowest to highest level): tables and subsystems (an example is shown in Figure 1-1). Tables are collections of records, or rows, where the data reside. All tables must contain a unique identifier for each record in the table called the primary key. For example, a table managing animal data would contain the age, sex, species, and strain, in addition to the database assigned unique primary key. In order to prevent tables from becoming too large, and to prevent data redundancy, more tables can be created. Tables can be related to one-another through a series of primary key-foreign key relationships, where a foreign key is a primary key entry from a foreign table that exists as part of a record within a table. For example, in Figure 1-1 a CAGE_ID foreign key (a primary key from the CAGE table) would exist within the ANIMAL table as part of a record for an individual animal so that there is knowledge of what cage was used for that animal. This allows cage-specific data to be contained in the cage table, separately from the animal. If all of the cage-specific

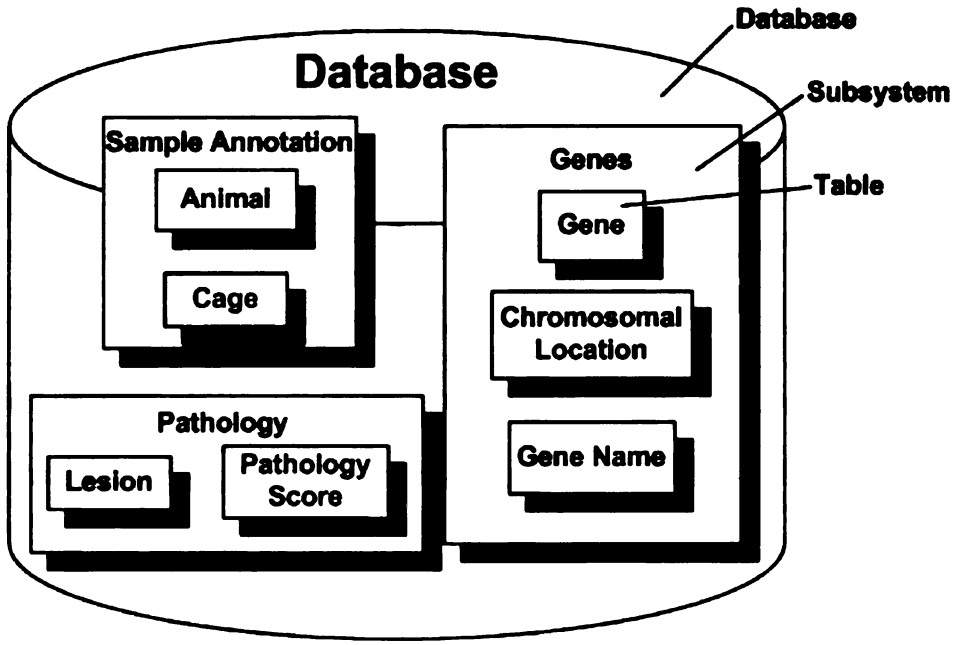

Figure 1-1: Anatomy of a Database. A database consists of two conceptual parts: the tables and the subsystems. Tables contain records, which can be thought of as rows. A series of related tables are grouped together into subsystems. A group of subsystems make up a database.

data were also contained within the animal table there would be the possibility for data redundancy, where the same data were entered repeatedly to describe the same cage conditions, increasing the likelihood of data entry errors.

A collection of tables that describe the same larger concept are placed within the same subsystem. For example, all of the data describing pathology data would be contained within the Pathology Subsystem. When subsystems are associated with particular technologies they serve to keep the database modular, ensuring the database remains scalable (i.e., can continue to grow as new technologies are developed).

Quality Assurance

Quality assurance methods are practices that ensure the high quality of some data or process. The simplest quality assurance mechanism utilized within most laboratories

is the use of standard operating procedures (SOPs) which ensures that the experiment, and hopefully the data, are reproducible. Generally, if data are reproducible, it is considered to be of high quality (Grant, *et al.*, 2003). However, reproducibility alone does not determine quality, since it is possible to have a highly reproducible low quality result. For instance, consistently extracting degraded total RNA from a sample is not a high quality result, regardless of how reproducible it is. Thus, use of variance *prima facie* for quality determination necessitates the presupposition that the process is of high quality, a generally inappropriate assumption.

Quality control tests are methods for monitoring data quality by using a quality assurance plan. Many quality control methods exist, such as control charts, and statistical testing methods. These methods identify samples that are beyond some variance-based threshold; identifying samples that lie outside the distribution of high quality samples with some given confidence (*NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 4-5-04). Examples of these techniques include the Shewhart plot and the Student's t-test.

The Shewhart plot graphically depicts trends in a process across time, coupled with variance-based quality thresholds. Once the process exits high quality, and breaks the variance-based threshold, it is said to be "out of control." The Student's t-test is used to determine whether samples come from the same distribution. For example, the one-sample t-test would compare measurements made from products generated by a process (i.e., microarrays) against a high quality standard, and determine, with some confidence, if they come from the same distribution.

The advent of multivariate classification methods based on supervised pattern recognition techniques have also been applied to the quality control problem. These methods require the use of a high, and occasionally low, quality dataset to either train a mathematical model, or to facilitate investigator comparison while visualizing the data. These methods include the use of Principal Components Analysis (PCA), k-means clustering, and Support Vector Machines (SVM).

PCA seeks to reduce the dimensionality of the data, from *n* dimensions to at most *n*-1 dimensions. The dimensions (or principal components) from PCA are generated such that the first dimension contains the largest proportion of the variance from the dataset, while the subsequent dimensions each represent the largest portion of the residual variance while remaining orthogonal to the previous dimension. Thus, the first three dimensions from the PCA represent the three dimensions that best explain the most variance. By visualizing the first two or three principal components within a coordinate plane or a three dimensional (3-D) space, it is possible to identify similar samples. In the case of quality control, if the test samples are analyzed concurrently with the high quality data, it is possible to perform quality control analysis by defining the high quality region or sphere based on the Hotelling's T-squared distance (Model, *et al.*, 2002); a multivariate t-test. Thus, any samples that exist beyond the Hotelling's T-squared quality threshold are demonstrated to be of low quality.

Supervised pattern recognition methods identify an investigator specified number of clusters from a dataset. For example, in assessing high quality and low quality within a data set, two clusters would be specified. Support Vector Machines and k-means clustering are both supervised methods, but work quite differently. Whereas the SVM

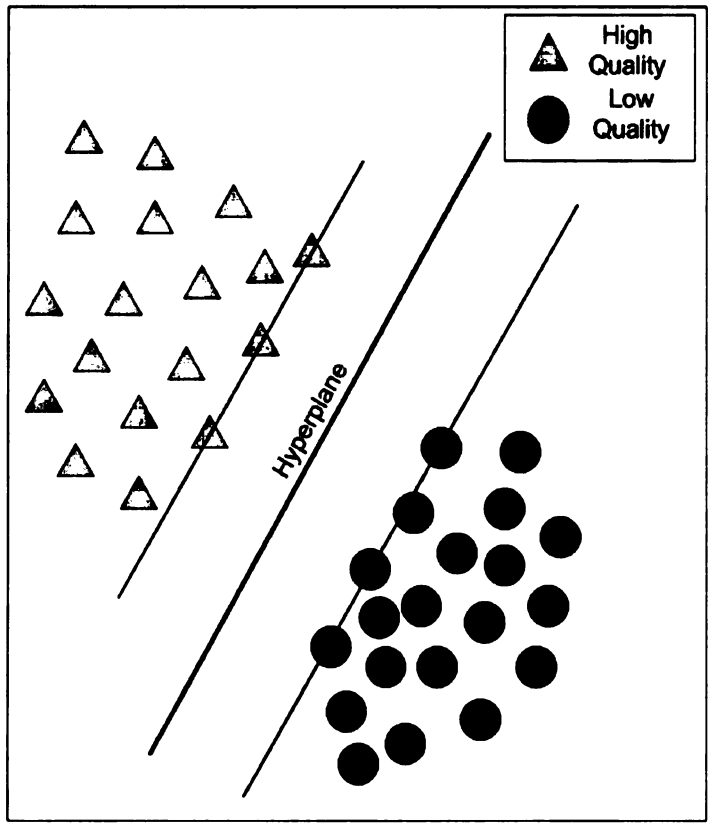

Figure 1-2: Support Vector Machine. The SVM algorithm projects the training data in a high dimensional space to better expand the differences between the classes (e.g., tumor vs normal, high vs low quality), and generate more optimal natural clusters. The algorithm then projects an optimal hyperplane through the data space that best discriminates the two classes. The mathematical model used to classify data following this training step represents this hyperplane in the high dimensional space.

creates a mathematical model trained on a defined historical dataset, the k-means clustering algorithm generates a cluster center based on clustering of the high and low quality historical datasets to determine the optimal pattern that best describe the high and low quality datasets.

The k-means algorithm requires that the cluster centers be specified along with the number of clusters. The cluster centers are used by the algorithm as "ground truth", meaning the cluster centers represent the ideal sample for that cluster. The algorithm

uses the cluster center as a point of comparison, where the sample is clustered with the center it is least far from, meaning the center that has the pattern most closely resembling the sample's pattern. In a typical quality control implementation of the k-means algorithm, the historical datasets will be clustered first, specifying that two clusters should be identified from this dataset. Once the clusters are defined, the cluster centers are extracted from the method. These centers will be supplied in the future when performing the k-means algorithm to cluster the samples.

The SVM method generates a mathematical model based on training data to classify data. The SVM is a binary classifier, but is capable of being extended to provide multi-class classification. Whereas the goal of the PCA is to reduce the dimensionality of the data, the SVM actually increases the dimensionality of the data (Duda, *et al.*, 2001). By increasing the dimensionality, natural clusters will form with larger interclass differences, making the classification problem easier (Figure 1-2). A hyperplane is projected within space, that separates the populations, and this represents the mathematical model. Classification of samples occurs by identifying their location with respect to the hyperplane (i.e., they will either exist on the high quality or the low quality side of the hyperplane). For quality assessment, samples from the high and low quality historical datasets are used to train the model.

Experimental Design

The experimental design utilized for an experiment holds immense gravity over the results and subsequent interpretation. Different designs yield varying degrees of precision and power for statistical inference, with varying degrees of cost (Fisher, 1962).

Although this provides the impetus for comparison of experimental designs, the choice of design is still at the discretion of the investigator.

When designing experiments, several factors must be considered and balanced including 1) the goal of the experimental manipulation, 2) the sources of variance and their interrelationships, and 3) the economics, both monetary and time, of the study. In some cases, it is possible that failure to balance these factors may result in less than optimal results, and in extreme cases the design may compromise the results to the point where they are inappropriate.

Tables 1-1 and 1-2 provide examples of two different experimental designs that could be employed for examining the effects of a chemical (Treatment = 1) in comparison to its vehicle control (Treatment = 0). Due to the logistics of the experiment, it must be performed across four days (Day = 1-4). A total of eight animals are used in these studies, four in the treated and four in the control groups. The only difference between these two designs is the arrangement of treatments with respect to the day.

One of the primary goals of an appropriate experimental design is to limit the number of confounding variables that may either complicate or negate the ability of the investigator to perform comparisons. A confounding variable is an explanatory variable that may interact with another variable. For example, in Table 1-1 there are a total of eight animals that will receive one of two treatments (0 = vehicle; 1 = treated) on two separate days. The null hypothesis being tested is that the treatment will have no effect, H_0: $\mu_0 = \mu_1$, where μ_i is the effect due to treatment (i = 0, 1). If it were expected that there were a significant and additive effect due to the day, d_j (j = 1, 2, 3, 4), then the experimental design used in Table 1-1 would exhibit a confound with respect to day (i.e.,

Table 1-1: Non-blocked Experimental Design with Day Confound

Animal	Day	Treatment	Expected Response*
1	1	0	$\mu_0 + d_1$
2	1	0	$\mu_0 + d_1$
3	2	0	$\mu_0 + d_2$
4	2	0	$\mu_0 + d_2$
5	3	1	$\mu_1 + d_3$
6	3	1	$\mu_1 + d_3$
7	4	1	$\mu_1 + d_4$
8	4	1	$\mu_1 + d_4$

day is a confounding variable on the treatment effect). To calculate the treatment effect, $\theta = \mu_0 - \mu_1$, between days 1 and 3:

$$(\mu_0 + d_1) - (\mu_1 + d_3) = (\mu_0 - \mu_1) + (d_1 - d_3).$$

This calculation becomes more complicated when treatments are considered across the entire design. It also demonstrates that the treatment variance cannot be distinguished from the variance due to the day.

The experimental design in Table 1-1 results in an n = 2 for each treatment. This is due to the fact that each animal within a day receives the same treatment. Thus, the experimental unit (i.e., the base treatment unit) is really the day, and not the animal. This is due to the confound within the experimental design, where the variance due to treatment and day are inseparable.

A more appropriate design for this experiment is listed in Table 1-2. Here, there are a total of four observations per treatment, instead of the two observations per treatment in Table 1-1. Day is no longer a confound within this design as each treatment variety (or treatment level) is represented within each day. This allows for the separation of the variance due to day and treatment. Furthermore, to calculate the treatment effect, $\theta = \mu_0 - \mu_1$, within a day:

Table 1-2: Complete Block Experimental Design

Animal	Day	Treatment	Expected Response*
1	1	0	$\mu_0 + d_1$
2	1	1	$\mu_1 + d_1$
3	2	0	$\mu_0 + d_2$
4	2	1	$\mu_1 + d_2$
5	3	0	$\mu_0 + d_3$
6	3	1	$\mu_1 + d_3$
7	4	0	$\mu_0 + d_4$
8	4	1	$\mu_1 + d_4$

$$(\mu_0 + d_1) - (\mu_1 + d_1) = (\mu_0 - \mu_1) - (d_1 - d_1) = (\mu_0 - \mu_1).$$

Similar mathematics would be used to calculate the treatment effect across the entire experiment. Thus, it is clear that by having each treatment variety present within each day, the variance due to the day variable can be factored out of the treatment effect. Also, as this design results in a larger *n* for the same number of animals, the statistical power will increase, meaning smaller changes due to treatment may be found significant at the same false positive rate.

In this example, the outcome being tested (i.e., the effect of treatment vs vehicle) and the economics of the experiment are the same. Both experiments require the same number of animals, the same number of days to complete the experiment, and the same amount of the treatment compound and vehicle. They only differ in the assignment of animals to treatments and days. However, the impact of the differences between the designs is great, with one experiment yielding results where the treatment is confounded by the day of treatment, whereas the other more accurately reflects the treatment effect. When a microarray experiment is considered, the number of possible confounding factors increases beyond those examined here.

Microarray Experimental Designs

Recently, the identification of appropriate microarray experimental designs has become a subject of great interest (Churchill, 2002; Dobbin, *et al.*, 2003; Dobbin and Simon, 2002; Simon, *et al.*, 2002; Tempelman, 2005; Townsend, 2003; Vinciotti, *et al.*, 2004). The experimental design used in a microarray experiment must include considerations of whether or not biological samples should be pooled, the number and types of technical replication, the arrangement of samples on the microarrays, and the analysis method. The choice of experimental design will impact the active list of genes, the hypotheses formulated from these lists, and ultimately the concept of the risk versus benefit from exposure to a chemical.

Table 1-3 lists many of the possible sources of variance in a microarray experiment. These sources can be categorized as either biological (i.e., inherent within the biological model) or technical. Generally speaking, it is best to block the sources of variance, such as in Table 1-2, such that each level of the treatment variable exists within every level of the confounding variables. For example, a cage-level confound that exists at the time of treatment, will persist through the entire experiment. Another common example of a confound is the assignment of different investigators to extract total RNA from tissue samples based on the treatment group. These are examples of confounds that could have been easily controlled through planning prior to the experiment. However, other confounds may exist which are difficult or impossible to overcome. For example, biological variance (variance due to the animal and the treatment) tends to be confounded with the microarray since it is rarely feasible to place a sample from every level of the treatment on every microarray. Thus, a complete block design (i.e., where every level of

Table 1-3: Sources of Variance

Type	Name
Array-wise Technical Variance	Microarray
	Feature
	Dye
	Microarray Print
	Subgrid
Non-array technical variance	Investigator
	Growth Conditions (cage/flask, husbandry, medium)
	Circadian
	Day
	Treatment formulation
	Dissection
Biological	Animal/Cell culture

treatment effect exists on every level of the confounding variables, such as in Table 1-2) for a microarray experiment is rarely feasible.

Although some of the sources of variance within a microarray study [such as the growth conditions, husbandry conditions, and the time of day at the time of sacrifice (e.g., circadian effects)] are relatively obvious, other sources include 1) array-wise technical variation, 2) non-array technical variation, and 3) biological variation (Shih, *et al.*, 2004). Array-wise technical variation includes all sources that are specific to the microarray process, from fluorescently labeling the sample, hybridization, and scanning. Sources of non-array technical variation include growth and husbandry conditions for the animals and cell cultures used in the experiments, or the local environment in the case of tumor samples and biopsies, circadian differences when the harvesting occurs at different times of day or different times within the light/dark cycle, differences in the treatment formulation (e.g., non-uniform suspension), and technical differences during dissection.

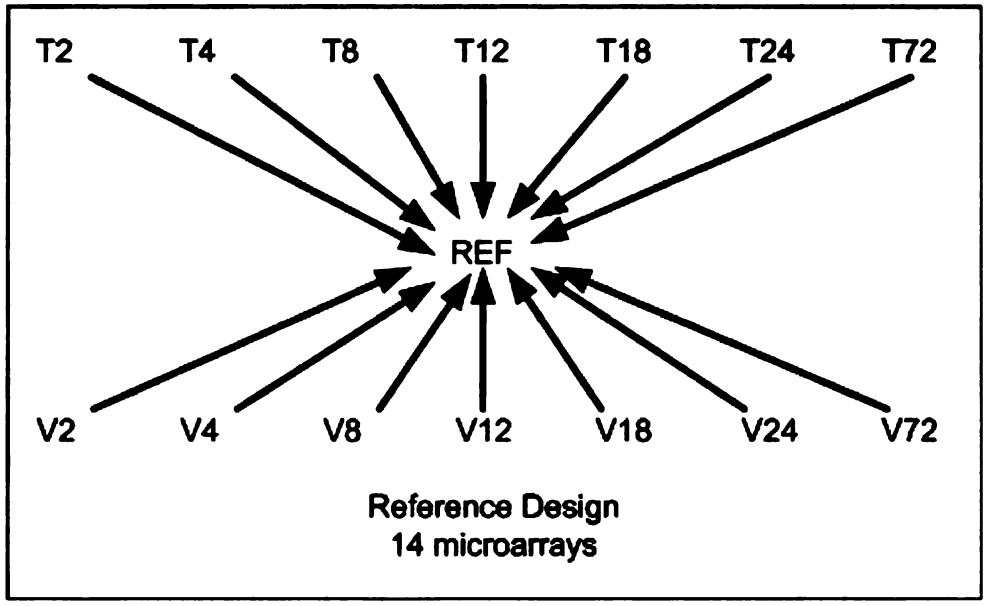

Figure 1-3: Reference Design. Each arrow represents a microarray, where the head denotes the Cy5 labeled sample, and the tail denotes the Cy3 labeled sample. T*n* represents a treated sample at the *n*th time-point; V*n* represents a vehicle sample at the *n*th time-point.

Due to the complexity of the microarray experiment, many of these sources of variance are compounded with each other in ways that are difficult to control experimentally. For example, a dye-bias has been demonstrated to exist within two-channel cDNA microarray data (Cox, *et al.*, 2004; Dobbin, *et al.*, 2003; Dombkowski, *et al.*, 2004; Eckel, *et al.*, 2005; Fare, *et al.*, 2003; Workman, *et al.*, 2002). There are several potential causes for bias including steric hinderence inhibiting dye incorporation, dye-dye interactions leading to the quenching of fluorescent signal (Cox, *et al.*, 2004), fluorophore exposure to ozone (Fare, *et al.*, 2003), and the sensitivity of the laser detector on the microarray scanner. This dye-bias may interact with subgrid effects on the microarray resulting in unreliable, inaccurate measurements if not controlled.

Initially, the most commonly used experimental design for two-channel array data was the reference design (RD) (Churchill, 2002; Kerr and Churchill, 2001a; Kerr and Churchill, 2001b; Vinciotti, *et al.*, 2004). The design consists of a pooled reference

sample that is labeled with the same dye, and hybridized to each microarray. The experimental samples are labeled with the same dye (other than that used to label the pooled reference) and hybridized to their respective arrays (Figure 1-3). The pooled reference sample, which has no biological significance, consists of an aliquot from each biological sample of interest. The purpose of the reference sample is to prevent division by zero errors when calculating fold change ratios. This design exhibits a confound between the biological treatments and the dye (i.e., the variance due to the dye is inseparable from the treatment variance) (Churchill, 2002; Yang and Speed, 2002), increasing the variance within the biological treatment, ultimately affecting the list of significant or active genes. By performing a dye swap, where replicate microarrays are performed with the dyes are reversed, the confound between the biological treatment and dye is alleviated. This is equivalent to the blocking procedure used in the example (Table 1-2) from the previous section. One of the primary concerns with the RD is the over collection of data concerning the relatively uninformative pooled reference compared to the treatment groups of interest (Kerr and Churchill, 2001a; Kerr and Churchill, 2001b; Yang and Speed, 2002). Another concern with the RD is that the amount of pooled reference required increases with each additional treatment group.

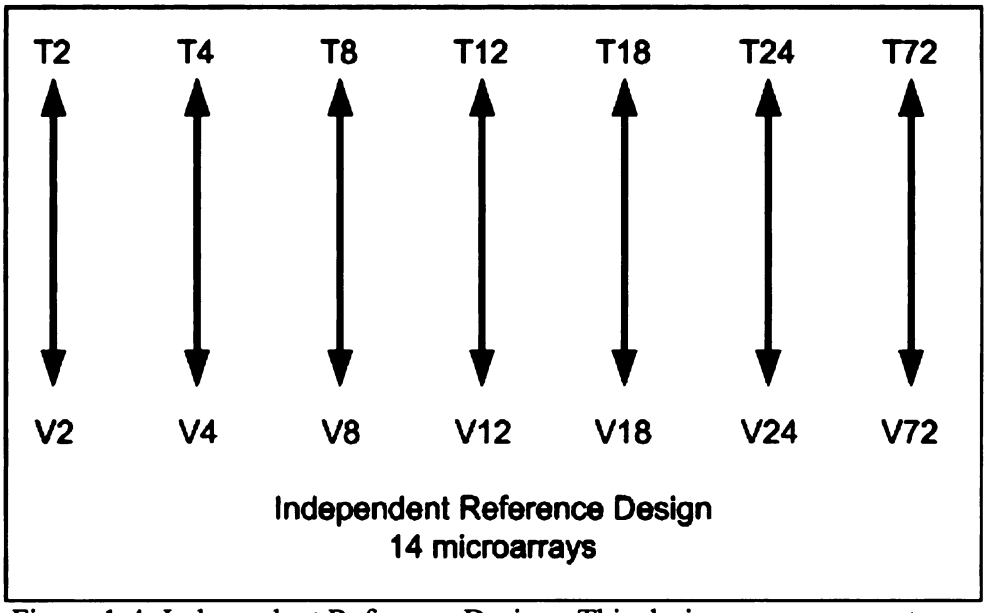

Figure 1-4: Independent Reference Design. This design encompasses two microarrays per time-point, with dye swap. A time-point is confounded within the same arrays using this design, although the dye is not confounded with treatment. T*n* represents a treated sample at the *n*th time-point; V*n* represents a vehicle sample at the *n*th time-point. Double headed arrows represent the dye swap, where each sample is labeled with both Cy3 and Cy5.

The independent reference design (IRD) was developed as an alternative to the reference design (Fielden, *et al.*, 2002b), and represents a direct design (Yang and Speed, 2002) where the comparisons of interest exist within the same microarray (Figure 1-4). In this case, comparisons are made within each time-point, between treated and vehicle samples. Thus, the variability due to the microarray is confounded with each time-point, however, no confound is exhibited between the treatment and dye. Although the IRD is more efficient than the RD from the standpoint that it does not collect data from an uninformative reference group, it exhibits a temporal confound (i.e., the time-point term is completely confounded with the microarray term) that contributes negatively towards its general utility.

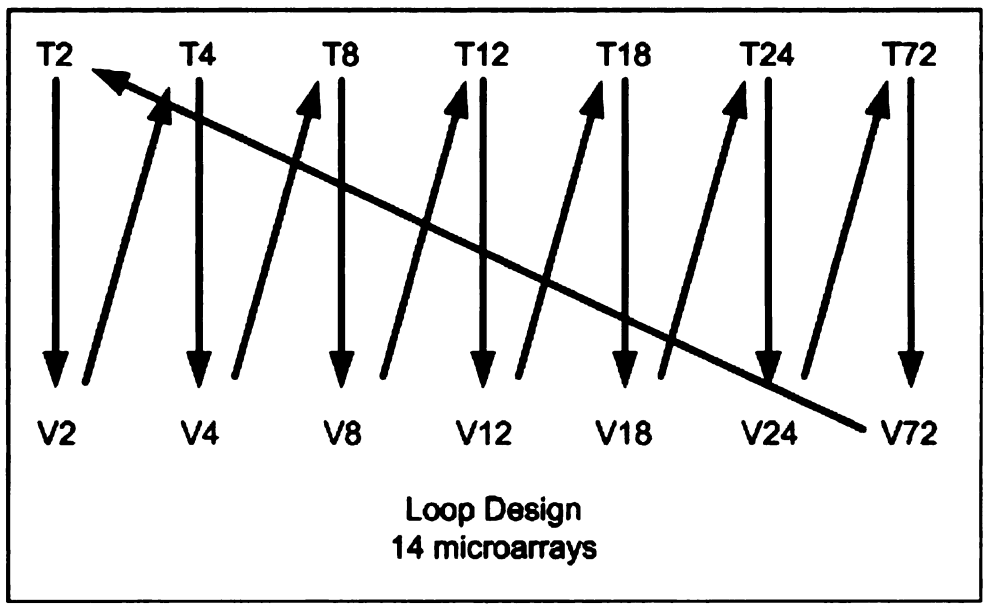

Figure 1-5: Loop Design. The Loop Design represents a balanced block design with respect to the dyes. Treatment and vehicle samples from the same time-point occur on the same microarray, as well as on the arrays with neighboring treatment groups. Each arrow represents a microarray where the head represents the Cy5 sample, and the tail represents the Cy3 sample. T*n* represents a treated sample at the *n*th time-point; V*n* represents a vehicle sample at the *n*th time-point.

Another alternative to the RD, that does not exhibit this temporal confound, is the loop design (LD; Figure 1-5) (Kerr and Churchill, 2001a; Kerr and Churchill, 2001b). The LD represents a hybrid direct and indirect design, where the comparisons of greatest interest occur on the same microarray, with the capability to make comparisons across the entire loop. As the size of the loop increases, the confidence with which one makes comparisons between distant nodes across the loop decreases (Kerr and Churchill, 2001a). Thus, with a large number of treatment varieties [e.g., where a treatment variety may be considered the treatment levels and the time levels; for a 2 treatment (treatment vs vehicle), 7 time-point experiment, there would be 7 x 2 = 14 treatment varieties] it is possible that comparisons within a treatment class (e.g., comparisons across time, but

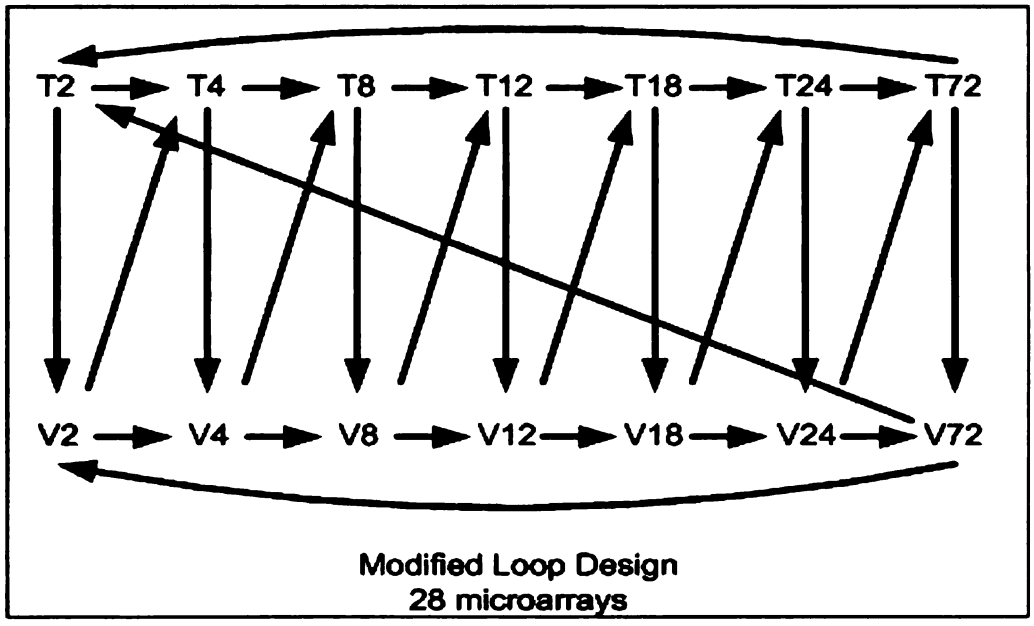

Figure 1-6: Modified Loop Design. Consisting of the loop design combined with two inner loops for each of the major classes of treatments, the modified loop design augments with the ability to perform comparisons across time, but within treated and vehicle groups. Each arrow represents a microarray where the head represents the Cy5 sample, and the tail represents the Cy3 sample. T*n* represents a treated sample at the *n*th time-point; V*n* represents a vehicle sample at the *n*th time-point.

within the treated or vehicle group) becomes less than optimal; the variance begins to increase as a function of the increased distance.

The modified loop design (MLD; Figure 1-6) (Boverhof, *et al.*, 2004) combats the size optimality problem of the LD by including two "inner" loops – one for each treatment group (e.g., treated and vehicle). These inner loops facilitate comparisons across time and within treatment class (e.g., treated or vehicle). The major drawback to the modified loop design is the number of microarrays required compared to the LD and IRD.

Another significant consideration when designing microarray experiments is the goal of the experiments. Typically, the goal in pharmaco- and toxicogenomics is the assessment of a biological response within a population. With this goal in mind, the

investigator must be mindful of the need to have a suitable number of randomized biological replicates, or else the estimated response will be inaccurate. This need for biological replication must also be balanced with the need for technical replicates (i.e., replication of a sample's measurement) to counteract large variances due to the microarray.

Genomics

Currently, genomics is dominated by the use of cDNA microarrays (Schena, *et al.*, 1995). The cDNA microarray is a glass slide with thousands of expressed sequence tags (ESTs) affixed to it. A microarray can be thought of as a distinct geographical entity, consisting of several islands arranged in a particular topography, typically a grid on a Cartesian plane (Figure 1-7). Each island, or subgrid, consists of hundreds of spots, or features, also arranged within a particular topography, again, typically a Cartesian plane. In the case of spotted arrays, every feature printed within a subgrid was spotted by the same print-tip, a needle used to transfer the cDNA. As a result, the features from the same print-tip typically illustrate some degree of covariance (Yang, *et al.*, 2002b).

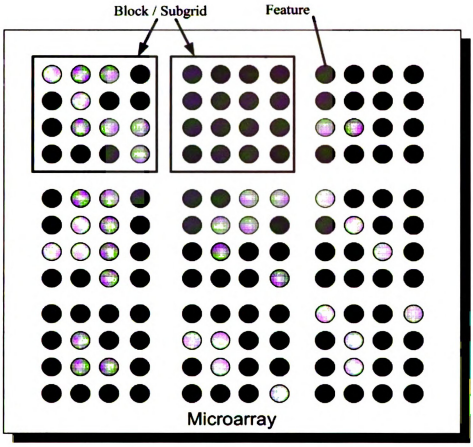

Figure 1-7: Schematic of a spotted microarray. The microarray consists of spots, or features, that are arranged within blocks, or subgrids. Each feature within a subgrid is printed by the same print-tip.

For gene expression experiments, total RNA is extracted from the samples of interest and labeled with a fluorescent dye (e.g., Cy3 or Cy5). These labeled samples, or labeled extracts, are mixed together according to the experimental design, and hybridized to the microarray. The microarray is scanned with a confocal microscope with detectors for the dyes, resulting in a Tagged Image File Format (TIFF) image representing the fluorescent intensities at each feature. The TIFF images for an experiment are quantified using software such as GenePix, resulting in a tabular representation of the data, where

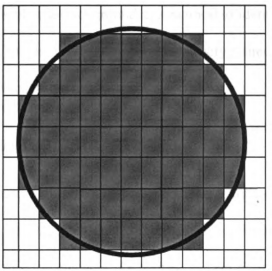

Figure 1-8: Pixel map of a microarray feature. Images are made up of pixels, small rectangles which contain only one color. This zoomed in diagram of a feature shows the pixels as rectangles. The circle drawn within the feature is the best circle that can be fitted by the software. The feature measurements are taken from within this circle as the weighted mean and medians of the pixel intensities (the weight is the percentage of the pixel within the circle).

each row represents a distinct feature, and columns represent the intensity in the fluorescent channels.

Due to a limitation within the TIFF standard, fluorescence values exist within the chromatic scale $x \in \{0..65{,}535\}$, where 0 represents a pixel of no signal, and 65,535 represents a pixel with the highest, saturated signal. Each feature on a microarray consists of several pixels, and due to their near circular morphology (imposed by the software; features may actually exist as other shapes), they fail to represent a whole number of pixels, complicating the quantification process (Figure 1-8). Software, such as GenePix, is employed to overcome these difficulties. These software packages report signal intensity per feature as distributions, where the median, mean, and standard deviations are calculated. They also calculate the mean, median, and standard deviation of the background intensity for each feature. As it has been shown previously that background subtractions obfuscate further data analysis (Qin and Kerr, 2004; Tran, *et al.*,

2002), our analysis protocols use background as a method to identify whether or not a feature is present (i.e., if the median feature signal intensity ≤ median background feature intensity, then the feature is absent).

Microarray Data Analysis

One important feature of microarray data analysis is the activity of normalizing the quantified data from the raw microarray images, and identifying genes with treatment-influenced expression changes. Generally, the act of quantifying the microarray data is automated through various software processes. These quantified data are exported to files, and generally uploaded to a database for data management and association with sample annotation information (e.g., treatment, animal, age, sex, physiological status). Captured quantified data are then normalized to reduce technical variation across the samples, while maintaining treatment effects (Cheadle, *et al.*, 2003; Eckel, *et al.*, 2004b; Quackenbush, 2002). Following normalization the data are analyzed using any one of a myriad of statistical techniques to identify treatment-influenced gene expression changes (i.e., active genes).

Microarray Image Quantification

Following image acquisition, investigators quantify the data from the microarray image using special software that detects each feature and reports the median and mean feature signal intensity, as well as the median and mean background intensity. For all microarray experiments reported within this body, median feature signal intensity values are used for analysis as they are robust to outlier pixels within a feature. Outlier pixels

typically result from the software package not being capable of accepting any feature morphology other than circular.

Following image quantification, data are submitted to a database and associated with their sample annotation information. MIAME supportive databases typically require submission of the TIFF images (a.k.a. raw data) in addition to the quantified data. Submission of the TIFF images is desirable as there is generally a lack of consensus on how to properly quantify microarray data, and re-quantification by others may be desirable when a more general consensus on data analysis is formed.

Microarray Data Normalization

Over the years several different microarray data normalization techniques have been developed, all with the goal of decreasing the technical variance within the assay, while not disturbing the treatment variance. Each normalization technique that has been developed focused on a different and seemingly important technical variant. These techniques can be grouped as the 1) local normalizations, 2) global normalizations, and 3) the hybrid techniques, encompassing the strengths of the first two groups.

The local normalization techniques operate on the assumption that a great deal of the variance within a microarray is due to some local subset of the data, most commonly the print-tip (Quackenbush, 2002). Examples of these methods include the lowess regression based on print-tip groups scaling for dye biases (Yang, *et al.*, 2002a; Yang, *et al.*, 2002b); and ratio-based normalization where the same cDNA is spotted in each print-tip group, and a spike-in control is used to calculate a local correction factor based on a ratio of 1 for each print-tip group (Lashkari, *et al.*, 1997); and mean or median centering of data across an array. These methods work to normalize the data within the local level

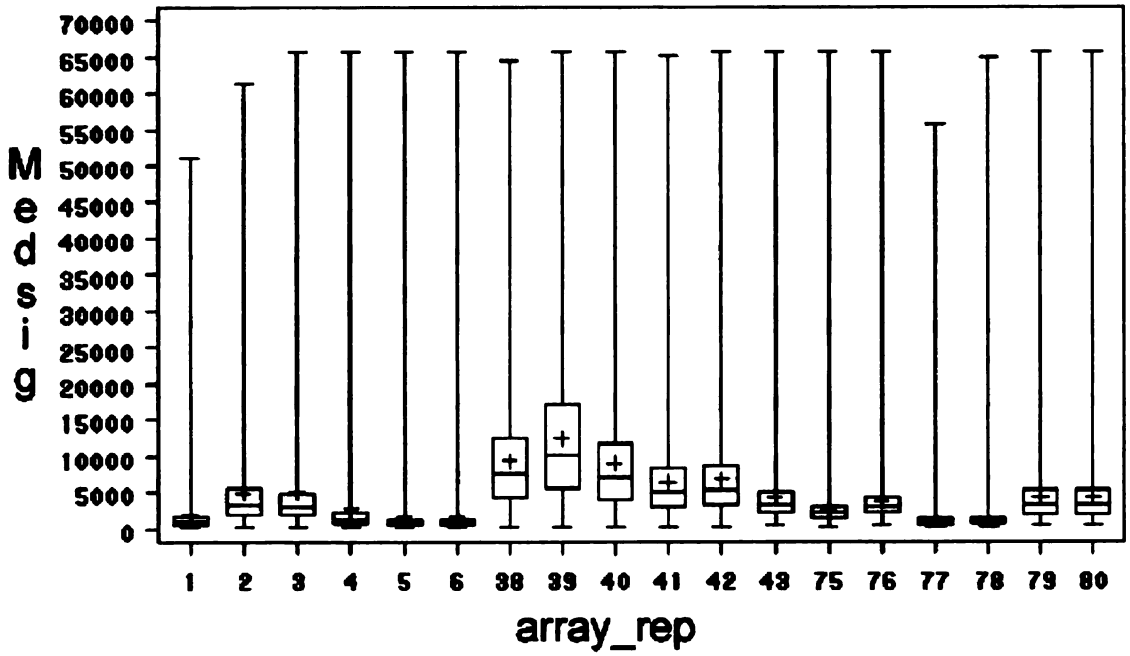

Figure 1-9: Distribution of median signal intensities. This figure illustrates a relatively common, although undesirable, phenomeon where the distribution of data within microarrays varies greatly. The y-axis represents the median signal feature intensity, while the x-axis represents the microarray. The boxes represent interquartile range; that distance from the 25th to the 75th percentile. The cross is the mean of the distribution, while the horizontal line within the box represents the median (50th percentile). The whiskers represent the remainder of the distribution, with the uppermost whisker representing the largest data point, and the lower whisker representing the smallest data point.

(e.g., subgrid or microarray) such that data across an array approximate the same distribution.

The underlying assumption of local normalization techniques is that similar distributions are approximated by all of the microarrays in the study prior to normalization. Thus, the purpose of the local normalization is to internally shift the data, or to make fine adjustments, without altering the overall distribution. These normalization methods would thus fail if used in the experiment illustrated in Figure 1-9, where the distributions across the microarrays differ substantially.

The global normalization procedures perform normalizations across the entire experiment. One example of this is the Z-score centering of data, forcing all of the data

within the experiment to conform to the same distribution (Fielden, *et al.*, 2002a). Another example is the General Linear Mixed Model (GLMM) approach, where the normalized data values are the residuals from the model (Wolfinger, *et al.*, 2001). The GLMM approach regresses the data to fit a specified linear model. The residual represents the difference between the predicted value from the model and the sample data from the microarray. The global model is more capable of normalizing the data across the microarrays than the local model, however, they typically perform less well in normalizing the local effects (e.g., subgrid) (Quackenbush, 2002; Yang, *et al.*, 2002a).

More recent normalization techniques attempt to build upon the strengths of the local and global types of normalization. One example of this hybrid technique is the semiparametric normalization (Eckel, *et al.*, 2004b). Here, the lowess regression, a local technique, is combined with a global approach – the lowess regression model is built on a subgrid and treatment basis, thus it normalizes data both within a subgrid and across microarrays. Thus, the response within subgrid and across the same treatment is modeled within the regression to normalize the data across the microarrays; decreasing the technical variation within the treatment groups, without sacrificing the biological variation. When this normalization technique is coupled with a design that incorporates dye-swaps, where each sample is labeled with both dyes, the normalization will also account for dye-biases.

Microarray Data Filtering

Datasets from cDNA microarrays yield a large amount of data, and these data must be trimmed to make it more reasonable for investigators to follow-up on the results. Generally, investigators are concerned with identifying genes that are the most changed

due to treatment. Numerous methods exist for doing this, including the common use of arbitrary fold change cut-offs, the t-test (Fielden, *et al.*, 2002b), Wilcoxon rank sum test (a nonparametric analogue of the t-test) (Efron and Tibshirani, 2002), the SAM method (Tusher, *et al.*, 2001), and empirical Bayes methods (Eckel, *et al.*, 2004a; Efron and Tibshirani, 2002).

Fold change cut-offs were the original method for determining active genes. Generally, a two-fold change was considered significant, and those genes would be used for further follow-up. However, fold change cut-offs tend to meet with significant resistance in the absence of more rigorous statistical methods, especially since there is no biological significance attributed to the fold change cut-off. For example, no doctrine dictates that 2-fold changes are more important than 1.5-fold or 2.5-fold changes. Furthermore, it is generally accepted that ratios track with fluorescence intensity (Yang, *et al.*, 2002a), and that fluorescence intensity is related to the number of copies of a message within the tissue. However, ratios yield less insight to the number of copies present than absolute intensity. For example, consider four cell populations, population A has 2 copies of a message, population B has 4 copies, population C has 12,000 copies, and population D has 24,000 copies. The ratio of B:A = D:C (i.e., the ratio in both cases = 2). The absolute difference in the number of copies is drastically different (2 vs 12,000), however the ratio is the same. Thus, it is difficult to make comparisons across or within experiments using the ratio alone.

The t-test and Wilcoxon rank sum tests provide the ability to make comparisons between treatment groups. The t-test makes the assumption that the data are independent, and identically distributed, and follow a normal distribution. The Wilcoxon test does not

assume a particular distribution, however it does require the shape of the two population distributions to be the same. The general idea of both tests is to identify whether or not the treated and comparator (typically vehicle) groups exist within different data distributions. These tests are performed on a per-gene basis, thus necessitating follow-up with an additional test to control the false positive rate.

Generally, investigators tend to term the genes that survive filtering "significant", implying some degree of statistical rigor has been satisfied with regard to the gene's expression compared to some other population (e.g., vehicle gene expression). However, in this body the term "active" is used, as opposed to "significant", as ranking statistics (typically the same statistics that others use for significance, without the use of a p-value) are used exclusively for ranking and prioritization of genes for further investigation and inclusion. Thus, the likelihood of these "active" genes having been treatment altered is greater than those who are further down the list. By defining these genes as active, there is no mention or interpretation as to the distribution that these genes come from; in other words, there is no statement that active genes must necessarily exist within a distribution other than that for vehicle treated genes. Thus, active genes are not statistically or biologically significant *ipso facto*. The only stipulation for a gene to be active is that there is a higher likelihood that the gene's change in expression is due to treatment; inclusion in the list only necessarily dictates that more rigorous follow-up experiments be pursued at a later time [e.g., quantitative real-time polymerase chain reaction (QRT-PCR)].

The SAM method is similar to the t-test except it performs an additional adjustment with respect to the standard error (Tusher, *et al.*, 2001). Microarray data tend

to illustrate variable standard errors, where the standard error for low intensity genes is much less than the standard error for high intensity genes. As the standard error acts as a penalty in the calculation of the t-statistic (i.e., the t-statistic is inversely proportional to the standard error), low intensity genes would tend to have larger t-statistics. The SAM method adjusts the standard error by a user-defined factor to adjust the standard error, and reduce this bias. The primary concern with the SAM method is that the factor may itself be chosen inappropriately, thus further biasing the t-statistic, and penalizing otherwise active genes.

The empirical Bayes methods (a.k.a. hierarchical Bayes methods) are an application of Bayes' theory to the problem of identifying active genes. Bayes' theory holds that the likelihood an event will occur is dependent upon the prior probability of that event happening (Gelman, *et al.*, 2004). The prior probability can be thought of as a historical probability, that is, it is the known probability that an event will occur based on past trials. For example, to determine the probability that a person has a disease using a diagnostic kit, it is necessary to know the historical probabilities of correct and incorrect diagnoses, especially with respect to the patient either actually having the disease or not. These historical probabilities are the prior probabilities. Generally, the prior probability must be implicitly stated; however there are mathematical means of deriving suitable and appropriate prior probabilities when they are unknown. For example, the empirical Bayes models do not require specification of the prior probabilities as they can be inferred mathematically from relationships within the existing dataset (Eckel, *et al.*, 2004a; Efron and Tibshirani, 2002; Gelman, *et al.*, 2004). The key difference between these methods and the previously mentioned ones is that the empirical Bayes methods

require familiarity with statistical model building and are not generally accessible to a larger biological audience without the assistance of trained investigators.

Data Interpretation

Once a microarray dataset has been distilled down to a list of the most active genes, it must be further analyzed to facilitate data interpretation. Biological interpretation of the data is a daunting task, and requires the integration of data from several sources, including up-to-date gene annotation. Investigators will need as much information as possible to accurately interpret the data, including gene names, abbreviations, and aliases for literature searches; cellular and extracellular locations; functional annotation; disease processes the gene participates in; and biological interaction data (e.g., protein-protein interactions). This information is oftentimes available in biological databases devoted to a particular purpose or data domain.

Biological databases are grouped by the type of data they manage into several categories. Figure 1-10 depicts a subset of the more common databases and groups them as they relate to genomic data integration; excluded from the figure are the metabonomics related domains which are currently in development. All of the databases exist in an extremely complex data exchange continuum, where some databases rely entirely upon others for their information, others are nearly independent of the rest, and the remaining host a smorgasbord of data integrated from several different levels. Generally speaking, however, genome sequences, from databases such as Ensembl (Clamp, *et al.*, 2003; Hubbard, *et al.*, 2005), Entrez Genomes (Wheeler, *et al.*, 2004), and the UCSC Genome Browser (Karolchik, *et al.*, 2003), can be thought of as the root of the universe. From these genomic templates, expressed sequence tags and cDNAs in GenBank (Wheeler, *et*

al., 2004) can be clustered together and associated with genes (i.e., UniGene (Wheeler, *et al.*, 2004)), and exemplary, representative sequences can be identified and mapped back to genes in the genome (i.e., RefSeq (Wheeler, *et al.*, 2004)). These elements are then annotated in databases such as Entrez Gene (Maglott, *et al.*, 2005), where functional information (Gene Ontology (Harris, *et al.*, 2004)), and genetic disease information (Online Mendelian Inheritance in Man; OMIM (Wheeler, *et al.*, 2004)) are integrated to give a more full picture of the gene's function.

The same elements from the sequence level databases are also represented on microarrays. This provides the relationship that facilitates functional annotation of active genes from microarray experiments. Microarray data are captured locally, for a laboratory or consortium, within laboratory information management systems (LIMS), and disseminated to the public through repository systems such as the Chemical Effects in Biological Systems Knowledgebase (CEBS) (Waters, *et al.*, 2003a), ArrayExpress (Brazma, *et al.*, 2003; Rocca-Serra, *et al.*, 2003), and the Gene Expression Omnibus (GEO) (Edgar, *et al.*, 2002).

Tie-in of data between the genomic and proteomic level is also capable through sequence relationships, from the mRNA to the protein translation. This facilitates further functional predictions, by analyzing the protein domains that might exist. Interaction data, from databases such as BIND (Biomolecular Interaction Network Database) (Bader and Hogue, 2000) and DIP (the Database of Interacting Proteins) (Xenarios, *et al.*, 2000), can also be gleaned from these proteomic databases to build new networks, facilitating the development of communication models and novel modes of action.

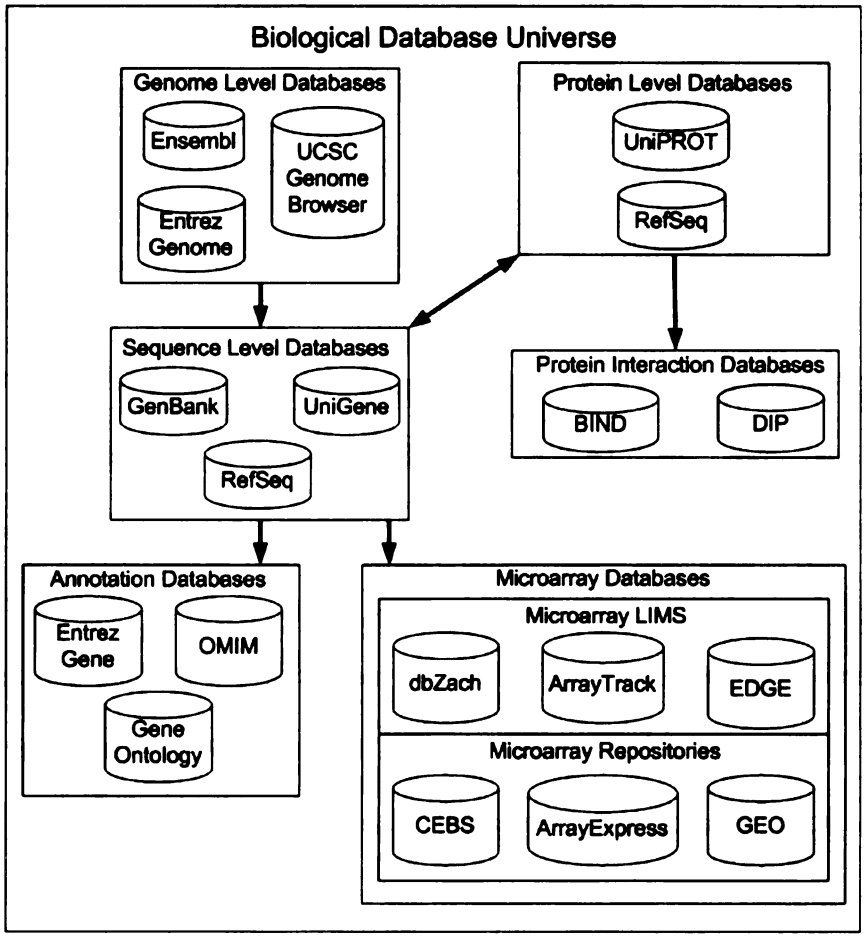

Figure 1-10: The Biological Database Universe. The biological database universe is ever growing, and this figure depicts six of those levels as they pertain to genomic data analysis and interpretation. The Genome Level Databases catalog data with respect to the full genome. The Sequence Level Databases catalog sequence reads from cells, including genomic sequence and expressed sequence tags (ESTs). Annotation Databases provide functional information about genes and their products. Protein Level Databases provide information on protein sequences, families, and domain structures. The Protein Interaction Databases provide interaction data concerning proteins, genes, chemicals, and small molecules. The Microarray Databases include local laboratory information management systems (LIMS) and data repositories. Arrows in the figure depict communication between the different domains, where information from one level may exist in another level to allow for cross-domain integration.

Genome Level Databases

Genome level databases manage, at the very least, genomic data. However, they differ in their integration of other types of data and often in their assignment of computationally defined genes. The three primary genome level databases are the Ensembl database (Clamp, *et al.*, 2003; Hubbard, *et al.*, 2005), the Entrez Genomes database (Wheeler, *et al.*, 2004), and the University of California Santa Cruz Genome Browser (Karolchik, *et al.*, 2003). All three databases use different techniques for predicting genes and gene structures (e.g., untranslated regions (UTR), regulatory regions, introns, and exons).

The Ensembl database utilizes several different methods for the prediction of genes and gene structures (Curwen, *et al.*, 2004). The method is biased towards the alignment of species-specific proteins and cDNAs, and using orthologous protein and cDNA alignments when necessary. The use of the protein and cDNA alignments against the genome sequence facilitates the identification of exonic and intronic sequences and UTRs (Figure 1-11). A putative transcription start site (TSS) can be obtained from this, defining the end of the upstream region.

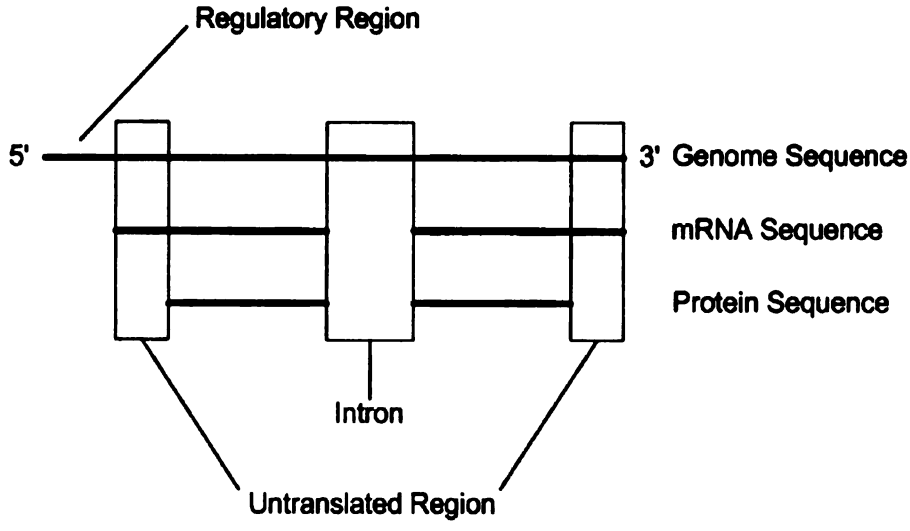

Figure 1-11: Ensembl Genome Annotation. This simplified view of the Ensembl genome annotation system illustrates their method for identifying gene structures, such as the untranslated region (UTR), exons, and introns by combining genome, mRNA, and protein alignments.

The National Center for Biotechnology Information (NCBI) Entrez Genomes database annotates genes based on the RefSeq database of reference, exemplary sequences. RefSeq sequences are aligned to the genomic sequence using the MegaBLAST algorithm; additional mRNA and ESTs are aligned to find more genes (http://www.ncbi.nlm.nih.gov/genome/guide/build.html#contig; accessed April 5, 2005).

The UCSC Genome Browser uses the NCBI genome builds for its annotation, however, previously the Genome Browser used similar annotation sources as the Entrez Genome. Today there are no differences between the human and mouse genome builds. However, the mouse genome reported at UCSC is the C67/Bl6 strain. Previously, the primary difference between the two methods was in their genome assemblies, where Entrez Genome used sequence entries from the GenBank database to drive assemblies, while the UCSC Genome Browser uses BAC clones, mRNA sequences, and a greedy algorithm (greedy algorithms divide a problem into parts, and identify the locally

optimized solution to each part independently, and combines those to form a larger solution, which will hopefully be the globally optimum solution) to furnish the assembly; resulting in differences in the genome assemblies (Rouchka, *et al.*, 2002). Without knowing the true assembly of the genome, it is difficult to ascertain which assembly was more or less correct than the others. This had vast implications with regards to SNP, promoter, and other data mining applications with regards to these differences (Rouchka, *et al.*, 2002). Today, NCBI and UCSC agree in their annotation as UCSC uses build from the NCBI or genome authorities (see http://genome.ucsc.edu/FAQ/FAQreleases for further details).

Sequence Level Databases

Sequence level databases manage data with respect to a particular sequence read of an EST or cDNA. These databases may deal with those sequences directly, as is the case for GenBank and RefSeq, or they may manage them on a larger scale, where multiple sequences are grouped together, as in UniGene. Generally, these databases provide the first level of annotation for microarray studies, as the sequences are directly represented on the microarrays.

GenBank Accession numbers are generally the most commonly used identifier for probes attached to microarrays. The GenBank Accession matches the probe to one sequence within the GenBank database (Wheeler, *et al.*, 2004); a database of submitted biological sequences (ESTs, cDNAs, etc). The UniGene database creates non-redundant gene clusters based on GenBank sequences (Wheeler, *et al.*, 2004). Clusters are built by sequence alignment, and annotated based on overall sequence alignment to genes in the Entrez Gene database.

The RefSeq database provides exemplary transcript and protein sequences based either on hand curation or based on information from a genome authority (e.g., the Jackson Labs) (Pruitt and Maglott, 2001; Wheeler, *et al.*, 2004). There are currently seven categories of RefSeq records (Table 1-4): 1) genome annotation, 2) inferred, 3) model, 4) predicted, 5) provisional, 6) validated, and 7) reviewed (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status; accessed April 7, 2005). The first category, genome annotation, includes mRNA and protein records that are aligned to the annotated genome. Inferred records are those that are predicted based on the genome analysis, but there is no mRNA/EST that exists within GenBank to qualify the prediction. Records labeled as "model" are predicted based on gene prediction methods, and may or may not have a known transcript associated with it. Predicted, represents protein and transcript sequences from genes with unknown functions. Provisional records represent

Table 1-4 RefSeq Categories

RefSeq Record Category	Source of annotation
Genome annotation	Records are aligned to the annotated genome
Inferred	Predicted to exist based on genome annotation, but no record in GenBank exists to qualify the prediction
Model	Predicted based on bioinformatics prediction methods; a known transcript may or may not exist
Predicted	Sequences from genes with unknown functions
Provisional	Sequences associated with genes of known function that have not been reviewed by NCBI personnel
Validated	Sequences associated with genes of known function that have undergone an initial review
Reviewed	Sequences representing genes of known function that have been completely reviewed by NCBI personnel

genes with known functions, but which have not been verified by NCBI personnel. Validated records have undergone an initial review, and are awaiting further review by NCBI personnel. The reviewed status is reserved for those RefSeq records representing genes of known function that have been reviewed by the NCBI personnel. For further, and updated information on the status codes used by RefSeq see: http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status.

RefSeq accession numbers follow a PREFIX_NUMBER format (e.g., NM_123456, or NM_123456789). All curated RefSeq transcript accessions are prefixed by an NM, while XM prefixes represent accessions which have been generated by automated methods. Some of the NM transcript accessions have also been generated by automated methods, but all NM transcripts are relatively mature (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status; accessed April 7, 2005).

Annotation Databases

Annotation databases provide functional information for genes, and may also catalogue the gene's structure. These databases serve as a launching point for mechanistic understanding and hypothesis generation from microarray data. Several domain specific annotation databases exist, including those that focus on particular species, such as the Mouse Genome Database (Eppig, *et al.*, 2005).

The Entrez Gene database is a part of NCBI's Entrez suite of bioinformatics tools. Entrez Gene is a source for information on annotated genes in several different genomes, including human, mouse, rat, and dog (Maglott, *et al.*, 2005). Annotated genes are defined as those that have a RefSeq identifier associated with them, or those that have

Table 1-5: Entrez Gene Annotation Categories and Sources (adapted from Maglott, *et al.*, 2005)

Annotation Categories	Source
Gene names and abbreviations/symbols	Publications and genome authorities
RefSeq Sequence	RefSeq database
Genome position and gene structures	Genome databases
Gene Function	Gene Ontology (GO) database, Gene References into Function (GeneRIF)
Expression Data	Gene Expression Omnibus (GEO), EST tissue expression from GenBank

been annotated by an annotation authority (e.g., Jackson Labs for mice). As such, entries within Entrez Gene may or may not have a RefSeq associated with them, and those entries with associated RefSeq accessions may have either the NM (mature) or the XM (non-reviewed) series. Thus, an Entrez Gene record may not have an exemplary RefSeq sequence associated with it.

Entrez Gene serves as a focal point for gene annotation, integrating data from many sources, including databases outside NCBI. Some of this data integration is achieved through hyperlinks to the appropriate database entries, while others are catalogued on the detail page for that gene. Table 1-5 (adapted from Maglott, *et al.*, 2005) lists several of the annotation categories, and their sources. The most basic form of gene annotation is the gene name and the abbreviation. These are necessary to begin functional annotation of the gene through the literature. The Entrez Gene database also integrates data from the RefSeq, Gene Ontology (GO), Gene Expression Omnibus (GEO), Gene References into Function (GeneRIF), and GenBank databases. The RefSeq sequences, both mRNA and protein, facilitate sequence based searching, such as identifying other genes that may be homologous, or identifying gene function based on

protein domains. The GO database catalogues genes by their molecular function, cellular location, and biological process. Information about the expression of genes can be obtained from the GenBank database, where the tissue localization for an EST is recorded, as well as the GEO – NCBI's gene expression repository (Wheeler, *et al.*, 2004). GeneRIFs provide curated functional data and literature references. GeneRIFs serve as a useful starting point, however, they typically do not provide the most up-to-date functional annotation from the literature. Investigators can facilitate GeneRIF updates by submitting suggestions directly to the NCBI through their update form: http://www.ncbi.nlm.nih.gov/RefSeq/update.cgi.

For human studies, the Online Mendelian Inheritance in Man (OMIM) database, the online version of the Mendelian Inheritance in Man (McKusick, 1998), provides linkages between human genes and diseases (Hamosh, *et al.*, 2002; Wheeler, *et al.*, 2004). The OMIM database is searchable through the NCBI Entrez system. Links to the OMIM database are provided within query output pages from the Entrez Gene database. For many of the diseases within OMIM, a synopsis of the clinical presentation is provided in addition to links to the genes associated with the disease. PubMed citations are also made available through the OMIM database, with hyperlinks to the PubMed database entries. Also, the OMIM contains information on known allelic variants and some polymorphisms (Hamosh, *et al.*, 2002).

The Gene Ontology (GO) (Harris, *et al.*, 2004) database is another source of gene functional annotative information. The database consists of an ontology (i.e., a catalogue of existents/ideas/concepts and their interrelationships (Cox, 1999)) where terms exist within a directed acyclic graph (DAG; Figure 1-12). DAGs are graphical

structures that cannot exist as loops, thus, a child node (i.e., an object or concept) may not also serve as its own predecessor (i.e., parent, grandparent, great-grandparent, etc...). Any child node within a DAG may have any number of parents, and any number of paths to get to the child. For example, Figure 1-12 shows two paths leading to the same child, GO:0045814: negative regulation of gene expression, epigenetic. Here it is evident that this epigenetic negative regulation of gene expression is both a regulation process and critical in development. GO entries that exist at the same level relative to the root, or starting node, do not necessarily reflect the same level of specificity. The level of specificity afforded must be taken on a per DAG basis, and not relative to the other DAGs. Thus, a 4^{th} order node (a node that is 4 levels below the root node) in one DAG has no specificity relationship with regards to a 6^{th} order node in a different DAG. At each mode within the GO there may exist a list of genes. As the annotation for a gene improves, it may change node associations. For example, if gene X were previously GO:0040029 (regulation of gene expression, epigenetic), and new experimental data suggested gene X was a negative regulator of gene expression through an epigenetic mechanism, it would be reassigned to GO:0045814 (negative regulation of gene expression, epigenetic).

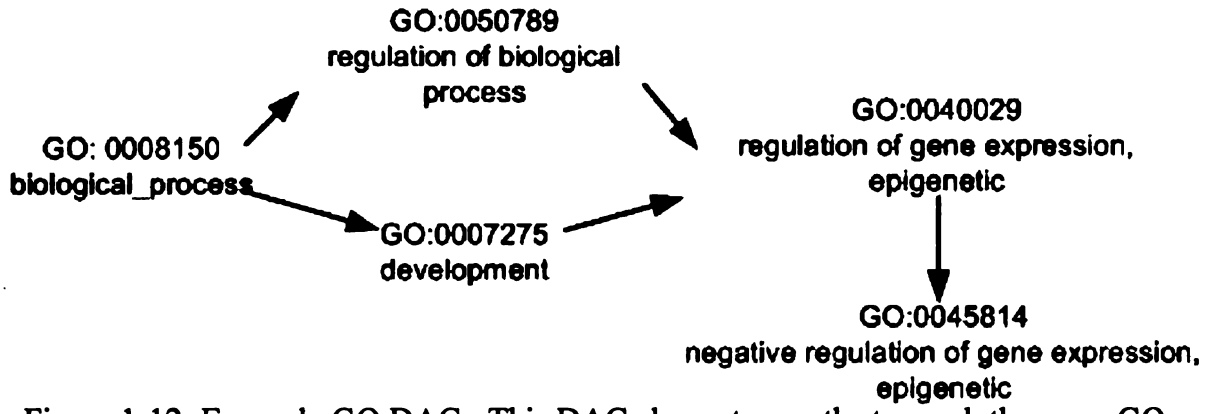

Figure 1-12: Example GO DAG. This DAG shows two paths to reach the same GO entry, GO:0045814. It is important to note that the DAG travels from the most general case, and becomes more specific with entries that are farther down the DAG.

Protein Level Databases

In the course of interpreting gene expression results, it is useful to consult protein databases to identify the proteins that may be encoded by the genes of interest. Some of the gene annotation databases mentioned above provide links to this information, such as Entrez Gene and the Ensembl databases. However, as the gene sequence level databases provided sequence anchoring for the higher level databases, so do the protein level databases, with respect to protein sequence.

Recently, several protein level databases were merged into one primary protein resource, the Universal Protein Resource (UniProt). UniProt combines the Swiss-Prot, TrEBML, and PIR-PSD databases into one resource, consisting of three related databases. The UniProt Archive (UniParc) is a database of non-redundant protein sequences obtained from 1) translation of sequences within the gene sequence level databases (e.g., GenBank), 2) RefSeq, 3) FlyBase, 4) WormBase, 5) Ensembl, 6) the International Protein Index, 7) patent applications, and 8) the Protein Data Bank (Bairoch, *et al.*, 2005). The UniProt Knowledgebase (UniProt) provides functional annotation of the sequences within the UniParc. Examples of the annotation include the

protein name, listing of protein domains and families from the InterPro database (Mulder, *et al.*, 2003), Enzyme Commission identifier, and Gene Ontology identifiers. Proteins represented within the UniParc and UniProt Knowledgebase are then gathered automatically to create the UniProt reference database,(UniRef) a database of reference, exemplar sequences based on sequence identity. Three different versions of the UniRef database exist, the UniRef100, UniRef90, and UniRef50, where the number denotes the percent identity required for sequences to be merged, from across all species represented in the parent databases, together into a single reference protein sequence. Thus, the UniRef50 requires only 50% identity for proteins to be merged together. The UniRef50 and 90 databases provide faster sequence searches for identifying probable protein domains and functions by decreasing the size of the search space.

The RefSeq database also contains reference protein sequences, similar in concept to the reference mRNA sequences. These are available through the Entrez Gene system, when querying for a gene. For more information on RefSeq, see the section on Sequence level databases.

Protein Interaction Databases

Protein interaction databases capture data on the interaction of proteins with other proteins, genes, and small molecules. The two protein interaction databases discussed here include the Biomolecular Interaction Network Data (BIND) and the Database of Interacting Proteins (DIP), however others include the Molecular Interaction database (MINT), and the IntAct database. Tools are also available to view the networks, such as Osprey (Breitkreutz, *et al.*, 2003) and Cytoscape (Shannon, *et al.*, 2003). By visualizing

the interaction data, with some notion of the gene expression data, investigators can begin to construct hypotheses to test mechanistic understandings.

Both the BIND (Alfarano, *et al.*, 2005) and DIP (Xenarios, *et al.*, 2000) databases manage data from protein interaction experiments, including yeast-two-hybrid and co-immunopreciptation experiments. Much of this data is submitted to the databases either directly or through database curators scouring the literature. The databases make their data available through interaction files which are typically available in the Protein Standards Initiative (PSI) Molecular Interaction (PSI-MI) XML format.

Visualization of these datasets is made possible through tools such as Osprey and Cytoscape. Both of these tools produce protein interaction networks based on input data, which may be from either of these databases, or from other sources. Cytoscape has the additional functionality of allowing users to input their gene expression data for overlay on the protein interaction map, through plug-ins (Shannon, *et al.*, 2003). Through these visualization tools, investigators may begin to identify pathways of interest that are putatively altered following treatment, facilitating the generation of new hypotheses, or identification of new drug targets.

Microarray Databases

Microarray databases typically come in two forms: 1) laboratory information management systems (LIMS), and 2) data repositories. The LIMS solutions are used on the local level to manage data within a laboratory or a consortium. The primary purposes of the LIMS are to ensure data are being properly managed, facilitate analysis, and archive data for long-term use. Data repositories serve to facilitate comparisons between

datasets from across laboratories, facilitate reanalysis of data, and complement interpretation from studies in non-genomic laboratories.

A general discussion of microarray LIMS is difficult as they are designed by groups to meet their individual needs. Typically they are developed based on the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma, *et al.*, 2001). This standard discusses the need for genomics investigators to include with their data enough information for other scientists to replicate the experimental protocols.

Data repositories typically follow the same basic philosophy: collect data submitted by investigators, process it to make it usable by others, and provide some means for comparison to the rest of the repository. Several journals require microarray submissions to adhere to the MIAME standard, while the MGED Society is pushing for journals to require microarray datasets be submitted to repositories as a condition of publication, similar to requirements that novel sequences be submitted to GenBank prior to publication (Ball, *et al.*, 2004a; Ball, *et al.*, 2004b). Two of the most common data repositories are the NCBI Gene Expression Omnibus (GEO) (Edgar, *et al.*, 2002) and the ArrayExpress (Brazma, *et al.*, 2003; Rocca-Serra, *et al.*, 2003) at the European Bioinformatics Institute (EBI). These serve as general repositories, capable of handling most gene expression data. Recently, specialized repository efforts have been undertaken, such as the Chemical Effects in Biological Systems (CEBS) Knowledgebase (Waters, *et al.*, 2003a; Waters, *et al.*, 2003b), which will serve to catalogue gene expression data from chemical exposures with the associated pathology data.

With the emergence of more pharmacology and toxicology domain specific LIMS for genomics, the International Life Sciences Institute (ILSI) Health and Environmental

Sciences Institute (HESI) Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment, in cooperation with the MGED Society, began work on a toxicology-specific MIAME standard (MIAME/Tox) (Mattes, *et al.*, 2004). This MIAME/Tox document is expected to further specify the minimum information that needs to be communicated to replicate a toxicogenomics experiment. It is expected that this document will facilitate data sharing among the toxicogenomics community (e.g., transmission of data from a toxicology LIMS for inclusion in CEBS).

Conclusion

It is evident that pharmacology and toxicology benefit from large scale omic technologies. Experimentally perturbing a system, and identifying changes in gene expression, may result in the generation of novel mechanistic hypotheses (Nuwaysir, *et al.*, 1999). However, to effectively harness genomic technologies it is also necessary to develop data management technologies to manage the massive amounts of data as they are generated. Through these efforts, improved quality assurance protocols and experimental designs may develop. These improvements, coupled with the maturation data annotation landscape, will ultimately lead to more informative mechanistic hypotheses.

Literature Cited

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. and Hogue, C. W. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res, **33 Database Issue**, D418-24.

Bader, G. D. and Hogue, C. W. (2000) BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics, **16**, 465-77.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res, **33 Database Issue**, D154-9.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S. A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J. and Winegarden, N. (2004a) Submission of microarray data to public repositories. PLoS Biol, **2**, E317.

Ball, C. A., Sherlock, G. and Brazma, A. (2004b) Funding high-throughput data sharing. Nat Biotechnol, **22**, 1179-83.

Boverhof, D. R., Fertuck, K. C., Burgoon, L. D., Eckel, J. E., Gennings, C. and Zacharewski, T. R. (2004) Temporal and dose-dependent hepatic gene expression changes in immature ovariectomized mice following exposure to ethynyl estradiol. Carcinogenesis, **25**, 1277-91.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, **29**, 365-71.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. and Sansone, S. A. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, **31**, 68-71.

Breitkreutz, B. J., Stark, C. and Tyers, M. (2003) Osprey: a network visualization system. Genome Biol, **4**, R22.

Cheadle, C., Vawter, M. P., Freed, W. J. and Becker, K. G. (2003) Analysis of microarray data using Z score transformation. J Mol Diagn, **5**, 73-81.

Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet, **32 Suppl**, 490-5.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Birney, E. (2003) Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res, **31**, 38-42.

Cox, C. (1999) Nietzsche: Naturalism and Interpretation, University of California Press, Berkeley, California.

Cox, W. G., Beaudet, M. P., Agnew, J. Y. and Ruth, J. L. (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. Anal Biochem, **331**, 243-54.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. and Clamp, M. (2004) The Ensembl automatic gene annotation system. Genome Res, **14**, 942-50.

Dobbin, K., Shih, J. H. and Simon, R. (2003) Statistical design of reverse dye microarrays. Bioinformatics, **19**, 803-10.

Dobbin, K. and Simon, R. (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics, **18**, 1438-45.

Dombkowski, A. A., Thibodeau, B. J., Starcevic, S. L. and Novak, R. F. (2004) Gene-specific dye bias in microarray reference designs. FEBS Lett, **560**, 120-4.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) Pattern Classification.

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004a) Empirical bayes gene screening tool for time-course or dose-response microarray data. J Biopharm Stat, **14**, 647-70.

Eckel, J. E., Gennings, C., Therneau, T. M., Boverhof, D. R., Burgoon, L. D. and Zacharewski, T. R. (2004b) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, in press.

Eckel, J. E., Gennings, C., Therneau, T. M., Burgoon, L. D., Boverhof, D. R. and Zacharewski, T. R. (2005) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, **21**, 1078-83.

Edgar, R., Domrachev, M. and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, **30**, 207-10.

Efron, B. and Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol, **23**, 70-86.

Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A., Anagnostopoulos, A., Baldarelli, R. M., Baya, M., Beal, J. S., Bello, S. M., Boddy, W. J., Bradt, D. W., Burkart, D. L., Butler, N. E., Campbell, J., Cassell, M. A., Corbani, L. E., Cousins, S. L., Dahmen, D. J., Dene, H., Diehl, A. D., Drabkin, H. J., Frazer, K. S., Frost, P., Glass, L. H., Goldsmith, C. W., Grant, P. L., Lennon-Pierce, M., Lewis, J., Lu, I., Maltais, L. J., McAndrews-Hill, M., McClellan, L., Miers, D. B., Miller, L. A., Ni, L., Ormsby, J. E., Qi, D., Reddy, T. B., Reed, D. J., Richards-Smith, B., Shaw, D. R., Sinclair, R., Smith, C. L., Szauter, P., Walker, M. B., Walton, D. O., Washburn, L. L., Witham, I. T. and Zhu, Y. (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res, **33**, D471-5.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y. and Wang, Y. (2003) Effects of atmospheric ozone on microarray data quality. Anal Chem, **75**, 4672-5.

Fielden, M. R., Halgren, R. G., Dere, E. and Zacharewski, T. R. (2002a) GP3: GenePix post-processing program for automated analysis of raw microarray data. Bioinformatics, **18**, 771-3.

Fielden, M. R., Halgren, R. G., Fong, C. J., Staub, C., Johnson, L., Chou, K. and Zacharewski, T. R. (2002b) Gestational and lactational exposure of male mice to diethylstilbestrol causes long-term effects on the testis, sperm fertilizing ability in vitro, and testicular gene expression. Endocrinology, **143**, 3044-59.

Fisher, R. A. (1962) The place of the design of experiments in the logic of scientific inference. Colloq. Int. Cent. Nat. Recherche Scientifique, **110**, 13-19.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC, Boca Raton, FL.

Glas, A. M., Kersten, M. J., Delahaye, L. J., Witteveen, A. T., Kibbelaar, R. E., Velds, A., Wessels, L. F., Joosten, P., Kerkhoven, R. M., Bernards, R., van Krieken, J. H., Kluin, P. M., van't Veer, L. J. and de Jong, D. (2005) Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. Blood, **105**, 301-7.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, **286**, 531-7.

Grant, G. R., Manduchi, E., Pizarro, A. and Stoeckert, C. J., Jr. (2003) Maintaining data integrity in microarray data management. Biotechnol Bioeng, **84**, 795-800.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V. A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, **30**, 52-5.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, **32**, D258-61.

Hood, L. and Perlmutter, R. M. (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol, **22**, 1215-7.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey,

R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Birney, E. (2005) Ensembl 2005. Nucleic Acids Res, **33 Database Issue**, D447-53.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet, **2**, 343-72.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D. and Kent, W. J. (2003) The UCSC Genome Browser Database. Nucleic Acids Res, **31**, 51-4.

Kavlock, R. J., Ankley, G., Blancato, J., Collete, T., Francis, E., Gray, E., Hammerstrom, K., Swartout, J., Tilson, H., Toth, G., Veith, G., Weber, E., Wolf, D. and Young, D. A. (2003) A framework for computational toxicology research in ORD.

Kerr, M. K. and Churchill, G. A. (2001a) Experimental Design for Gene Expression Microarrays. Biostatistics, **2**, 183-201.

Kerr, M. K. and Churchill, G. A. (2001b) Statistical design and the analysis of gene expression microarray data. Genet Res, **77**, 123-8.

Kristensen, V. N., Sorlie, T., Geisler, J., Langerod, A., Yoshimura, N., Karesen, R., Harada, N., Lonning, P. E. and Borresen-Dale, A. L. (2005) Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: clinical implications. Clin Cancer Res, **11**, 878s-83s.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. and Davis, R. W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A, **94**, 13057-62.

Lindsay, M. A. (2003) Target discovery. Nat Rev Drug Discov, **2**, 831-8.

Luyendyk, J. P., Mattes, W. B., Burgoon, L. D., Zacharewski, T. R., Maddox, J. F., Cosma, G. N., Ganey, P. E. and Roth, R. A. (2004) Gene expression analysis points to hemostasis in livers of rats cotreated with lipopolysaccharide and ranitidine. Toxicol Sci, **80**, 203-13.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, **33 Database Issue**, D54-8.

Mattes, W. B., Pettit, S. D., Sansone, S. A., Bushel, P. R. and Waters, M. D. (2004) Database development in toxicogenomics: issues and efforts. Environ Health Perspect, **112**, 495-505.

McKusick, V. A. (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edition. Johns Hopkins University Press, Baltimore, MD.

Mischel, P. S., Cloughesy, T. F. and Nelson, S. F. (2004) DNA-microarray analysis of brain cancer: molecular classification for therapy. Nat Rev Neurosci, **5**, 782-92.

Model, F., Konig, T., Piepenbrock, C. and Adorjan, P. (2002) Statistical process control for large scale microarray experiments. Bioinformatics, **18 Suppl 1**, S155-63.

Moggs, J. G., Tinwell, H., Spurway, T., Chang, H. S., Pate, I., Lim, F. L., Moore, D. J., Soames, A., Stuckey, R., Currie, R., Zhu, T., Kimber, I., Ashby, J. and Orphanides, G. (2004) Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth. Environ Health Perspect, **112**, 1589-606.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R. and Zdobnov, E. M. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res, **31**, 315-8.

Nicholson, J. K. and Wilson, I. D. (2003) Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov, **2**, 668-76.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C. and Afshari, C. A. (1999) Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog, **24**, 153-9.

Paules, R. (2003) Phenotypic anchoring: linking cause and effect. Environ Health Perspect, **111**, A338-9.

Pruitt, K. D. and Maglott, D. R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res, **29**, 137-40.

Qin, L. X. and Kerr, K. F. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res, **32**, 5471-9.

Quackenbush, J. (2002) Microarray data normalization and transformation. Nat Genet, **32 Suppl**, 496-501.

Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Contrino, S., Vilo, J., Abeygunawardena, N., Mukherjee, G., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A. and Sansone, S. A. (2003)

ArrayExpress: a public database of gene expression data at EBI. C R Biol, **326**, 1075-8.

Ross, J. S., Schenkein, D. P., Pietrusko, R., Rolfe, M., Linette, G. P., Stec, J., Stagliano, N. E., Ginsburg, G. S., Symmans, W. F., Pusztai, L. and Hortobagyi, G. N. (2004) Targeted therapies for cancer 2004. Am J Clin Pathol, **122**, 598-609.

Rouchka, E. C., Gish, W. and States, D. J. (2002) Comparison of whole genome assemblies of the human genome. Nucleic Acids Res, **30**, 5004-14.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, **270**, 467-70.

Selvanayagam, Z. E., Cheung, T. H., Wei, N., Vittal, R., Lo, K. W., Yeo, W., Kita, T., Ravatn, R., Chung, T. K., Wong, Y. F. and Chin, K. V. (2004) Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling. Cancer Genet Cytogenet, **154**, 63-6.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, **13**, 2498-504.

Shih, J. H., Michalowska, A. M., Dobbin, K., Ye, Y., Qiu, T. H. and Green, J. E. (2004) Effects of pooling mRNA in microarray class comparisons. Bioinformatics, **20**, 3318-25.

Simon, R., Radmacher, M. D. and Dobbin, K. (2002) Design of studies using DNA microarrays. Genet Epidemiol, **23**, 21-36.

Tempelman, R. J. (2005) Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol, **105**, 175-86.

Townsend, J. P. (2003) Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics, **4**, 41.

Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P. and Cho, K. W. (2002) Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic Acids Res, **30**, e54.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, **98**, 5116-21.

Ulrich, R. and Friend, S. H. (2002) Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov, **1**, 84-8.

Vinciotti, V., Khanin, R., D'Alimonte, D., Liu, X., Cattini, N., Hotchkiss, G., Bucca, G., De Jesus, O., Rasaiyaah, J., Smith, C. P., Kellam, P. and Wit, E. (2004) An experimental evaluation of a loop versus a reference design for two-channel microarrays. Bioinformatics,

Waters, M., Boorman, G., Bushel, P., Cunningham, M., Irwin, R., Merrick, A., Olden, K., Paules, R., Selkirk, J., Stasiewicz, S., Weis, B., Van Houten, B., Walker, N. and Tennant, R. (2003a) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics, **111**, 15-28.

Waters, M. D. and Fostel, J. M. (2004) Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet, **5**, 936-48.

Waters, M. D., Olden, K. and Tennant, R. W. (2003b) Toxicogenomic approach for assessing toxicant-related disease. Mutat Res, **544**, 415-24.

Weinshilboum, R. and Wang, L. (2004) Pharmacogenomics: bench to bedside. Nat Rev Drug Discov, **3**, 739-48.

Wetmore, B. A. and Merrick, B. A. (2004) Toxicoproteomics: proteomics applied to toxicology and pathology. Toxicol Pathol, **32**, 619-42.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A. and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res, **32**, D35-40.

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 expression data via mixed models. J Comput Biol, **8**, 625-37.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H. H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol, **3**, research0048.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000) DIP: the database of interacting proteins. Nucleic Acids Res, **28**, 289-91.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J. and Quackenbush, J. (2002a) Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol, **3**, research0062.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002b) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, **30**, e15.

Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. Nat Rev Genet, **3**, 579-88.

Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J. and Golub, T. (2001) Molecular classification of multiple tumor types. Bioinformatics, **17 Suppl 1**, S316-22.

CHAPTER TWO

Burgoon, L.D., Boutros, P.C., Dere, E., and Zacharewski, T.R. (2005) dbZach: A MIAME-Compliant Toxicogenomic Supportive Relational Database. In submission.

Abstract

The comprehensive elucidation of mechanisms of toxicity that will support mechanistically-based quantitative risk assessment requires the development of innovative computational infrastructure and approaches including enterprise data management systems in order to facilitate data integration and reduce uncertainties in the source-to-outcome continuum. dbZach (http://dbzach.fst.msu.edu) is a modular relational database with associated data insertion, retrevial and mining tools that manages toxicogenomic and traditional toxicology data, and facilitates data integration, analysis, and sharing between collaborating investigators or with public repositories. It consists of four Core Subsystems (i.e. Clones, Genes, Sample Annotation and Protocols), four Experimental Subsystems (i.e. Microarray, Affymetrix, Real-Time PCR (RTPCR), and Toxicology), and three Computational Subsystems (i.e. Gene Regulation, Pathways, Orthology) that are supportive of the Minimum Information About a Microarray Experiment (MIAME) standard. Its modular structure allows data management to be extended to other emerging technologies and model systems including ecologically relevant species. dbZach provides daily ongoing support for a number of *in vivo* and *in vitro* toxicogenomic microarray studies and is currently populated with human, mouse and rat data. The source code will be made available for examination and implementation to interested parties under license. The terms of the license are currently being developed. All distribution and licensing information will be made available on the dbZach website (http://dbzach.fst.msu.edu).

Introduction

In order to improve the quantitative risk assessment of chronic and subchronic exposure to synthetic and natural chemicals and their complex mixtures, uncertainties within the source-to-outcome continuum must be minimized. Emerging technologies and computational toxicology approaches will provide informative mechanistic data needed to further improve quantitative predictive models. However, disparate data including chemical, exposure, adsorption, distribution, metabolism, excretion, toxicologic and omic data, must be integrated in order to develop comprehensive computational models that consider all of the available data. The development of enterprise data management systems is an integral step to support emerging computational toxicology methods and to facilitate revision of mechanistically-based quantitative risk assessment.

Relational databases and knowledgebases capable of supporting toxicology and quantitative risk assessment efforts are emerging (Bushel, *et al.*, 2001; Mattes, *et al.*, 2004; Tong, *et al.*, 2003; Waters, *et al.*, 2003). In addition to ensuring proper data management and storage, relational databases facilitate data quality assurance, analysis, sharing and deposition into public repositories. Furthermore, they provide a platform for complex queries across disparate data and support the development and use of data mining applications. For example, properly designed relational databases may prove to be indispensable in the reevaluation of historical data in light of new results, identify relationships across several different data domains (e.g., gene expression, metabolite, gross observations, histopathology) to identify predictive agglomerative biomarkers with greater predictive accuracy, and identify orthologous response genes between model and ecologically relevant species to reveal conserved mechanisms of toxicity.

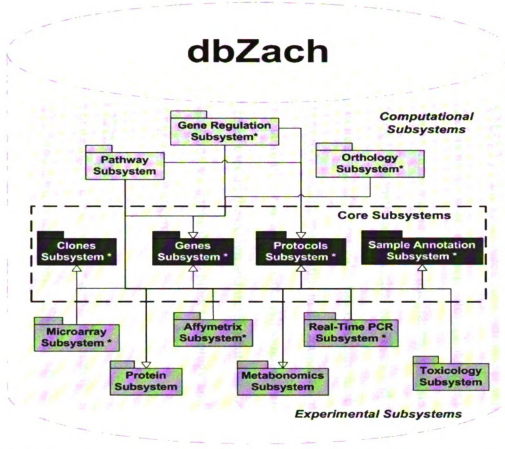

Figure 2-1: Subsystem Interactions. The dbZach database is divided into three primary modules (Core Subsystems, Experimental Subsystems, Computational Subsystems), each consisting of several subsystems. All of the Core Subsystems are required for the database to function properly as each subsystems within dbZach connects to at least one of the tables within this module. The asterisk denotes a populated subsystem.

Here, we discuss the dbZach System, a database and associated software suite that was developed to support data management of ongoing traditional toxicology (e.g., histopathology, clinical chemistry/pathology, gross observations) and toxicogenomic (i.e., genomics, proteomics, and metabonomics) studies (Figure 1). It is supportive of the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma, *et al.*, 2001), and the Microarray and Gene Expression (Spellman, *et al.*, 2002) Markup Language (MAGE-ML) which facilitates the electronic sharing of stored data with other databases. dbZach is not a public data repository, but rather an intralaboratory framework for the storage, management, integration, analysis, and mining of data including toxicology, histopathology, clinical chemistry and microarray data. Data integration facilitated by dbZach provides infrastructure for building computational toxicology tools to reduce the uncertainties in the source-to-outcome continuum associated with quantitative risk assessment. Although developed to support our research efforts, the schemas used in the design and implementation of the database and its associated applications are applicable to other toxicology and biomedical research programs requiring data management.

Database Design

dbZach is designed to be modular and to accurately reflect biological concepts and relationships. Modularity ensures new subsystems for nascent technologies can be incorporated without requiring changes to the preexisting backend. Each separate module of the database is termed a subsystem, and each subsystem manages data for a technology (e.g., quantitative Real-Time PCR, spotted microarray, Affymetrix), a

biological concept/discipline (e.g., cDNA clones, genes, toxicology, pathway, gene regulation), or MIAME required ancillary annotation (e.g., protocols, sample annotation).

Relationships between tables representing definitive biological concepts (e.g., animals and organs) are structured to capture their biological relationships (Figure 2). For example, the animal table records only that data which is specific to the animal itself, such as its arrival date, age at arrival, sex, and the cage identifier. A separate table records information about harvested organs, such as the organ name, wet/dry weights, etc. The two tables are connected through a one-to-many relationship, where one animal may have data from one or more organs associated with it.

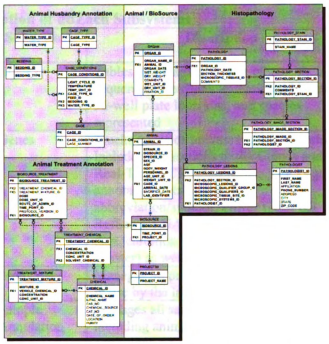

Figure 2-2: Relationships Between Animal Husbandry, Treatment, and Histopathology. Animal husbandry and treatment data are defined in different sets of tables. As cages may hold more than one animal, a one-to-many relationship exists between the CAGE and ANIMAL tables. This separation allows data specific to the cage (e.g., bedding, feed type, water type) to be separated from the animal. Similar logic follows for treatments, and histopathology data. Relationships between tables are depicted using the crow's feet symbols (the line symbols between tables) , where the parent table (e.g., CAGE) is represented with either a double line symbol (required relationship), or a circle with a cross symbol (not required relationship), and the child table (e.g., ANIMAL) is represented with a crow's foot (a circle with prongs leading from it). In the one-to-many relationship, there is one parent that may contain many children (e.g., one cage may contain many animals). In practice, the one-to-many (i.e., parent-to-child) relationship is realized through a primary key (i.e., unique identifier from the parent table) to foreign key relationship (e.g., the CAGE_ID in the CAGE table is the primary key, while the CAGE_ID in the ANIMAL table is a foreign key).

Another example of the persistence of biological relationships within the database is the management of histopathology data. Animals may consist of many organs within the database. Organs may exist as an agglomeration of sections. Each section may be scored by a pathologist, and a lesion may be identified on a per section basis. Pathologist scores and remarks concerning each lesion are related back to the animal annotation through the section and organs. Thus, chemical treatment/exposure annotation is not

Table 2-1: Description of dbZach Subsystems

Database Subpart	Subsystem	Description	Status*
Core	Clones	Tables that manage cDNA clones, their sequences, GenBank accessions, 384- and 96-well plate locations, and species represented	Populated
	Genes	Tracks genes of interest for the laboratory. Includes connections to the Clones and Real-Time PCR subsystems. Tracks gene annotation data (name, abbreviation/symbol, Entrez Gene ID, RefSeq mRNA Accession, Gene Ontology data, chromosomal location, UniGene Cluster) as well	Populated
	Protocols	Manages all protocols, and their versions, in use by the laboratory.	Populated
	Sample Annotation	Manages all sample annotation data including animal husbandry, cell culture conditions, organs, biological fluids, and biological sample type.	Populated
Experimental	Real-Time PCR	Manages primers, primer sets, PCR plates, and data from quantitative real-time PCR reactions.	Populated
	Microarray	Manages labeled extracts, microarrays, clones and feature locations, quality data, raw image files, quantified data, normalized and statistical data.	Populated
	Toxicology	Manages toxicology and pathology data.	Testing
	Protein	Manages protein annotation and proteomic data	Development
	Metabonomic	Manages metabolite annotation and metabonomic data	Development
	Affyemtrix	Manages all Affymetrix data generated from an experiment	Populated
Computational	Gene Regulatory	Manages sequences upstream of gene start sites, gene regulatory sequences, and their annotation	Populated
	Orthology	Manages orthologous gene relationships between species	Populated
	Pathways	Manages known and newly discovered pathways by modeling the relationships between endogenous and exogenous chemicals, proteins, and genes	Testing

* Populated: subsystems ready for querying. Testing: subsystem tables have been put in place, and are being tested to ensure all appropriate relationships are captured. Development: the database subsystem is currently being developed.

provided at the histopathology level, but rather the animal. This allows any data

associated with the animal to also be associated with experimental manipulations performed on the animal (e.g., treatment, surgeries, husbandry), and optimizes database design and performance by reducing redundancy where this same information would be associated with each experimental level (e.g., histopathology, clinical chemistry, gross observations, etc) individually.

Currently the database consists of 13 subsystems divided in four Core Subsystems (i.e. Clones, Genes, Sample Annotation and Protocols), six Experimental Subsystems (i.e. Microarray, Affymetrix, Real-Time PCR (RTPCR), Toxicology, Metabonomics, and Proteomics), and three Computational Subsystems (i.e. Gene Regulation, Pathway, Orthology). A brief description of each subsystem and its current status is summarized in Table 2-1. More detailed descriptions of each subsystem are provided below.

Core Subsystems

The Core Subsystems include the Clones, Genes, Protocols and Sample Annotation Subsystems. These subsystems satisfy MIAME requirements and are needed for the functionality of the Experimental and Computational Subsystems.

The Clones Subsystem consists of tables that manage the cDNA clones represented on microarrays. Each clone is associated with a GenBank accession number. The mapping of a clone to a GenBank record is a one-to-many relationship to account for multiple high probability BLAST matches (Figure 2-3). The subsystem also relates a clone to its location within 96- and/or 384-well storage plates.

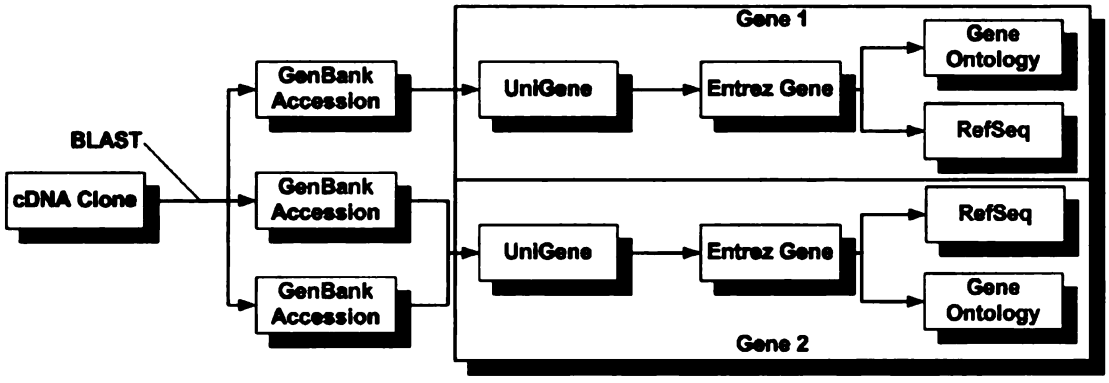

Figure 2-3: cDNA Clones to Gene Annotation. BLAST analysis of sequenced ESTs represented on the microarrays may match multiple, high probability GenBank Accession numbers. GenBank Accession numbers are mapped to genes using the UniGene and Entrez Gene databases. Each GenBank Accession is mapped to only one UniGene record, however, not all of these accessions map to the same gene.

The Genes Subsystem manages gene annotation data for genes associated with cDNA clones, real-time PCR primers, and pathways. Annotation data include chromosomal locations, Gene Ontology data, NCBI Entrez Gene and RefSeq identifiers, and NCBI UniGene Cluster numbers. GenBank accessions are associated with gene records through the UniGene database. As a clone may be represented by more than one GenBank accession, it is also possible for a clone to map to many genes (Figure 2-3). dbZach updates clone-gene relationships following each UniGene build based on GenBank Accession relationships with GenBank and cross-references to the Entrez Gene database.

In compliance with the MIAME standards, all protocols and standard operating procedures used within the laboratory are managed by the Protocols Subsystem. Changes to existing protocols are maintained as separate versions. Protocol versions are associated with data from all of the various subsystems, and may be used in analyses as appropriate to investigate differences in methods, and to examine the effect of protocol variations on data.

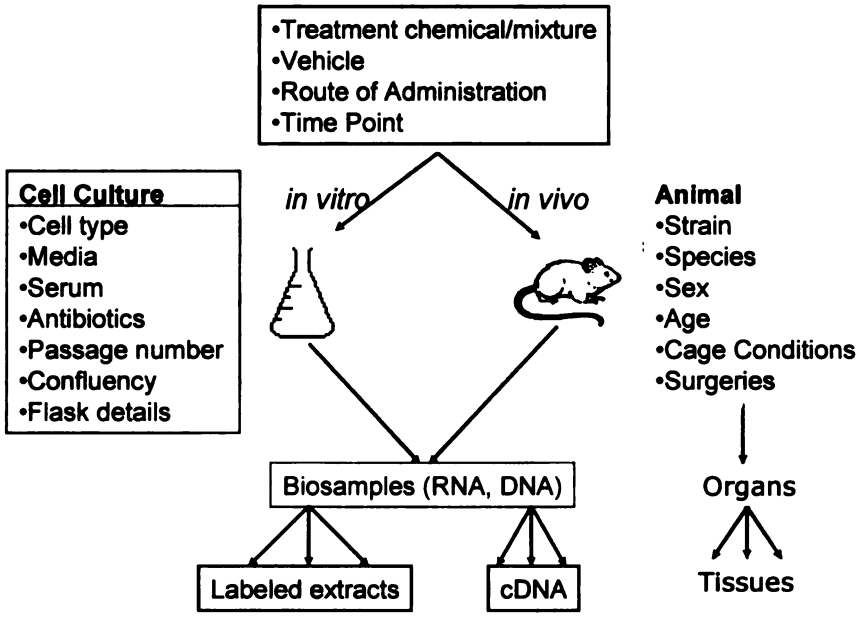

Figure 2-4: *In Vivo* and *In Vitro* Sample Annotation Tracks. The management of *in vivo* and *in vitro* sample annotation data is tracked separately. This minimizes table sizes which improves efficiency, and allows for more complete annotations to be tracked. For example, animals are not grown within a medium, nor are surgeries performed upon a cell culture sample. Yet, both categories of information are necessary for complete sample annotation, and not appropriately captured using one large table for both *in vivo* and *in vitro* data.

The Sample Annotation subsystem manages all biological sources of data, such as animals and cell culture samples. Unlike other database efforts that ascribe to the MIAME standards, dbZach manages in vitro and in vivo information about biological sources in different tracks, allowing more detailed information to be managed while decreasing the number of columns per table, thus enhancing technical efficiency and simplifying the annotation used for describing study designs and experimental conditions (Figure 2-4).

Experimental Subsystems

The Microarray Subsystem can manage *n*-channel (i.e., there is no limit on the number of concurrent dyes per microarray set by the database) microarray data including

cDNA and oligonucleotide platforms utilizing any number of fluorescent dyes. Currently, the subsystem manages the raw TIFF microarray images from the array scanner, quality control data, quantified data from the image, normalized data, and the statistical analysis output used in the identification of significant/active genes. This subsystem leverages relationships with other tables within the Sample Annotation, Clones, and Protocols subsystems to provide additional annotation.

The Real-Time PCR Subsystem manages critical data for performance of quantitative real-time PCR (QRT-PCR) assays. These include the sequence of forward and reverse primers, including the central probe for TaqMan assays, the layout of assay plates, raw data files from the assay equipment, and expression data. The primers are associated with the template used for their design, and also with the Genes Subsystem, to provide up-to-date gene annotation data. This also facilitates *in silico* comparisons to determine the correlation between the microarray and QRT-PCR gene expression data (Boverhof, *et al.*, 2005; Fong, *et al.*, 2005).

The Toxicology Subsystem is responsible for the management of all traditional toxicology data, including histopathology, *in vitro* assays, clinical chemistry, and cell morphology. The Toxicology Subsystem currently uses the National Toxicology Program Pathology Code as the controlled vocabulary for pathology data. Toxicology data are associated with the source organisms in the Sample Annotation subsystem allowing toxicology parameters to be tracked back to specific animals and/or treatment conditions.

Computational Subsystems

The Computational subsystems include the Orthology, Gene Regulatory and Pathway Subsystems. These subsystems manage data that facilitate mechanistic interpretation, cross-species comparison, network elucidation, and quantitative risk assessment. Tables within these subsystems are populated with data that have been generated using computational means.

The Orthology Subsystem facilitates species comparisons. Currently, the subsystem catalogues orthologous genes across human, mouse, and rat species. The architecture allows other species, such as ecologically relevant models, to be incorporated as necessary. Orthology data are currently obtained from the Ensembl database, however, the architecture also enables data from other sources to be used, including the Comparative Toxicogenomic Database (http://www.niehs.nih.gov/oc/factsheets/ctd.htm). This facilitates comparisons among databases, as well as maximizing the identification of orthologous genes in an effort to identify conserved responses and mechanisms of toxicity between species thus minimizing uncertainties associated with extrapolations.

To facilitate the generation of new hypotheses concerning gene regulation with respect to the gene expression data, genomic sequence for likely regulatory regions (proximal and distal promoters, 5'untranslated, 3' untranslated) for all RefSeq identified genes from the UCSC Genome Browser have been included in the Gene Regulatory Subsystem. When leveraged with position weight matrices and probabilistic short sequence motif identification, this data allows novel hypotheses concerning gene regulation to be computationally examined for regulatory motifs (Sun, *et al.*, 2004) and subsequently verfied using chromatin immunoprecipitation (ChIP) and ChIP-on-chip

technologies (Buck and Lieb, 2004) that facilitates the development of comprehensive gene expression regulatory networks (Luscombe, *et al.*, 2004).

The Pathways Subsystem catalogs all protein-protein, protein-DNA, chemical-protein, chemical-DNA, and metabolism pathways. The current schema will handle data from other databases, and those pathways that are internally defined. For example, two-hybrid data from data repositories such as BIND (Bader and Hogue, 2000) and DIP (Xenarios, *et al.*, 2000) can be included in the database, as can data from the literature utilizing text mining algorithms, or pathways determined internally through pharmacological experiments.

Implementation

Platform Independent

The dbZach system was not designed for any particular RDBMS or operating system. The database and tools have been tested under the Oracle 9*i* and IBM DB2 database engines, and the Java2 v 1.5.0 runtime environment (JRE). The database schema may be implemented on any platform; however, the tools require a database that is compatible with the Java Database Connectivity (JDBC) package. Some of the input tools require users to populate template Microsoft (MS) Excel files which may be accomplished using either MS Excel or the open source office/productivity software, OpenOffice (http://www.openoffice.org) which runs on most major operating systems, including Linux, Windows, and Mac OS X.

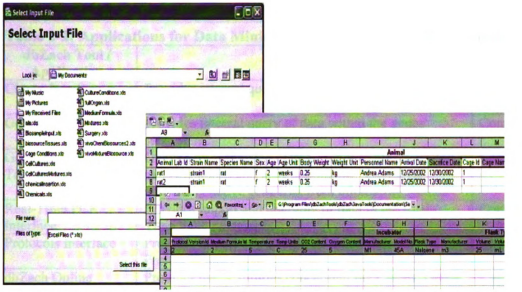

Figure 2-5: Bulk Data Insertion. Bulk data insertions are accomplished using a template spreadsheet (MS Excel format). Submitters fill out the spreadsheet with all of the necessary information, and select it using the Select Input File graphical user interface (GUI). This mechanism is used for uploading Sample Annotation, Microarray, and Toxicology data. Shown here are examples from the Sample Annotation Interface.

Bulk Data Insertions

Data are inserted into dbZach using template spreadsheet files written in MS Excel (Figure 2-5). Spreadsheets were chosen over more cumbersome graphical user interfaces to facilitate bulk uploads, and build upon user familiarity with spreadsheets. Furthermore, users can populate template spreadsheets by simply cutting-and-pasting data from one sheet to another.

An additional advantage of spreadsheets is that they simplify the numerous one-to-many relationships present within the data. For example, it is far easier to visualize and enter data from one-to-many relationships in a single spreadsheet than in most graphical user interfaces. Users can easily copy data from cell-to-cell in a spreadsheet, decreasing user-based errors, such as typographical errors, for GUI input fields, or mouse-click errors in the case of GUI combo boxes. Moreover, the use of spreadsheets

Table 2-2: Applications for Data Mining, Upload, and Interaction with dbZach

dbZach Tool / Application	Description	Status*
Clones Interface	Provides internal query of data within clones subsystem	Complete
Genes Interface	Provides internal query of data within the genes subsystem	Complete
Gene Annotation Tool (GAT)	Queries dbZach using cDNA clones to obtain related gene mappings and gene annotation. Useful for annotating active gene lists from microarray studies	Complete
Real-Time PCR Interface	Provides import and querying of data in the Real-Time PCR subsystem.	Complete
Protocols Interface	Provides import of protocol information to the Protocols subsystem	Complete
dbZach Online	Provides query abilities concerning the Clones, Genes, primers, and protocols in dbZach	Complete
Microarray Interface	Provides microarray data import capabilities	Complete
Pathology Interface	Provides pathology data import capabilities	Complete
Query Control Center	Tool that provides query capabilities for all data within the database	Complete / Development
Visualization Control Center	Tool that provides multiple biological data visualization capabilities in 2-D and 3-D, including pattern recognition	Complete / Developement
Toxicogenomics Correlation Tool (TCT)	Provides visualization capability for ontological pair (i.e., pairwise combination of genes, proteins, metabolites) expression and significance levels	Development
MAGE-ML Exporter	Tool that exports data within dbZach in MAGE-ML format for submission to data repositories	Development
Audit and Report Tool (ART)	Family of tools that generate audit and report tools for data submitted to dbZach	Complete / Development

* Complete: tools and applications that are complete and have met the quality standards of the laboratory. Complete / Development: plug-in capable applications with partial functionality available to the laboratory with new plug-ins in development. Development: tools and applications that are currently in development or testing.

also decreases the amount of time users spend interacting with the database, and away from the bench since data entered into spreadsheets can be uploaded without monitoring and continuous interaction.

Although efforts were made to minimize the complexity of data uploads, human error is always possible. To ensure data are entered into dbZach appropriately, a series of audit and report tools have been developed for investigators to double check their

uploaded data. This serves to minimize curatorial errors by individuals who are not familiar with the data, and has prevented the loss of time in analyzing incorrect data.

Database Querying

Users can interact with dbZach through the web, using dbZach applications and tools, or by using Structured Query Language (SQL) queries. Limited data are available through the web, which is accessed primarily by collaborators. These data include information regarding the current complement of genes represented on our cDNA microarrays, and primers available for real-time PCR analysis.

Most database queries performed by investigators within the lab occur through special interfaces and applications referred to as dbZach Tools. All of the tools have been written in Java2 and employ the use of the Swing library of classes for GUI development. Table 2-2 lists and summarizes the availability of current dbZach Tools and their primary functions.

Data Mining Applications

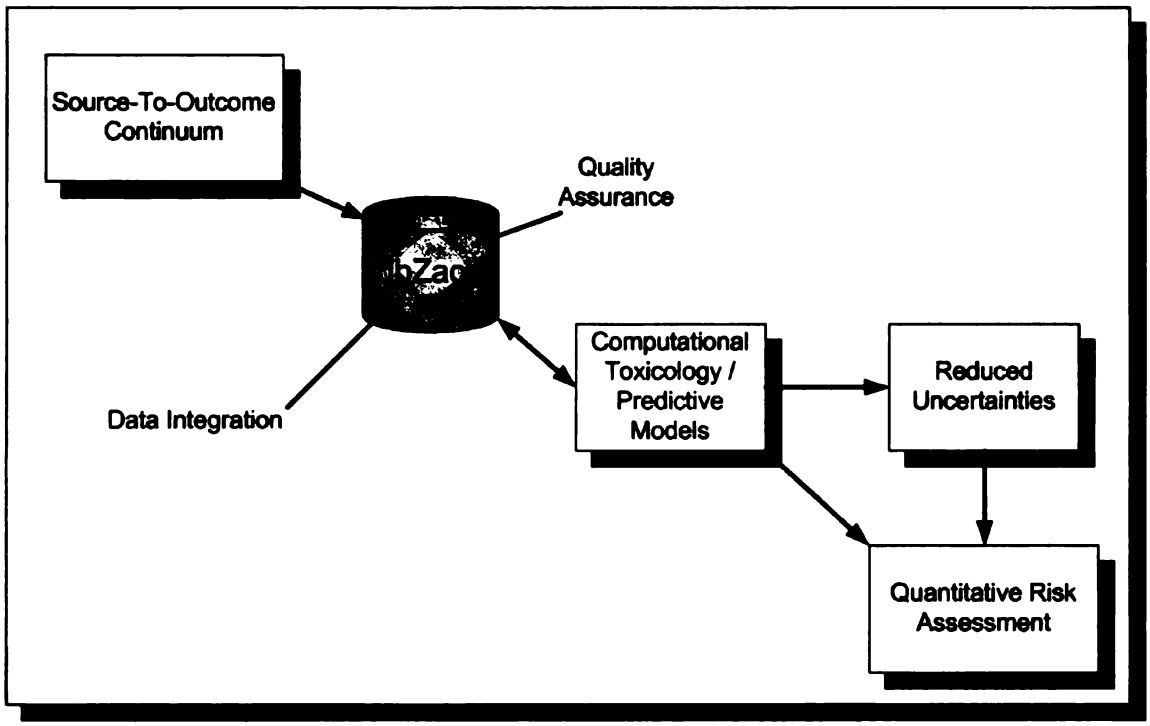

Figure 2-6: dbZach Facilitates Computational Toxicology Analysis of Source-To-Outcome Data. Data from the source-to-outcome continuum are managed within dbZach and subjected to quality assessment prior to analysis. Analyzed data is subsequently integrated using computational toxicology tools and predictive models to facilitate data interpretation and quantitative risk assessment. Comprehnsive analysis using computational tools support the development of predictive models reducing uncertainties within the source-to-outcome continuum, facilitating the development of more accurate quantitative risk and safety assessments.

In addition to the standard data query tools, a series of dbZach Data Mining Applications have been developed. Data mining involves identifying relationships and correlations within datasets. It is the first step towards data interpretation, where preexisting information is applied to data mining outcomes in order to generate new knowledge. Methods for identifying these relationships vary, and span everything from statistical analysis to data visualization. Current and future data mining applications are listed in Tables 2-1 and 2-2. All of these applications have been written in Java2, and may use the statistical language R.

Database Applications in Toxicogenomics

Database systems engender the goals of quantitative risk assessment through integration of data from the source-to-outcome continuum, and by facilitating the development of computational tools (Figure 2-6). Ultimately, these methods will reduce uncertainties associated with linkages in the source-to-outcome continuum, and facilitate the development of more accurate quantitative risk and safety assessments.

The dbZach system has been used within our laboratory to facilitate comparative experiments across species and toxicants, the generation of new visualization tools to facilitate data interpretation, the exploration of data sharing methodologies with collaborators and public repositories such as the Chemical Effects in Biological Systems (Waters, *et al.*, 2003) and Array Express (Brazma, *et al.*, 2003; Rocca-Serra, *et al.*, 2003), and the development of novel quality assurance methods to ensure consistent high quality data within and across studies. Each of these is briefly discussed below, and provide the foundation for further developments in computational and predictive toxicology.

Comparative Toxicogenomics

Comparative toxicogenomics is the comparison of toxicogenomic data between domains, such as chemicals, chemical classes, and species. Examples include comparing toxicogenomic responses within a chemical class to define a signature of response, and identifying conserved mechanistic responses across species, as well as identifying those genes that exhibit differential regulation between treatments and/or species that may provide information regarding differences in susceptibility to toxicity.

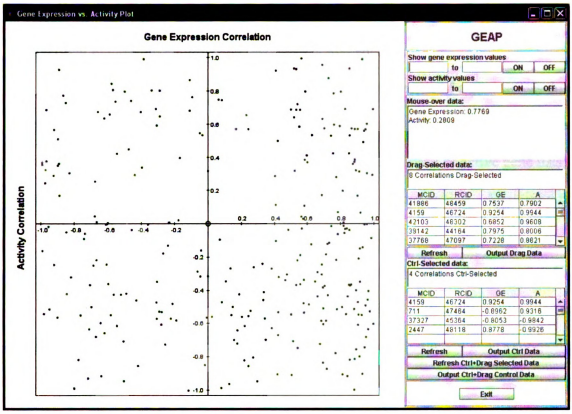

Figure 2-7: Orthologous Gene Expression and Activity Profiles. Active gene lists for mouse and rat uterine gene expression following the same treatment (i.e. 100 ug/kg ethynyl estradiol, oral gavage) were interrogated for orthologous gene pairs that exhibit comparable expression and temporal activity profiles. Pearson correlation coefficients were calculated for gene expression and temporal activity data for every pair of orthologous mouse and rat genes. The x-axis represents the correlation coefficients for gene expression, a measure of how well the temporal patterns parallel each other. The y-axis represents the correlation coefficients for activity, a measure of p1(t) similarity for each orthologous gene in the pair. Orthologous pairs that correlate well in both variables are represented in the first quadrant (upper right quadrant), while pairs that are inversely correlated, in both variables are represented in the third quadrant (lower left quadrant). Pairs that are poorly correlated in both variables will appear close to the origin of the graph.

The Orthology Subsystem catalogues orthologues between species, such as mouse and rat. This subsystem is connected to the Microarray Subsystem through relationships with the Genes Subsystems, and from there the Clones Subsystem. Thus, as microarray data are catalogued by clone identifiers within dbZach, gene expression data are easily compared across species. By calculating the correlation coefficient of gene expression

across the orthologous gene pairs, it is possible to identify genes with similar and divergent expression patterns and responses across species.

However, from gene expression data, the activity profile of a gene across dose or time can also be calculated. Error based statistics, such as the empirical Bayes posterior probability, are used to rank and prioritize genes for the creation of active gene lists; those genes that exhibit expression that is most different from vehicle. As activity is determined on a per time or dose basis, an activity profile is generated as a binary signature across the dose/time. Alternatively, the signature can also be represented as the probability values across dose/time. By correlating the activity response between the species for each orthologous gene pair it is possible to determine the gene expression similarity of orthologous genes.

Expression and activity correlations are useful tools that may be fused for further benefit. The "activity index" (AI) is a measure of correlation between expression profiles, while the "significance index" (SI) is a measure of the correlation between the significance/active levels, for the ontological pair (i.e., any pair wise combination of genes, proteins, metabolites). The Toxicogenomics Correlation Tool (Figure 2-7) visualizes the AI (x-axis) vs SI (y-axis) in a 2-D coordinate plane. As the values exist within the set {-1...1}, the Cartesian plane can be broken up into four quadrants where the quadrant-coordinate pair [denoted: (AI, SI)] mappings are: Quadrant 1 (1Q) is (+AI, +SI); Quadrant 2 (2Q) is (-AI, +SI); Quadrant 3 (3Q) is (-AI, -SI); Quadrant 4 (4Q) is (+AI, -SI).

The AI is determined primarily by the shape of the expression profile, where similar profiles yield more positive indices, and opposite profiles yield more negative

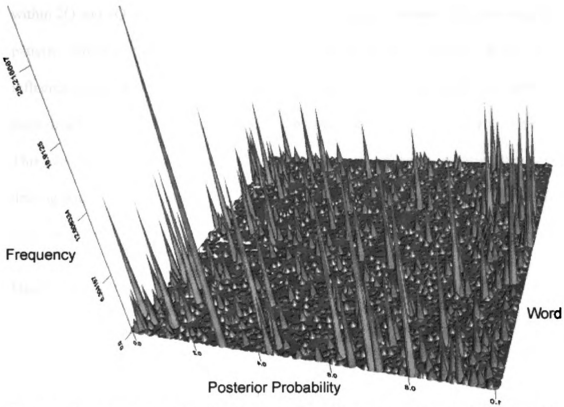

Figure 2-8: Visualization Control Center. The Visualization Control Center (VCC) provides the ability to visualize data in different dimensions. The VCC is a plug-in capable tool written in Java2. The visualization capabilities of the VCC are independent of the data types, thus, any appropriate data may be visualized in 3-D using this tool. In this example, a rotated 3-D cone-plot of posterior probabilities (x-axis; due to image rotation it appears as the mirror of the x-axis), short sequence motifs (z-axis), and their frequency of occurrence (y-axis) within an active gene set relative to an inactive random set of genes is shown. Large cones to the far right identify the frequent occurrence of short sequence motifs in the regulatory region of responsive genes that may be involved in gene regulation when compared to a subset of non-responsive genes. Mousing over the cone identifies the motif the occurs frequently in the responsive gene set. Additional functionality that is under development include pattern recognition techniques.

indices. The SI is determined primarily by the degree to which treated and vehicle samples are different, and the variance within the groups. Thus, pairs within 1Q illustrate similar expression profiles and significant changes under the same conditions, while 3Q pairs exhibit opposite expression profiles, and significant changes under the "opposite" conditions. For example, gene pairs where gene 1 is significantly altered at early time points, and gene 2 is significantly altered at late time points will have a negative SI. Pairs

within 2Q and 4Q are more difficult to interpret. 2Q pairs illustrate opposite expression patterns, however their significance follows similar dynamics. This may be due to the influence of variance upon mean estimation for calculation of the expression pattern. 4Q pairs illustrate similar expression patterns, however the significance patterns are opposite. This may occur as a result of variability effecting the significance patterns. When dealing with orthology data, 2Q and 3Q pairs may also suggest the pairs do not represent true orthologues.

Data Visualization

Visualization techniques project and transform data to facilitate the identification of relationships. Commonly used visualization methods include agglomerative hierarchical clustering, k-means clustering, and 2-D and 3-D scatterplots.

For example, Figure 2-8 is a 3-D scatterplot produced by the Visualization Control Center (VCC). The VCC, coupled with the Query Control Center (QCC), a universal data querying tool for dbZach, projects data into three dimensions to facilitate the identification of trends. This serves as an initial, exploratory data mining activity, however, pattern recognition algorithms can also be added to extend the functionality of this tool.

Data Sharing

dbZach facilitates the sharing of data at the intra- and interlaboratory levels. Biomedical researcher-friendly graphical user interfaces (GUIs) allow investigators within the laboratory to query for any data within the database (Figure 2-9). Interlaboratory sharing is facilitated by export of data using the emerging Microarray and

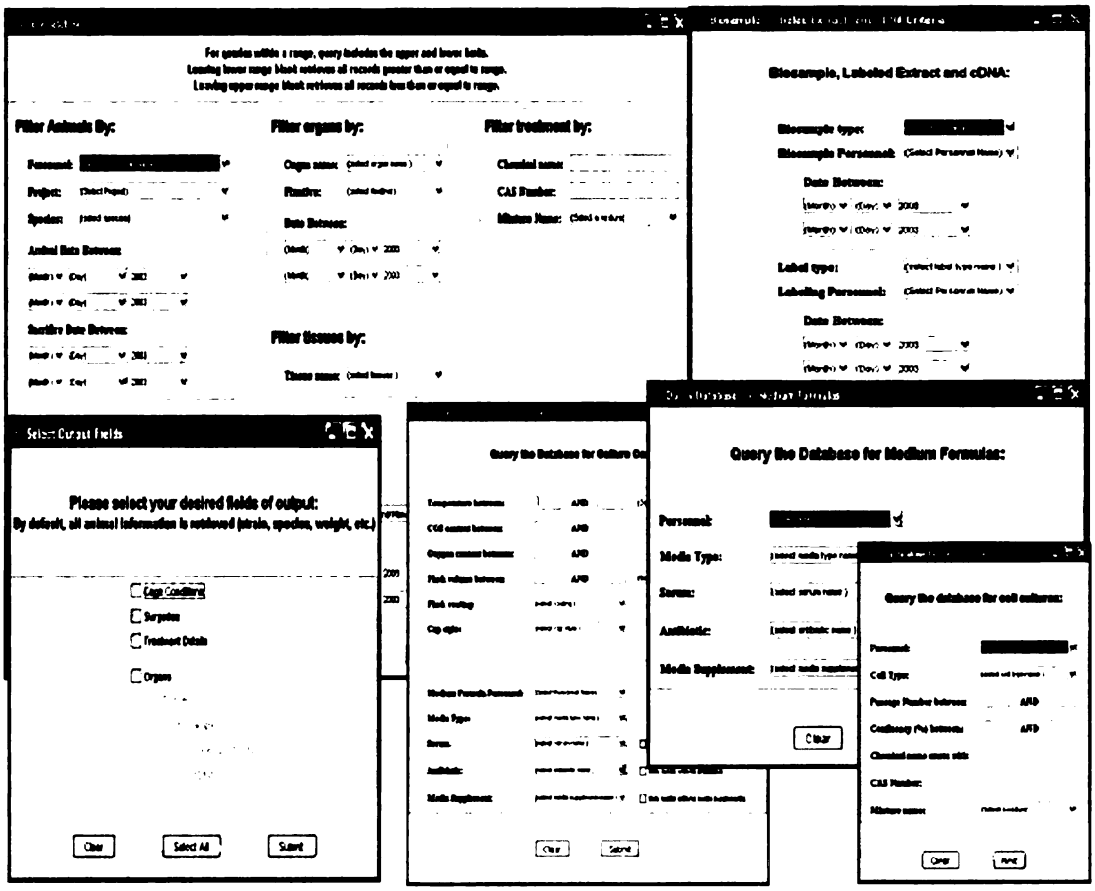

Figure 2-9: Intralaboratory Data Access GUIs. Investigators within the laboratory have access to all of the data within dbZach. Users may access data either through direct query using SQL, a domain-specific GUI tool such as the Gene Annotation Tool (GAT) and the Sample Annotation Interface, or the generic Query Control Center (QCC). The Sample Annotation Interface GUI (shown here) and the QCC are menu-driven bulk query systems (i.e., users filter their queries down using guided menu items), whereas the GAT is not primarily menu-driven, rather, it takes spreadsheets as input to perform queries on specific terms, such as clone identifiers.

Gene Expression (MAGE) Markup Language (MAGE-ML) (Spellman, *et al.*, 2002) that facilitates the electronic transfer of data between databases including public repositories.

The dbZach System encompasses several intralaboratory GUI-based applications developed in Java that ensures data sharing tools may be used across different platforms (e.g., Windows, Mac OS X, Linux). Several domain-specific GUI applications have been developed in the past, such as the Clones Interface, Genes Interface, and the Gene

Annotation Tool (GAT). These interfaces facilitate bulk query of the database, and are useful for obtaining general information about the clones represented on microarrays within the database, the annotation available for genes within the database, and the functional annotation of microarray data, respectively.

Data sharing at the consortium level can occur through the mechanisms outlined above, or by leveraging database replication technology. Database replication allows the contents of remote databases to be replicated within the central system, and the contents of the central system to be replicated at the remote systems. Database replication typically provides faster query return times and less stress on the central database server.

The dbZach MAGE Exporter facilitates interlaboratory data sharing by leveraging the MAGE file format. This same application also facilitates the deposition of microarray data to repositories, such as ArrayExpress (Brazma, *et al.*, 2003; Rocca-Serra, *et al.*, 2003), the Chemical Effects in Biological Systems (CEBS) Knowledgebase (Waters, *et al.*, 2003), and the Gene Expression Omnibus (GEO) (Edgar, *et al.*, 2002).

Quality Assurance

Quality assurance takes two forms with respect to data within the database, 1) audits, and 2) traditional quality assurance. The goal of the audit is to ensure data within the database faithfully represent what was supposed to be entered. The goal of traditional quality assurance is to ensure data conform to the quality standards of the organization and the scientific community.

Data audits are performed within dbZach prior to analysis. Audit and Report Tool (ART) applications are designed for particular data domains, such as the Microarray Audit and Report Tool (MART). These tools produce detailed multilevel audit reports

for data producers and submitters to verify uploaded data are correct and appropriately related with other data. Prior to analysis, data analysts will also invoke the appropriate ARTs (e.g., Sample Annotation, Toxicology and Pathology) to further verify the validity of the data they are about to analyze.

Databases also serve as a rich source of information for generating quality assurance protocols. As the volume of information within the database increases, a large pool of training data becomes available for the generation of models for quality assurance and process control which provide non-biased quality assessments.

dbZach Status

Tables 2-3 through 2-6 provide brief summaries of the data within dbZach, as of March 2, 2005. Similar up-to-the-minute status reports are also available at http://dbzach.fst.msu.edu:8050/dbZachCurrentStats/Statistics. Table 2-3 lists the current number of cell culture entries, broken down by cell line name and species. Table 2-4 presents the number of animals represented within dbZach by species, and a sampling of the tissues collected from these species, along with their counts. Table 2-5 provides an overview of the number of clones represented on the current in-house microarrays for mouse, rat, and human. Also, the number of genes represented on these arrays, and the number of genes represented by more than 2 clones is provided to indicate the level of redundancy present on the microarrays. Table 2-6 provides details regarding the number of microarrays and the number features from these arrays, from *in vitro* and *in vivo* experiments by species. In total, dbZach currently manages 31.4 million features from approximately 2500 microarrays.

Table 2-3: Number of Cell Culture Entries

Cell Type Name	Species	Cell Culture Entries
HL1-1*	human	25
HepG2	human	56
905K-1†	human	15
hepa1c1c7 c1	mouse	3
hepa1c1c7 c4	mouse	3
hepa1c1c7 wt	mouse	472
hepa1c1c7 c12	mouse	3
H4IIE	rat	48

* Human liver stem cell line
† Human kidney stem cell line

Discussion

Laboratories engaged in toxicogenomics can benefit from databases by not only facilitating the management and sharing of large, multivariate, disparate datasets but also in the generation of novel hypothesis. Furthermore, databases reduce data redundancy while providing a modular data integration solution. This modularity serves to increase the return on investment as subsystems may be seamlessly added or "plugged-in" without further redevelopment of the backend in order to support management and integration of data from nascent technologies.

Relational databases also support the generation of quality assurance protocols to ensure high quality conclusions are derived from the data. For example, historical datasets may be defined within the database and used for training statistical learning theory models (e.g., Support Vector Machines) to identify high and low quality microarrays. However, to support these efforts database developers must incorporate data auditing methods, such as multi-level reporting (e.g., where frequency data are reported on a per experiment basis, and 2-way tables illustrating cross-tabulated frequencies) to identify problems with data submissions.

Table 2-4: Animal and Organ Entries

Species	Number of Organisms	Liver Sections	Kidney	Mammary	Uterus
Mouse	685	1,518	854	1,135	633
Rat	396	316	256	256	396

Table 2-5: Clone and Gene Information

Species	Clones	Genes	Genes Rep'd by >2 clones
Human	10,068	6,025	214
Mouse	13,362	7,952	568
Rat	8,567	3,022	35

The greatest utility of databases is the ability to effectively mine, or uncover, relationships across large, complex data domains and experiments, that are not intuitively obvious. For example, a single database query can identify all genes that are active following the same treatment in several different tissues, building a hypothesis for a putative biomarker of exposure to a specific chemical class. Using similar logic, queries on histopathology data may identify chemicals that yield similar and conserved histological events across tissues and/or species. This would provide evidence of functional consequences resulting from conserved mechanisms of action and would support cross species extrapolations in quantitative risk assessment by reducing uncertainties inherent in the source-to-outcome continuum.

Data integration facilitates querying across data domains and engenders systems toxicology, the iterative development of computational models that are predictive of outcome based on expssure data, or can predict dose based on response. For example, data integration methods are used for phenotypic anchoring of "omic" observations. By further integrating with orthology data, it is possible to identify conserved responses to chemical exposures across species. Through integration of multi-technology responses

Table 2-6: Count of Microarrays and Features

	In Vitro			*In Vivo*		
Category	**Human**	**Mouse**	**Rat**	**Human**	**Mouse**	**Rat**
Microarray	170	786	0	N/A	1,156	358
Features	1,833,141	10,836,346	0	N/A	15,635,954	3,081,216

(e.g., genomics, proteomics, metabolomics), with species, histopathology, and other toxicologic response data, novel multidimensional data analysis (e.g., data fusion) and visualization methods may be used to develop computational models that can predict exposure levels, response outcomes and identify mechanistically-based biomarkers of exposure and toxicity.

The growing interest in data sharing (Ball, *et al.*, 2004b; Brazma, *et al.*, 2001) and calls for increased use of data repositories (Ball, *et al.*, 2004a), require investigators to consider effective methods for data exchange. Databases, such as dbZach, which are capable of exporting MIAME-compliant data in MAGE-ML, not only provide effective sharing mechanism that maintain the integrity of the data, but also provide significant time savings when submitting data to public repositories and other interested investiagtors. These methods are less error prone than web-interaction based submissions as data within the database are written directly to a file without human intervention.

Conclusion

Databases support the integration of disparate data to facilitate analysis and foster the development of new analysis techniques. The dbZach System currently provides integration of toxicology, gene expression (microarray and real-time PCR), gene functional annotation, orthology, and gene regulation data. These capabilities are currently being extended to include metabonomic data, with proteomic and biological

pathway data slated as the need arises. By combining data integration and quality assurance capabilities, and leveraging new analysis and visualization technologies, such as data fusion (Joint Directors of Laboratories, 1991) and other advanced statistical and machine learning approaches, uncertainties within the source-to-outcome continuum will be reduced, ultimately engendering mechanistically-based quantitative risk assessment.

Acknowledgements

The authors would like to acknowledge Dr. Rob Halgren, Dr. Yan Sun, Shane Doran, Shraddha Pai, Raeka Aiyar, Jigger Vakharia, Rebecca Rotman, Bonny Lau, Andrea Adams, Jung-sup Lee, Willis Lang, Rahul Sarkar, and Stacy Hung for their efforts in developing code associated with this project. This work was supported by NIEHS grants ES 04911-12, ES 011271, and ES 011777.

References

Bader, G. D. and Hogue, C. W. (2000) BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics, **16**, 465-77.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S. A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J. and Winegarden, N. (2004a) Submission of microarray data to public repositories. PLoS Biol, **2**, E317.

Ball, C. A., Sherlock, G. and Brazma, A. (2004b) Funding high-throughput data sharing. Nat Biotechnol, **22**, 1179-83.

Boverhof, D. R., Burgoon, L. D., Tashiro, C., Chittim, B., Harkema, J. R., Jump, D. B. and Zacharewski, T. R. (2005) Temporal and dose-dependent hepatic gene expression patterns in mice provide new insights into TCDD-mediated hepatotoxicity. Toxicol Sci,

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, **29**, 365-71.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. and Sansone, S. A. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, **31**, 68-71.

Buck, M. J. and Lieb, J. D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics, **83**, 349-60.

Bushel, P. R., Hamadeh, H., Bennett, L., Sieber, S., Martin, K., Nuwaysir, E. F., Johnson, K., Reynolds, K., Paules, R. S. and Afshari, C. A. (2001) MAPS: a microarray project system for gene expression experiment information and data validation. Bioinformatics, **17**, 564-5.

Edgar, R., Domrachev, M. and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, **30**, 207-10.

Fong, C. J., Burgoon, L. D. and Zacharewski, T. R. (2005) Comparative Microarray Analysis of Basal Gene Expression in Mouse Hepa 1c1c7 Wild-type and Mutant Cell Lines. Toxicol Sci, in submission.

Joint Directors of Laboratories. (1991) Data Fusion Lexicon.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature, **431**, 308-12.

Mattes, W. B., Pettit, S. D., Sansone, S. A., Bushel, P. R. and Waters, M. D. (2004) Database development in toxicogenomics: issues and efforts. Environ Health Perspect, **112**, 495-505.

Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Contrino, S., Vilo, J., Abeygunawardena, N., Mukherjee, G., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A. and Sansone, S. A. (2003) ArrayExpress: a public database of gene expression data at EBI. C R Biol, **326**, 1075-8.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr. and Brazma, A. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol, **3**, RESEARCH0046.

Sun, Y. V., Boverhof, D. R., Burgoon, L. D., Fielden, M. R. and Zacharewski, T. R. (2004) Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. Nucleic Acids Res, **32**, 4512-23.

Tong, W., Cao, X., Harris, S., Sun, H., Fang, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., Perkins, R., Shi, L. and Casciano, D. (2003) ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. Environ Health Perspect, **111**, 1819-26.

Waters, M., Boorman, G., Bushel, P., Cunningham, M., Irwin, R., Merrick, A., Olden, K., Paules, R., Selkirk, J., Stasiewicz, S., Weis, B., Van Houten, B., Walker, N. and Tennant, R. (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics, **111**, 15-28.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000) DIP: the database of interacting proteins. Nucleic Acids Res, **28**, 289-91.

CHAPTER THREE

Burgoon, L.D., Eckel-Passow, J.E., Gennings, C., Boverhof, D.R., Burt, J.W., Fong, C.J., Zacharewski, T.R. (2005) Protocols for the Assurance of Microarray Data Quality and Process Control. In submission.

Abstract

Microarrays represent a powerful technology that provides the ability to simultaneously measure the expression of thousands of genes. However, it is a multi-step process with numerous potential sources of variation that can compromise data analysis and interpretation if left uncontrolled, necessitating the development of quality control protocols to ensure assay consistency and high data quality. In response to emerging standards, such as the Minimum Information About a Microarray Experiment (MIAME) standard, tools are required to ascertain the quality and reproducibility of results within and across studies. To this end, an intralaboratory quality control protocol for spotted microarrays was developed using cDNA microarrays from *in vivo* and *in vitro* dose-response and time-course studies. The protocol combines: 1) diagnostic plots monitoring the degree of feature saturation, global feature and background intensities, and feature misalignments with 2) plots monitoring the intensity distributions within arrays with 3) a support vector machine (SVM) model. The protocol is applicable to any laboratory with sufficient data sets to establish historical high and low quality data.

Introduction

Microarray technology provides the ability to simultaneously measure the expression of thousands of genes in a cell, tissue or model of interest. However, numerous potential sources of experimental variation (Hessner, *et al.*, 2004; Jarvinen, *et al.*, 2004) have raised concerns regarding assay consistency, and data quality which confounds the ability to compare data sets between independent investigators and undermines the utility of intralaboratory (i.e., local), interlaboratory (i.e., collaborative center), or global scale (i.e., public repository) data sharing and exchange efforts (Miles, 2001; Ulrich, *et al.*, 2004). Consequently, quality assurance and control protocols that assess the reproducibility of data by identifying deviations or abnormal trends in assay performance and data quality are required.

Although several quality assurance and control methods have been proposed, criteria for differentiating high- from low-quality microarrays is lacking, leaving assessment open to interpretation. Many methods attempt to address this impediment through a variance-based statistical method, however they suffer from a lack of training, as the method solely tests the hypothesis of deviation from the rest of the population, and fail to judge data based on prior knowledge. Therefore, arrays that are technically of low quality (i.e., high background, low feature signal intensity, misaligned features, or inappropriately distributed feature intensity values) can still be labeled as high-quality, if they belong to a larger population of low-quality arrays.

In lieu of these more complicated quality assurance and control methods, data quality has been reported in terms of sample clustering by assessing whether biological replicates cluster together (Grant, *et al.*, 2003). Although this methodology determines

whether or not biological replicates exhibit similar behavior, it provides minimal insight into the technical quality of the assay (i.e., are these microarrays of high-quality). For example, similarly treated biological replicates may cluster together, or yield similar patterns, in light of poor technical quality (e.g., high background, narrow dynamic range). Moreover, this method may yield false-negative results in a background of extensive biological variation.

In addition, quality assessments can be stratified to the feature (Hautaniemi, *et al.*, 2003; Wang, *et al.*, 2001), subgrid or block (Gollub, *et al.*, 2003), or microarray (Model, *et al.*, 2002; Petri, *et al.*, 2004) level. Although examination of each stratum is crucial, a comprehensive analysis strategy based on all strata would be advantageous. Thus, the most robust, comprehensive quality assurance and control protocol would incorporate aspects of training by using historical datasets (HDS) of known quality, provide analysis at all microarray quality strata, and diagnose possible sources of poor quality data that could be corrected and addressed to minimize future problems (i.e., quality assurance).

In this report, a three step intralaboratory quality-control protocol is proposed to assess spotted microarray data quality as a first step towards ensuring publicly accessible data is of high quality. Global feature and background signal intensities as well as signal-to-noise ratios are first assessed to identify problems with raw microarray data quality. The feature identification process, commonly referred to as gridding, is then computationally examined to identify potentially misaligned features, which can be corrected to minimize potential downstream errors in normalization and functional assignment. Finally, a more in-depth assessment of raw and normalized data distributions is utilized to ensure that a sufficient dynamic range has been achieved for

subsequent analyses. 388 time course and dose response two color cDNA microarray data sets are used to establish high- and low-quality historical data sets and to demonstrate the utility of the protocol.

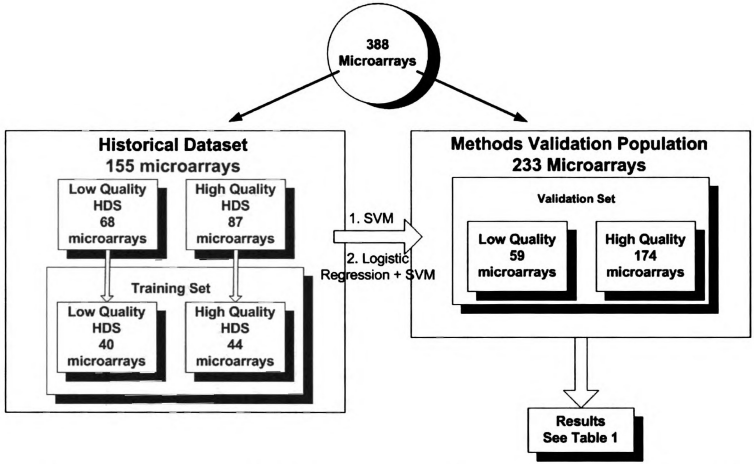

Figure 3-1: Historical, training, and validation data sets. The complete dataset of 388 microarrays was divided into two sets, the historical dataset (n = 155) and the validation set (n = 233). Both of these datasets were further subdivided into high and low quality datasets. The high- and low-quality historical datasets were further subdivided into their respective training sets by random sampling. The SVM models (SVM and logistic regression + SVM) were trained on the same training set data, and validated against the same validation datasets. The results of the validations are summarized in Table 3-1.

Materials and Methods

Creation of the Historical Datasets, Test, and Validation Sets

388 datasets, derived from *in vivo* and *in vitro* dose-response and time-course experiments using sequence verified cDNA microarrays were used to create both high- and low-quality historical datasets. Further details on microarray assay procedures are available at http://dbzach.fst.msu.edu/. All animal husbandry and sample collection procedures were approved by the Michigan State University All University Committee on Animal Use and Care. Microarrays were scanned using an Affymetrix 428 scanner, and images were quantified using GenePix v5.0 or v5.1.

Global statistics are calculated as:

$$\bar{x}_d = \frac{1}{n}\sum_{i=1}^{n} x_{di}$$

where d represents the dye (Cy3 or Cy5), n represents the number of features on the array, and x_{di} represents the median feature intensity (either feature signal or background from the image analysis software) for the d^{th} dye and the i^{th} feature.

The historical dataset consists of 155 microarrays that were further classified as high (n = 87) or low (n = 68) quality based on corroboration by quantitative real-time PCR ($p < 0.05$ for the correlation of the gene expression pattern of selected genes), low feature background intensity, congruent distributions of data points, and detection of comparable numbers of features. The background feature intensity does not have a threshold per se, rather it is based on visual inspection for high overall signal and anomalies such as smears, waves and excessive dust, the ratio of signal to background being greater than 20, the number of identified features, where at least 95% of the

Table 3-1: Comparison of the predictive accuracy of support vector machine (SVM) models for microarray quality predictions.

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
All Predictor Variables*	0.96	0.98	0.99	0.89
Regression Predictive Variables†	0.99	0.95	0.98	0.97

* All Predictor Variables includes six variables: Cy3 and Cy5 mean global feature intensities; Cy3 and Cy5 mean global background intensities, Cy3 and Cy5 signal-to-noise ratios (ratio of the two above listed values).
† Regressed variables are the predictive variables identified using a step forward logistic regression of the above six variables. These are: Cy3 global mean feature intensity, Cy3 global mean background intensity, Cy5 global mean background intensity, Cy5 signal-to-noise ratio.

features are detectable, and the distribution of intensity values must be comparable across the experiment. Examples of high and low quality images for each criteria are provided as supplementary data to further assist in defining the thresholds we initially used to establish our historical training set (HDS). Arrays not found to have the desired characteristics were categorized as low quality. Quality assignments are not a weighted vote approach, but rather an all or nothing voting scheme, where high quality arrays must meet all of the qualifications listed, and are specific to our HDS. The training set was derived from a random sampling of both high and low quality datasets to form a high (n = 44) and low (n = 40) quality training sets.

The validation dataset consisted of the 233 arrays not included in the historical dataset. The quality of these arrays was assessed the same as the historical dataset, resulting in 174 high- and 59 low-quality arrays (Figure 3-1).

Division 1 Analysis

Predictive variables include any parameter that is of interest to the investigator that may be indicative of quality. For example, these variables may include 1) the mean feature intensity across the array for each dye, 2) the mean background intensity across the array for each dye, 3) the mean ratio of the feature and background intensities, 4) atmospheric ozone concentration, and 5) laser intensity.

A step-forward logistic regression procedure was used to identify the most predictive variables for training the support vector machine (SVM). The dependent variable for the logistic model is a binary variable that reflects whether the microarray is of high- or low-quality, while the independent variables are the predictor variables from the historical dataset. The step-forward logistic regression enters predictor variables into the model one-at-a-time so long as it meets the significance threshold from the chi-square test. The HDS used for training the SVM is adjusted to reflect only the logistic regression predictive variables ($p < 0.05$).

The SVM is then trained using step-forward logistic regression predictive variables from the combined high quality (HQ-) and low quality (LQ-HDS). As microarray data become available (i.e., scanned and quantified) the resultant SVM model was used to classify microarrays as either high- or low-quality. High-quality microarrays continue through the protocol, while low-quality microarrays were flagged for repeat experiments. All data were stored for future inclusion into the HDS.

The logistic regression was performed using the LOGISTIC procedure in SAS v8.2, while the SVM training and analysis were performed using the e1071 package in R

v1.8.1 using a radial basis kernel. Details of SVM and PROC LOGISTIC implementation are given in their respective documentation.

Division 2 Analysis

Feature alignment was assessed using a loess nonparametric regression procedure that was developed as a normalization method to estimate bias on a per-array, print-tip or subgrid, and channel basis, and is visualized by MA-plots. Feature alignment is analyzed using a variant of the standard MA-plot (Yang, *et al.*, 2002), referred to as a modified MA-plot (Eckel, *et al.*, 2004). With respect to the modified MA-plot the true signal intensity for the i^{th} feature is either estimated as the average signal intensity across all arrays, dyes, and treatments ($\hat{\mu}_i$) or as the signal intensity across all arrays and dyes for each of the *j* treatment groups separately ($\hat{\mu}_{ij}$) for a particular experiment. The choice between using $\hat{\mu}_i$ versus $\hat{\mu}_{ij}$ is discussed in detail in (Eckel, *et al.*, 2004). Thus, the estimated true signal intensity is a substitute for the A-term in the modified MA-plot. The M-term estimates the bias associated with using $\hat{\mu}_i$ or $\hat{\mu}_{ij}$ to estimate the true signal intensity such that *M* is equal to the difference between each signal intensity with its corresponding estimated true signal intensity. After computing the estimated true signal intensity and the bias, a modified MA-plot is constructed separately for every array and a nonparametric regression smoother is fit to each print-tip on the corresponding array individually. If the nonparametric regression smoother for a particular print-tip, or for a subset of print-tips, is an obvious outlier, feature alignment is investigated. All procedures were performed in SAS v8.2.

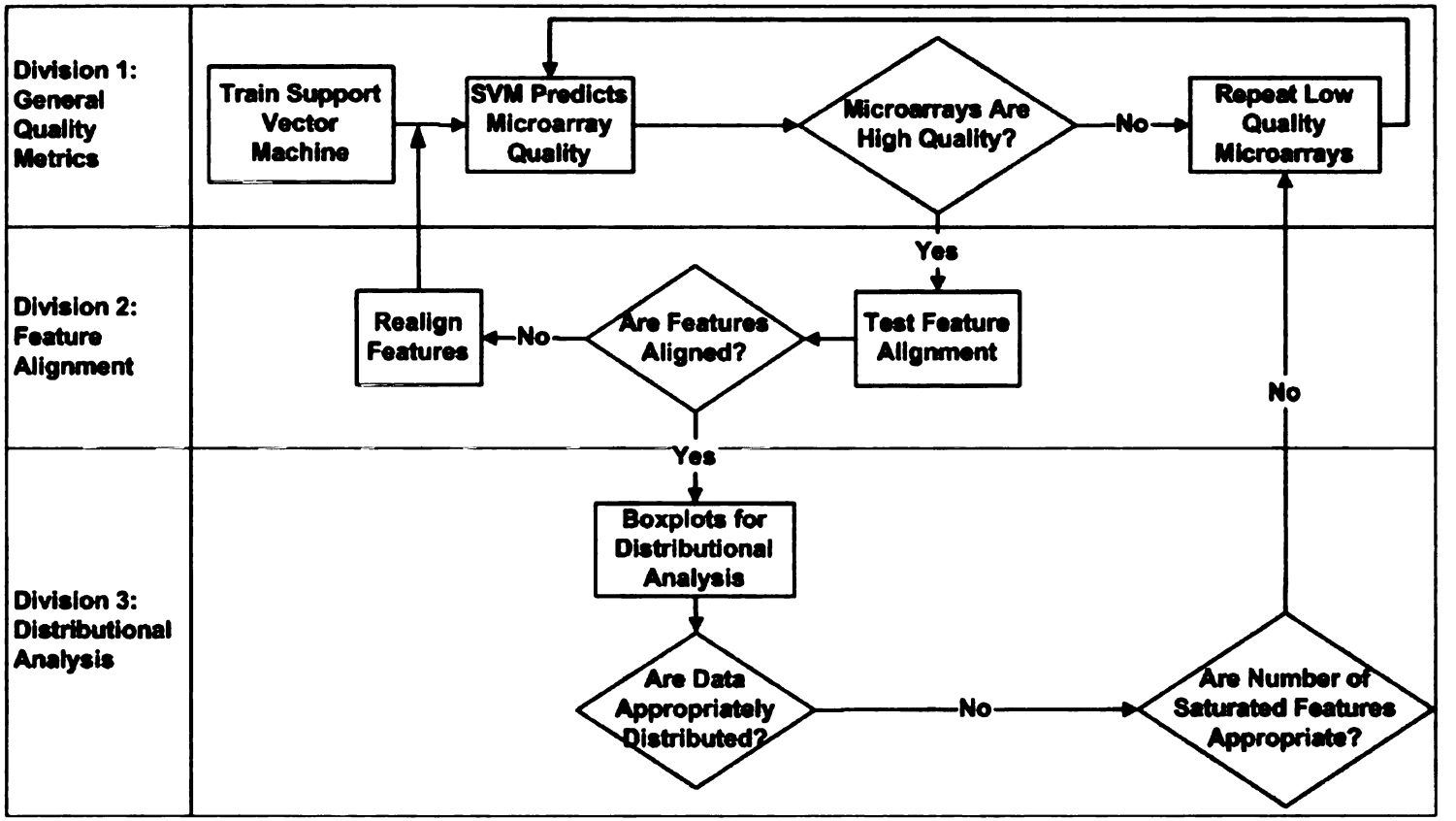

Figure 3-2: Microarray quality control protocol. General Quality Metrics, Feature Alignment, and Distributional Alignment divisions are depicted with two additional divisions that place the protocol into context with the overall data management infrastructure. General Quality Metrics (Division 1) analysis uses a Support Vector Machine (SVM) model trained on the historical data set (HDS), which includes a combination of the high-quality and low-quality data sets (HQ-HDS and LQ-HDS, respectively). Feature Alignment (Division 2) conducts a loess analysis based on treatment, dye, and microarray variables using the raw intensity values from each array to determine if a subgrid has been misaligned during the quantification process. Distributional Analysis (Division 3) combines box-and-whisker plots with standard line plots to identify trends in data distributions and the number of saturated and unidentified features.

Division 3 Analysis

Intensity distribution was assessed using box-and-whisker plots on a per-array basis. Line plots demonstrating trends in global mean feature intensity, global mean background intensity, and the count of saturated features were created depicting upper control limits (UCL) and lower control limits (LCL) for each metric. Acceptable numbers of saturated features have been historically established in this laboratory to be 1-2% of the total number of features. To assist in quality analysis it is generally useful to group microarrays performed on the same date together when plotting to identify temporal trends. All procedures were performed in SAS v8.2.

Results

Figure 3-2 provides an overview of the microarray data quality-control protocol which is divided into General Quality Metrics (Division 1), Feature Alignment (Division 2), and Distributional Alignment (Division 3). Two additional divisions are included to place the protocol into context within the overall data management scheme.

Establishment of High- and Low-Quality Historical Datasets

High- and low-quality historical datasets were created to anchor quality assessments to arrays of known quality to prevent inappropriate assessment of arrays as high-quality due simply to low variance within the study. High quality was defined empirically based on corroboration by a complementary technology (e.g., quantitative real-time PCR (QRTPCR)), low feature background intensity, congruent distribution of

data points, and detection of a comparable number of identified features. For example, among high quality arrays, QRTPCR corroborates greater than 80% of the gene expression trends exhibited by arrays (Boverhof, *et al.*, 2005; Boverhof, *et al.*, 2004a; Boverhof, *et al.*, 2004b; Burt, *et al.*, 2005; Fong, *et al.*, 2005a; Fong, *et al.*, 2005b; Kwekel, *et al.*, 2005; Sun, *et al.*, 2004). Arrays not found to have the desired characteristics in all of the above categories were labeled as low quality.

The HQ-HDS is based on a random sampling of the high-quality microarrays from all investigators within our laboratory (HQ-HDS: $n = 87$), and a LQ-HDS similar to the HQ-HDS, but representing a random sampling of the low-quality microarrays (LQ-HDS: $n = 68$) from an overall total of 388 time-course and dose-response two-color cDNA microarrays. Each HDS consists of the Cy3 and Cy5 global mean feature signal intensity (where global refers to the entire microarray), Cy3 and Cy5 global mean background signal intensity, and the Cy3 and Cy5 global signal-to-noise ratio (SNR; ratio of the global mean feature signal intensity to the global mean background signal intensity) for each array in the dataset.

Table 3-2: Logistic regression odds ratio for significant predictor variables

Predictor Variable	Odds Ratio (95% Confidence Interval)
Cy3 Global Mean Signal Intensity	1.003 (1.001, 1.004)
Cy3 Global Mean Background Intensity	0.979 (0.950, 1.009)
Cy5 Global Mean Background Intensity	0.980 (0.969, 0.992)
Cy5 Signal-to-Noise Ratio	1.601 (1.113, 2.305)

Division 1: Support Vector Machines Predict Microarray Quality

Division 1 analysis utilizes the HQ- and LQ-HDSs to develop and train a SVM model that best discriminates quality classes utilizing all six classification variables. The SVM model accurately classified (100%) a random sampling of low- (n=40) and high-quality (n=44) data sets from the HDS, here after referred to as the training set. Since this is a binary system the term positive is used to denote high-quality microarrays, while negative is used to denote low-quality microarrays. The positive predictive value (PPV) is the proportion of predicted high-quality arrays relative to the number of true high-quality arrays. The negative predictive value (NPV) is similar to the positive predictive value except it is calculated with respect to low-quality arrays. The SVM model accurately predicts high-quality microarrays when using a validation set (a randomly selected subset of the HDS, not including arrays from the training set) of 59 low-quality and 174 high-quality data sets, with a PPV of 99%, but performed less effectively when predicting low-quality microarrays, with a NPV of 89% (Table 3-1). In other words, 99% of the true high-quality arrays were accurately predicted to be of high-quality, while only 89% of the true low-quality arrays were accurately predicted to be of low quality.

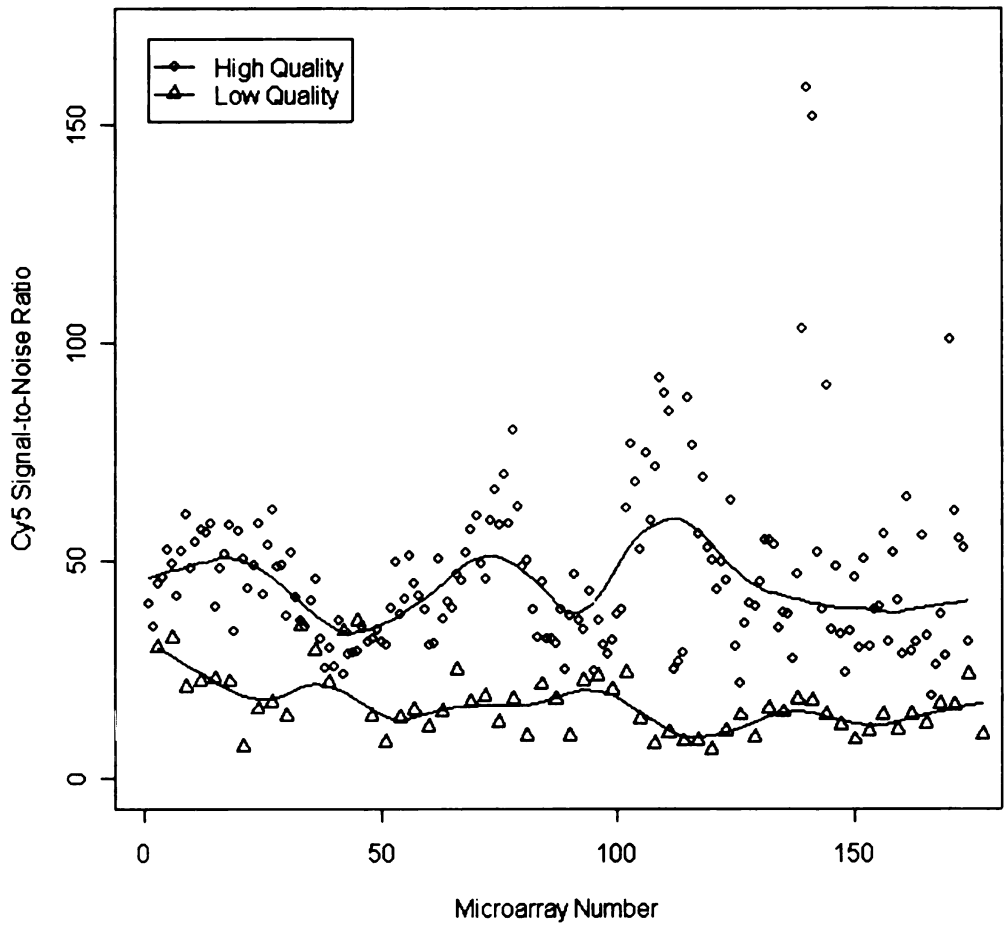

Figure 3-3: Cy5 signal-to-noise ratio is the most powerful predictor of high and low quality microarrays. High- and low-quality microarray data exist as two separable populations. The two lines represent a loess fit to the two different populations, and highlight the difference between the populations. High-quality microarrays tend to exhibit a larger Cy5 signal-to-noise ratio than their low-quality counterparts. The microarray number on the x-axis represents an identification number.

Logistic Regression Improves Predictive Accuracy of the SVM

Step-forward logistic regression identified Cy3 and Cy5 global (whole array) mean background intensity, Cy3 global mean feature intensity (mean of the feature median signal intensity), and the Cy5 global signal-to-noise ratio (ratio of global mean feature intensity and global mean background intensity for Cy5) as the most predictive

variables from the HDS, and were used to train a more discriminating SVM model. Cy3 and Cy5 global backgrounds were negative predictors of high-quality, as would be expected, while Cy3 global mean feature intensity and the global signal-to-noise ratio for Cy5 were positive predictors of high-quality microarrays. The most discriminate variable is the global Cy5 signal-to-noise ratio (odds ratio, OR = 1.60) (Table 3-2 and Figure 3-3). The loess fit lines illustrate the degree of difference between the two data populations (LQ- and HQ-HDS).

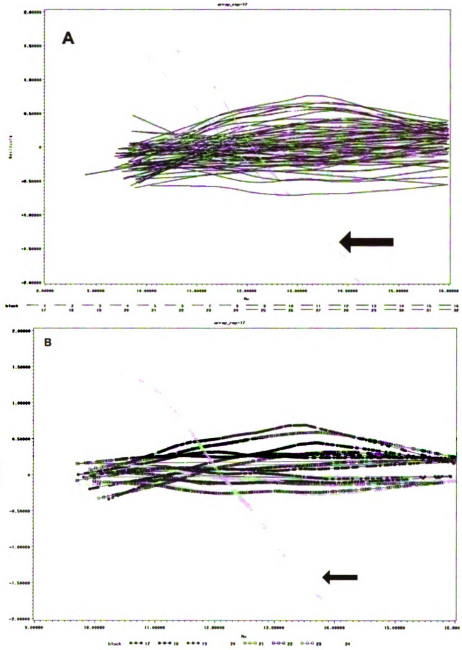

Figure 3-4: Loess analysis of microarray data identifies microarrays with misaligned grids. (A) Loess analysis of the raw intensity values from each array identified one misaligned subgrid on this microarray as evidenced by the lines with large, sharp slopes (arrow). Each subgrid is represented by two lines, one for each dye. (B) Subgrids 17-24 were identified as possibly problematic in A, and plotted in B for better resolution, identifying subgrid #24 as the putatively misaligned subgrid. The investigator verified the misalignment using the quantification software and corrected it prior to further analysis.

By training the SVM using just the predictive variables identified using the step-forward

logistic regression model, the PPV remained relatively stable at 98%, while the NPV improved from 89% to 97% (Table 3-1). These results suggest that assessment using all available variables to train the SVM model contributes to noise that compromise array quality predictions made on the validation set.

Division 2: Nonparametric Regression Methods Detect Grid Misalignments

A nonparametric regression procedure is utilized for detecting grid misalignments. MA-plots have been used to visualize microarray normalization implemented on the print-tip level (Dudoit, *et al.*, 2002; Eckel, *et al.*, 2004). In addition to aiding in normalization, MA-plots assist with the identification of misaligned grids. Nonparametric regression methods, initially introduced to estimate bias, are also capable of identifying misaligned microarray quantification grids on a per-array basis provided that most of the microarrays under study are correctly aligned, and that misalignment is an infrequent, aberrant event (Eckel, *et al.*, 2004). Whereas most of the microarray grid blocks (a geographical region on the microarray where all features are printed by the same print-tip) have a slight nonlinear relationship, misaligned blocks will exhibit a significantly greater slope than correctly aligned blocks such that they appear as obvious outliers in the MA-plot (Figure 3-4A and B).

Arrays demonstrating misaligned features are identified for follow-up and realignment. The realignment of the block will result in the alteration of the global intensity values for that array and as a result are resubmitted for Division 1 analysis. During the realignment process, it may be possible to diagnose possible causes of the misalignment, such as high background, dust contamination, or robotic printing error,

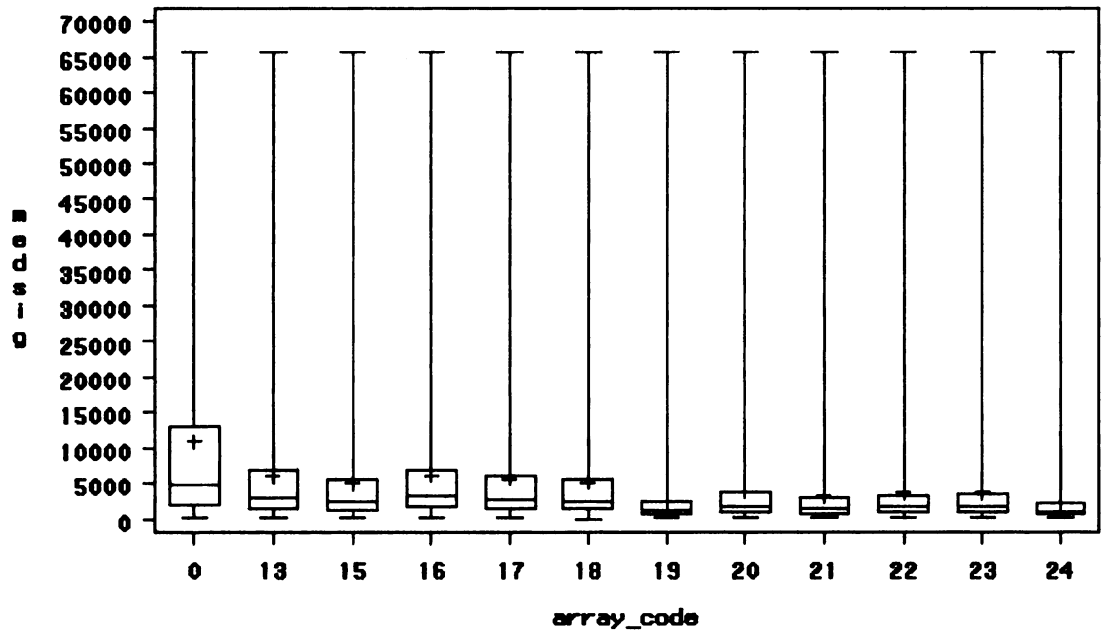

Figure 3-5: Illustration of the box-and-whisker plot to examine the distribution of feature intensities. Boxes represent the interquartile range, with the 75th percentile at the top and the 25th percentile at the bottom. The boxplot of the HQ-HDS population of median Cy3 signals (array_code = 0), illustrating a broad range of values, from eight randomly selected HQ-HDS arrays. Ideally the 75th percentile would be in the range of 7,000-13,000 units, with an interquartile range of approximately 5,000-9,500 units. The arrays under study (array_code > 0) exhibit some compression (Cy3 channel shown here), as indicated by compressed interquartile ranges (i.e. boxes), with microarrays 19-24 exhibiting the greatest compression issues. The line in the middle of the box represents the 50th percentile, or median, while the plus represents the mean. The pluses for arrays 20-24 lie on the 75th percentile line of the box. Whiskers represent the rest of the distribution, with their terminations representing the lowest and highest feature intensity values. The x-axis represents the individual microarray, while the y-axis represents the feature intensity values.

facilitating corrective action to minimize future occurrences thus improving assay performance and consistency

Division 3: Identifying Compressed and Similar Data Distributions in Microarray Data

Division 3 identifies microarrays with compressed or non-uniform dynamic range. Box-and-whisker plots were used to analyze feature intensity distributions on a per-

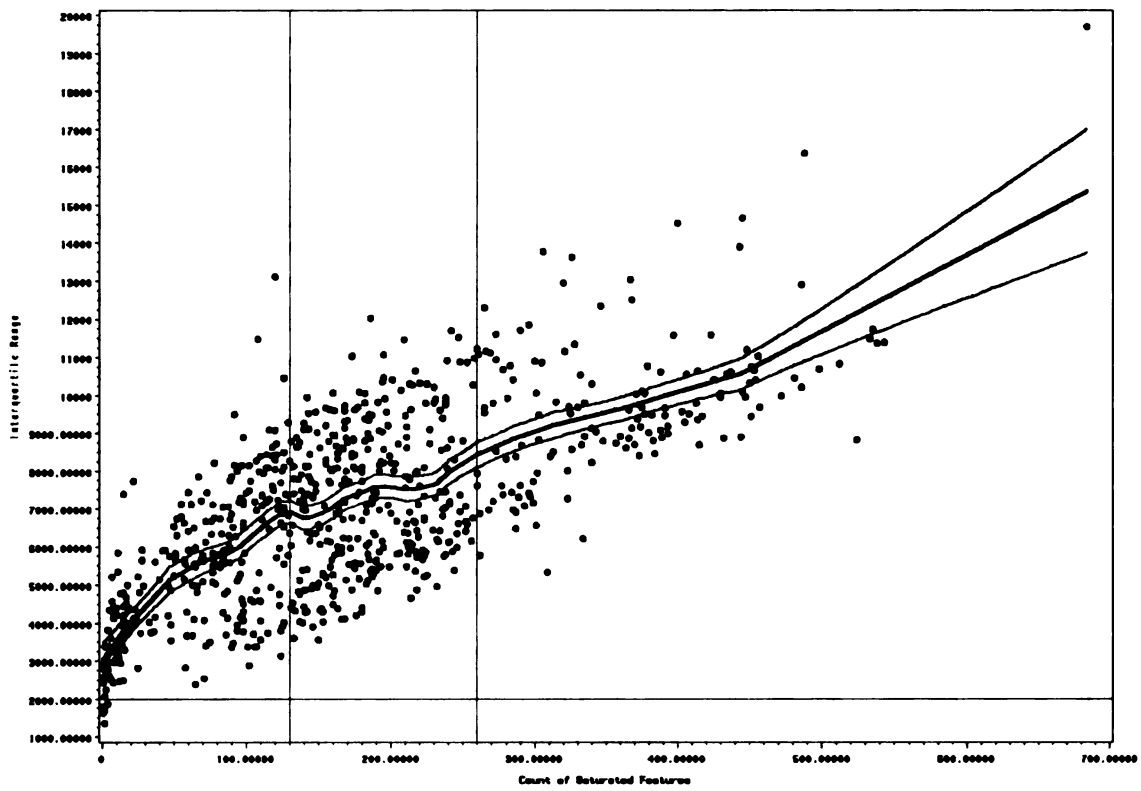

Figure 3-6: Interquartile range increases as a function of the number of saturated spots. The interquartile range is a measure of data spread, calculated as the difference between the 75th and 25th percentiles. The interquartile range increases with increasing number of saturated features, suggesting lower numbers of saturated features contribute to compressed ranges. By increasing the number of saturated spots compression is minimized. The lines on the plot represent the loess best fit line and the 95% confidence intervals.

microarray basis (Figure 3-5). Based on empirical observations, optimal distributions have the following characteristics: 1) a 25th percentile of approximately 700-2,000, 2) a 75th percentile of approximately 7,000-10,000 (i.e., interquartile range spanning intensities of 5,000-9,300 units), 3) a median of approximately 3,000-6,000, and 4) a mean within the interquartile range defined by the boxed region in Figure 3-5. The distribution of mean Cy3 median feature intensity values for the HQ-HDS is shown in Figure 3-5 (array_code = 0). Based on these criteria, microarrays 19-24 fail to show appropriate distributions because the 75th percentile is lower than the recommended range

of 7,000 to 10,000 (array_codes: 19-24). Microarrays 13-18 approach appropriate distributions, since the 75th percentile of the feature intensity distribution is closer to the recommended 75th percentile (i.e. 7,000-10,000) which is more consistent with the empirically defined recommendations based on the HQ-HDS (Figure 3-5). The distributions for all of these arrays are not optimal, as illustrated by the compressed feature intensity dynamic range as reflected in the constricted boxes (Figure 3-5).

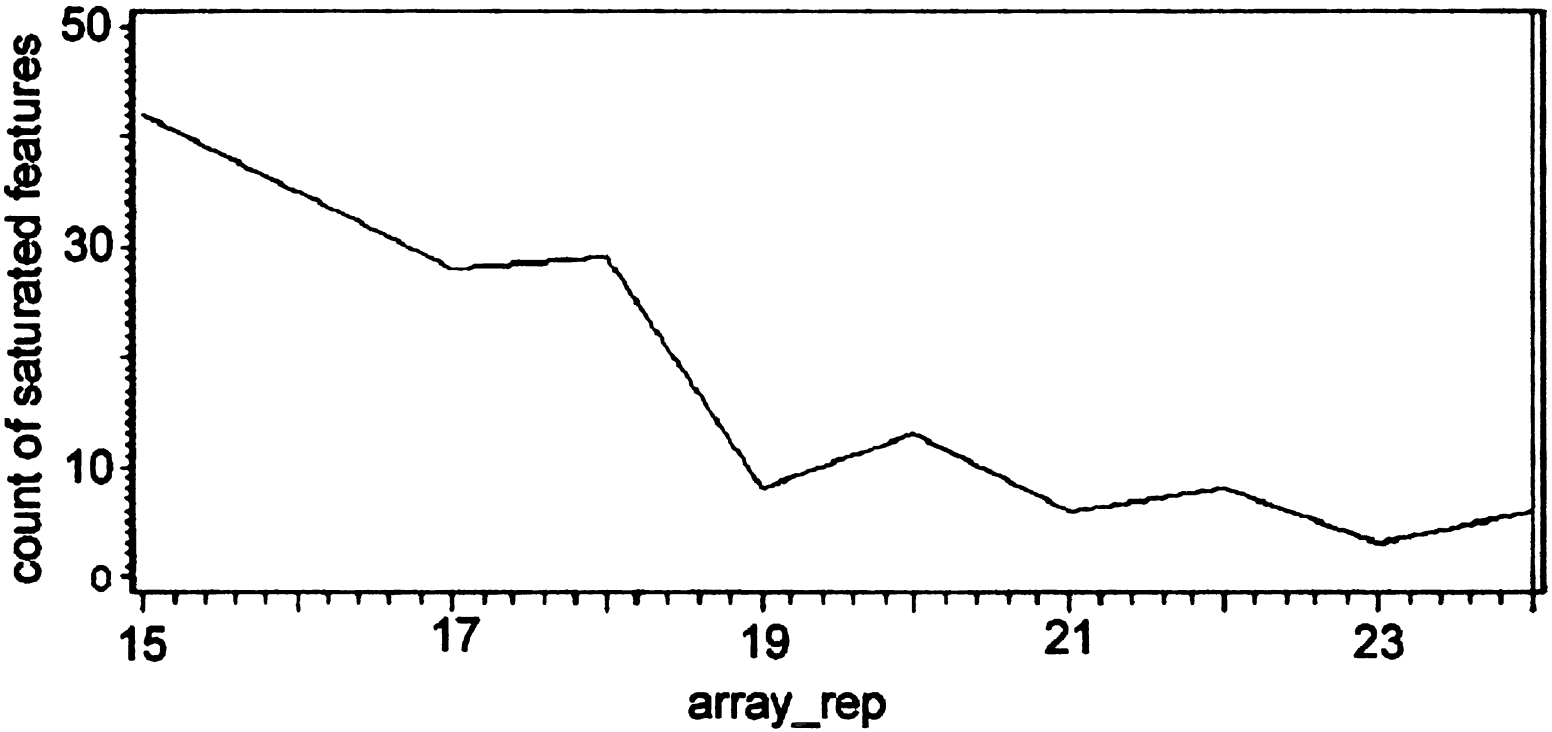

Figure 3-7: Saturated features correlate with compressed distributions. The microarrays depicted are the same shown in the box-and-whisker plot in Figure 5. The largest degree of distributional compression in Figure 5 corresponds to microarrays 19-24, the ones with the lowest number of saturated features.

Table 3-3: Applied Assumptions for Intralaboratory Quality Control and Assurance Protocol[a]
1. Test and training data sets were obtaining using the same, pre-agreed standard operating procedure (SOP)
2. Test and training data sets used the same microarray platform
3. Microarray scanning is performed using the same equipment
4. Image analysis (including segmentation and background calculation methods) used the same approach for test and training data sets
5. Same normalization methods were used for test and training data sets (Division 3 analyses)
a. The data sets available for this manuscript were insufficient to test the necessity of each assumption, and therefore, the necessity of each one was not tested.

As the interquartile range and number of saturated features are positively correlated (Figure 3-6), the number of saturated features serves as a useful surrogate marker to ensure comparable data distributions are achieved during array scanning. Figure 3-7 shows the number of saturated features per array for the microarrays shown in Figure 3-5 (array_codes > 0). Typically this plot includes the upper- and lower-control limits (empirically defined to be 2 and 1%, respectively). However, on this plot all of the microarrays (15-24) are well below the LCL (in the range of 0.1 – 0.5% of the features). Consequently, microarrays 19-24 have severely compressed dynamic range, as reflected by the low number of saturated features.

Implementation

The protocol is an initial step to provide investigators a non-biased data quality assessment tool that would facilitate the sharing of high quality data, albeit on a lab–to-lab basis. It is meant to be implemented locally, with a focus on intralaboratory or collaborative project quality assessments as opposed to broad quality assessments of data sets within public repositories. It is assumed that investigators have, at the very least,

practiced some form of feature quality control, such as that found in image quantification software (e.g., GenePix (Axon Instruments), AnalyzerDG (Molecularware)) or which can be implemented separately (Hautaniemi, *et al.*, 2003; Wang, *et al.*, 2001), prior to implementation of these methods. A more detailed listing of assumptions is provided in Table 3-3. The primary goal is to ensure arrays are of comparable quality, and to minimize unnecessary technical variation that may skew future results. As such, these techniques are platform independent, but do not support cross-platform quality comparisons within a study or across a public repository.

To implement the full protocol, an internally established historical dataset of high and low quality microarrays must be available in order to assess quality metrics of interest for Division 1 analysis. The predictive variables presented in this study are specific to our HDS; implementations of the general method by other groups may identify additional variables, although significant overlaps are likely. The logistic regression procedure is used to pare down the list of putative predictor variables, and the support vector machine is used to create a model to classify arrays as either high or low quality based on identified predictive variables. It should be noted that the logistic regression is used as a guide to determine which variables are predictive in the SVM. The logistic regression, a linear procedure, may not adequately model a non-linear prediction surface without the use of higher order terms (e.g., quadratic, cubic), thus the SVM is superior for non-linear estimates. Ultimately, the investigator must decide which variables are most predictive when used in the SVM. Investigators may also be required to use an alternative kernel in the SVM procedure to ensure optimal discrimination.

Division 2 and 3 analyses may be implemented without the use of the HDS, and may be implemented independent of Division 1, and each other. Division 2 and 3 analyses may be implemented using any statistical software that supports LOESS and boxplot creation, such as R or SAS.

Discussion

Quality control measures are performed to ensure that extreme or unusual variation and other technical issues do not overshadow biological and treatment variance. Although the goal of normalization is to minimize technical variation across samples, most normalization techniques will be more successful if less technical variation is present prior to normalization. Therefore, quality control techniques are used to identify technical variation arising from assignable causes due to the process. If the variability exceeds a chosen threshold, low-quality data sets can be identified and eliminated or corrected prior to further analysis while addressing sources of undesirable variation in future studies, thus improving assay performance and consistency. Normalization on the other hand corrects for variability that arises from assignable causes.

By controlling the quality of the data, assurances can be made that the results from these studies are due more to biological variation, and less to technical variation. Furthermore, by decreasing the technical variation, more accurate estimates of gene expression may be made, while making more power available for gene filtering and prioritization using statistical methods. This has direct impacts on knowledge that is exchanged through data sharing via scientific publications and public data repositories.

A streamlined and standardized process of microarray quality control has been developed that encompasses several complementary techniques. The protocol combines

a trained SVM model and nonparametric regression model with more classical techniques such as box-and-whisker and line plots. Although, it is possible to approach the line plot using a Shewhart plot, where control limits are defined based on the variance (*NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 4-5-04), for our purposes empirically defined control limits are preferred. Several different variables, including the feature signal and background intensity levels, signal-to-noise ratios, grid alignment, data distribution and dynamic range, and the number of saturated and undetected features are used to assess data quality on a per array basis, thus providing a streamlined, high-throughput analysis method to identify quality assurance issues that require intervention.

Specificity of the SVM model increased when using the logistic regression predictive variables, with negligible effects on sensitivity. These measures are properties of the test, but fail to address questions regarding the predictive nature of the model based on a population of microarrays. The PPV and NPV take into account the occurrence of high- and low-quality microarrays within the population in addition to the sensitivity and specificity. However, quality assignments by the SVM improved when only the most predictive variables, as determined by the step-forward logistic regression model, were used (Table 3-1). Collinearity between the Cy5 signal-to-noise ratio and the Cy5 background was not exhibited. The PPV remained stable while the NPV improved by 8% when using the parameters identified by the logistic regression model. By using the most predictive variables, noise within the system decreased, allowing for greater discrimination between high- and low-quality groups. With respect to the protocol,

microarrays that are of high-quality progress to Division 2 analysis while the samples from the low-quality microarrays are flagged to repeat the hybridization.

In our laboratory the step-forward logistic regression model identified four predictive variables (Table 3-2). However, these variables may differ among labs, and are expected to be technology/platform and protocol dependent. In this study, the global Cy5 signal-to-noise ratio was the most discriminating predictive variable (odds ratio (OR) = 1.60), providing the highest degree of stratification between the high- and low-quality microarrays (Figure 3-4). This degree of stratification is not entirely surprising as Cy5 is reported to be more susceptible to environmental factors, such as ambient ozone levels, than Cy3 (Fare, *et al.*, 2003). Thus, it is not surprising that the SVM continues to identify low quality arrays with questionable Cy5 backgrounds that are not apparent visibly.

Division 2 analyses focus on grid alignment using MA plots, and plotting the data on a per-block or subgrid basis to identify block misalignments. This streamlines the process of realignment which can be reassessed in Divisions 1 and 2, and minimizes the need to conduct costly, time consuming, and potentially unnecessary repeat hybridizations.

Division 3 analyses are concerned with data distributions, and ensuring a proper dynamic range. Appropriately and similarly distributed data are considered to be of high-technical quality and are forwarded for further analysis. Data distributions are assessed using box-and-whisker plots, where the highest intensity value should be at saturation (65,535 units). Data exhibiting appropriate distributions have yielded comparable results to those verified by quantitative real-time PCR (Boverhof, *et al.*, 2004a). Most problems with compressed interquartile range and distributions are linked to inappropriate

photomultiplier tube (PMT) gain settings. The PMT gain should be set to obtain a comparable number of saturated features (our experience is that 1-2% is appropriate) in order to achieve similarly shaped data distributions across all arrays (i.e., 75^{th} percentile of approximately 7,000-10,000 units, and 25^{th} percentile of approximately 700-2,000 units, with a mean within the interquartile range).

The most reliable indicator of obtaining appropriate dynamic ranges during the scanning process is the number of saturated features, and not the PMT value. We advocate shifting the PMT value in order to obtain a proper data distribution, and sacrificing the overall background intensity. Ideally, the background signal intensity will be low enough so that shifts in PMT will not adversely affect the number of identifiable features. Thus, it is not advisable to standardize the PMT gain value for an entire microarray experiment, as it is expected that optimal PMT gain values will vary by microarray. Following scanning, diagnostic plots can be used to determine if the number of saturated features meet the criteria (1% and 2% as the lower control limit (LCL) and upper control limit (UCL), respectively, are typically used). Abbreviated and compressed data distributions can manifest problems in downstream analysis and normalization, and may compromise subsequent statistical analysis of gene expression changes.

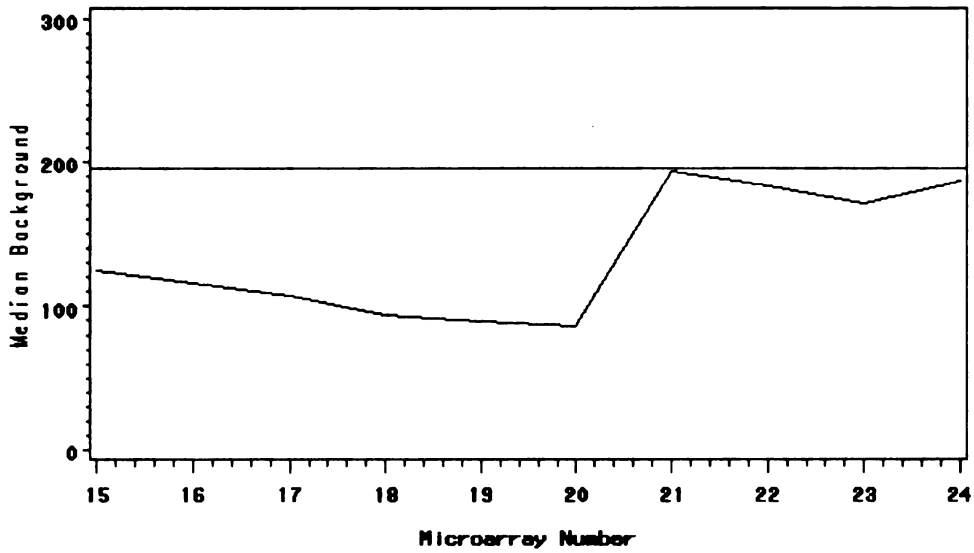

Figure 3-8: Background should be sacrificed for more saturated features. The microarrays depicted are the same shown in the box-and-whisker plot in Figure 5. The arrays with the largest Cy3 background are arrays 21-24. The reference line represents the mean Cy3 background for the HQ-HDS. In this case, the investigator was more concerned with obtaining a low Cy3 background than an optimal number of saturated features. Cy3 background should be sacrificed to increase the number of saturated features as the mean background for those arrays is below the mean for the HQ-HDS.

For example, arrays 19-24 exhibit the greatest degree of data compression (Figure 3-5) and highlight the correlation between the number of saturated features and the compressed distribution (Figure 3-7). The low background levels for these microarrays (Figure 3-8) is a likely contributing factor since the PMT gain was purposefully set low to minimize background intensity, resulting in the constricted interquartile range. Instead, PMT levels should have been increased to achieve 1-2% feature saturation to increase the probability of obtaining an appropriate and uniform distribution (dynamic range) across all microarrays within the study.

Following these quality control methods, only high-quality data should proceed to normalization and higher-order analyses. However, all microarray data should be stored in an appropriate database, including low-quality microarray data, for future refinement

of the HDSs. This ensures the quality of work being generated within a laboratory to be of their highest quality. However, it does not facilitate comparisons to the general body of publicly available data. By ensuring data being produced at the laboratory level is of the best local quality, investigators ensure the reproducibility of their results. However, the burden of quality assessment by the public user and peer reviewers still remains a challenge that is beyond the scope of these methods.

Conclusions

This protocol serves as an initial step to assess intralaboratory or collaborative group data quality for studies conducted using the same spotted microarray platform. Quality control ensures data integrity and is essential to facilitate subsequent analysis and meaningful interpretation that support conclusions, future hypotheses and knowledge-based decision making. It provides complementary QA/QC methods that include automated, high-throughput quality assessment using SVMs. Combining this protocol with other methods such as biological replicate clustering (Grant, *et al.*, 2003), and spot quality control assessments provides a more complete quality-control protocol that ensures the integrity of cDNA and oligonucleotide microarray data. The adoption of such measures is necessary to instill confidence in data uploaded to public repositories, an emerging requirement for a growing number of prestigious journals. However, the development of an enterprise solution that assesses data quality across platforms and between independent groups available within public repositories is needed in order to realize comprehensive knowledge extraction from publicly available complex data sets.

Acknowledgements

The authors would like to thank Gary Jahns for providing constructive comments on this work. This work has been supported by NIH grants ES11271, ES12245 and Superfund P42 ES04911. Support for LDB and JEE was provided by T32 ES07255 and NCI grant R25 CA92049, respectively. TRZ is partially supported by the Michigan Agriculture Experiment Station.

References

Boverhof, D. R., Burgoon, L. D., Tashiro, C., Chittim, B., Harkema, J. R., Jump, D. B. and Zacharewski, T. R. (2005) Temporal and dose-dependent hepatic gene expression patterns in mice provide new insights into TCDD-mediated hepatotoxicity. In submission,

Boverhof, D. R., Fertuck, K. C., Burgoon, L. D., Eckel, J. E., Gennings, C. and Zacharewski, T. R. (2004a) Temporal and dose-dependent hepatic gene expression changes in immature ovariectomized mice following exposure to ethynyl estradiol. Carcinogenesis, **25**, 1277-91.

Boverhof, D. R., Tam, E., Harney, A. S., Crawford, R. B., Kaminski, N. E. and Zacharewski, T. R. (2004b) 2,3,7,8-Tetrachlorodibenzo-p-dioxin induces suppressor of cytokine signaling 2 in murine B cells. Mol Pharmacol, **66**, 1662-70.

Burt, J. W., Burgoon, L. D., Humes, D., Kwekel, J. C., Harney, A. S. and Zacharewski, T. R. (2005) Effects of estrogen on immature, ovariectomized mice: A multi-approach, tissue-by-tissue comparison. In preparation,

Dudoit, S., Yang, H. Y., Callow, M. J. and Speed, T. (2002) Statistical methods for identifying differntially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, **12**, 111-139.

Eckel, J. E., Gennings, C., Therneau, T. M., Boverhof, D. R., Burgoon, L. D. and Zacharewski, T. R. (2004) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, in press.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y. and Wang, Y. (2003) Effects of atmospheric ozone on microarray data quality. Anal Chem, **75**, 4672-5.

Fong, C. J., Burgoon, L. D., Gupta, G., Humes, D. G. and Zacharewski, T. R. (2005a) Temporal Gene Expression Analysis of Mouse Hepa-1c1c7 Cells Treated with 17beta-Estradiol by cDNA Microarray. In preparation,

Fong, C. J., Burgoon, L. D. and Zacharewski, T. R. (2005b) Comparative Microarray Analysis of Basal Gene Expression in Mouse Hepa 1c1c7 Wild-type and Mutant Cell Lines. In preparation,

Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M.,

Brown, P. O., Botstein, D. and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res, **31**, 94-6.

Grant, G. R., Manduchi, E., Pizarro, A. and Stoeckert, C. J., Jr. (2003) Maintaining data integrity in microarray data management. Biotechnol Bioeng, **84**, 795-800.

Hautaniemi, S., Edgren, H., Vesanen, P., Wolf, M., Jarvinen, A. K., Yli-Harja, O., Astola, J., Kallioniemi, O. and Monni, O. (2003) A novel strategy for microarray quality control using Bayesian networks. Bioinformatics, **19**, 2031-8.

Hessner, M. J., Meyer, L., Tackes, J., Muheisen, S. and Wang, X. (2004) Immobilized probe and glass surface chemistry as variables in microarray fabrication. BMC Genomics, **5**, 53.

Jarvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O. P. and Monni, O. (2004) Are data from different gene expression microarray platforms comparable? Genomics, **83**, 1164-8.

Kwekel, J. C., Dalgleish, H. A., Burgoon, L. D., Harkema, J. R. and Zacharewski, T. R. (2005) Analysis of gene expression during uterine Induction and regression in immature, ovariectomized rats following treatment with ethynyl estradiol. In preparation,

Miles, M. F. (2001) Microarrays: lost in a storm of data? Nat Rev Neurosci, **2**, 441-3.

Model, F., Konig, T., Piepenbrock, C. and Adorjan, P. (2002) Statistical process control for large scale microarray experiments. Bioinformatics, **18 Suppl 1**, S155-63.

Petri, A., Fleckner, J. and Matthiessen, M. W. (2004) Array-A-Lizer: A serial DNA microarray quality analyzer. BMC Bioinformatics, **5**, 12.

Sun, Y. V., Boverhof, D. R., Burgoon, L. D., Fielden, M. R. and Zacharewski, T. R. (2004) Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. Nucleic Acids Res, **32**, 4512-23.

Ulrich, R. G., Rockett, J. C., Gibson, G. G. and Pettit, S. D. (2004) Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. Environ Health Perspect, **112**, 423-7.

Wang, X., Ghosh, S. and Guo, S. W. (2001) Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Res, **29**, E75-5.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, **30**, e15.

CHAPTER FOUR

Burgoon, L.D., Fong, C.J., Eckel-Passow, J.E., Gennings, C., Zacharewski, T.R. (2005) An Empirical Comparison of Genomic Experimental Designs for Temporal Studies. Bioinformatics, in submission.

Abstract

Motivation:

To effectively utilize microarrays, studies must be appropriately designed to ensure that the biological question of interest is properly addressed. Little guidance exists concerning experimental designs to identify active genes in a single dose, temporal experiment involving time-matched vehicle controls. To this end, two-color microarray assays were conducted to generate separate and independent temporal datasets from one *in vivo*, and two *in vitro* studies that incorporate the independent reference, loop, and modified loop designs.

Results:

All three designs resulted in different active gene lists, with varying degrees of overlap. The modified loop design included the most technical replicates, and consistently exhibited the largest active gene list. As the choice of experimental design significantly affected the overall biological interpretation of the data, the modified loop design is preferred when using the same number of biological replicates due to the larger number of technical replicates and to facilitate temporal treatment comparisons.

Introduction

Experimental design holds immense gravity with respect to the analysis methods, results, and interpretation of a study. Inappropriate designs and analysis methods may confound interpretation and lead to inappropriate hypotheses. For example, the use of an incongruous experimental design may lead to the generation of an inaccurate expression signature for toxicity in drug candidate screening or a hazard identification program.

Several experimental designs have emerged for the analysis of temporal gene expression effects including the reference design (Yang and Speed, 2002), independent reference design (Fielden, *et al.*, 2002), loop design (Kerr and Churchill, 2001a; Kerr and Churchill, 2001b; Yang and Speed, 2002), and modified loop design (Boverhof, *et al.*, 2004) (Figure 4-1).

The reference and independent reference designs (RD and IRD, respectively) are the most intuitive of the experimental designs. The key feature of the reference design is that all microarrays receive the same reference sample such that it is consistently represented with the same dye on each array. This implies that both treated and vehicle samples (i.e., the primary samples of interest) are cohybridized with the same reference sample on each array. Oftentimes, investigators will perform dye-swaps, where each sample is labeled with both dyes an even number of times to avoid possible dye biases (Cox, *et al.*, 2004; Irwin, *et al.*, 2004).

With regard to a temporal experiment, the IRD (Figure 4-1B) differs from a RD in that each time point has a matched vehicle-control. To model the effect due to the dyes, a dye-swap design is encouraged where each sample is balanced with respect to dyes (i.e., Cy3 and Cy5). Although the IRD requires half the number of microarrays, and therefore

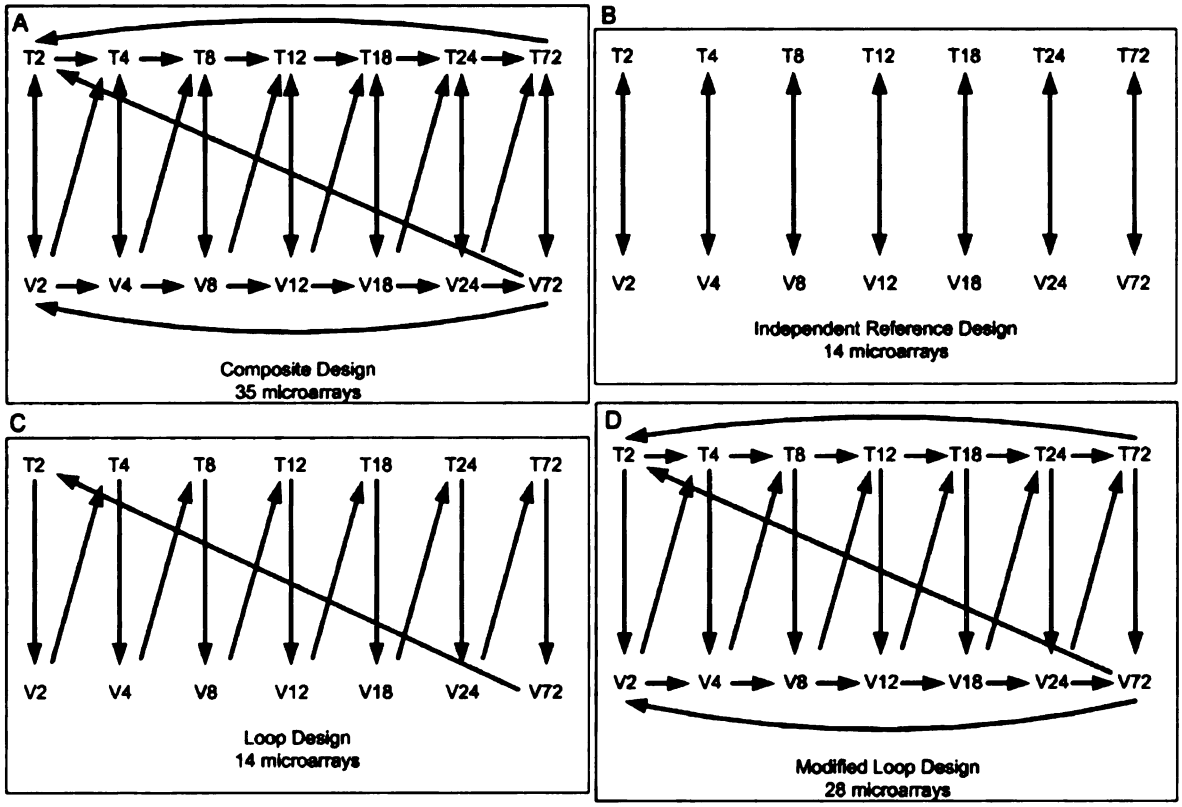

Figure 4-1: Microarray Study Designs. The composite design (A) is a combination of the independent reference (B), loop (C), and modified loop designs (D). The composite design is a non-redundant merger making it an economical method for study design comparisons. Each arrow represents a microarray, with the heads and tails each representing a different dye (e.g., Cy3 or Cy5). The independent reference design is the simplest design where comparisons are made between treated and vehicle samples where each time-point is treated independently. The loop design is an interconnected, balanced design where each treatment/vehicle sample is labeled with each dye equally. The modifications in the modified loop design are two additional loops, one for each treatment variety (i.e., treatment or vehicle). These additional loops serve to increase the technical replication and enhance the ability to make temporal comparisons within a treatment variety. T represents treated and V represents vehicle varieties while the numbers indicate the time-point.

less starting material than the reference design, temporal confounds exist that may compromise analyses across time (i.e., testing hypotheses that there are no changes in treatment effect across time).

The loop design (LD; Figure 4-1C) was developed as an alternative to the reference design that provides balanced measurements across the design (i.e., the same

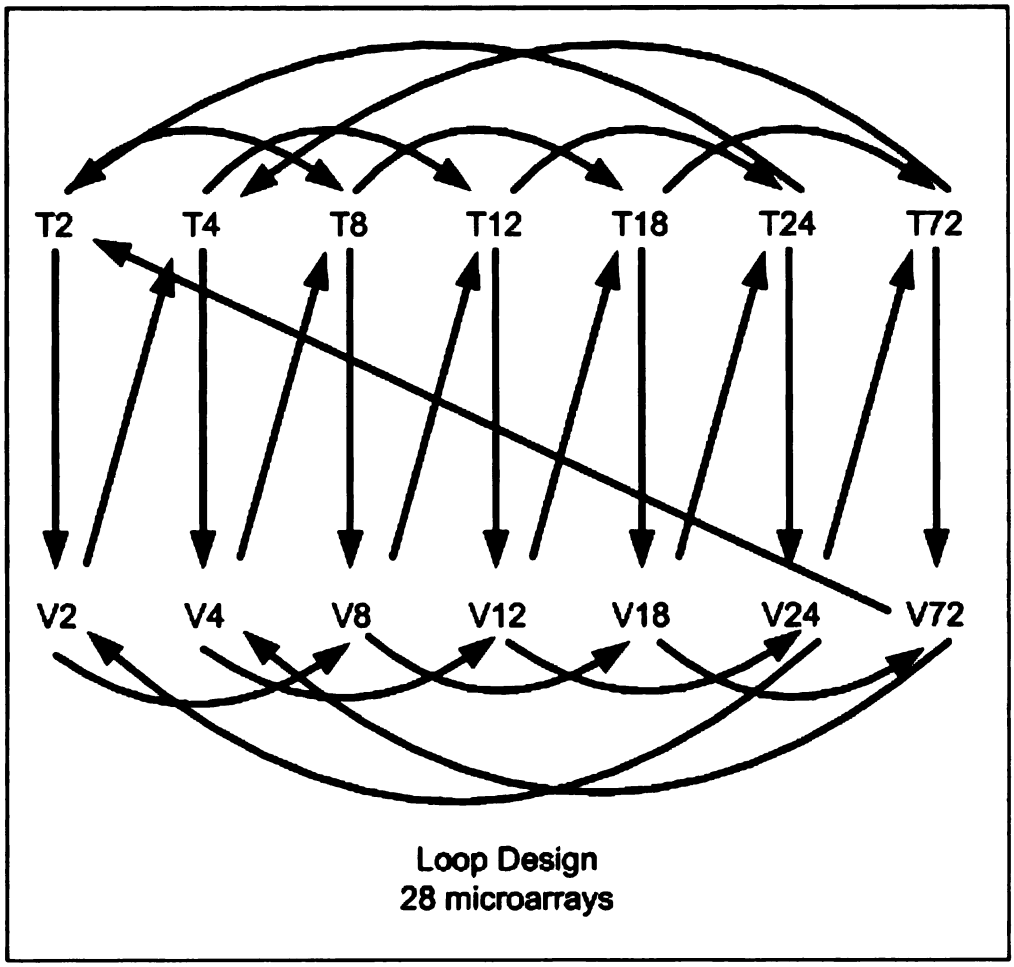

Figure 4-2: A-optimal Study Design. The A-optimal design for the 28 variety (i.e., 2 treatment varieties x 7 time-points) experiment represents the design exhibiting the lowest average error for arrays and varieties. This design does not make comparisons among adjacent time-points within variety, unlike the modified loop design.

number of measurements are made per treatment group). This alleviates the need to generate massive quantities of the reference sample as required with the reference design, and minimizes acquisition of large amounts of data from the generally uninformative reference sample (Kerr and Churchill, 2001a). The loop design is also more economically feasible than the independent reference design, requiring half the number of microarrays while accounting for dye effects.

The modified loop design (MLD; Figure 4-1D) consists of the standard loop design, augmented by two "inner" loops, one for each class of treatment variety (i.e.,

treatment and vehicle). For time-course experiments with treatment and vehicle comparisons, each treatment-by-time combination is referred to as a variety. In other words, for a 2 × 4 factorial case (2 treatments × 4 time points), there are eight varieties. For time-course experiments the loop design is A-optimal (i.e., the design that exhibits the smallest average variance for comparisons of interest) for four or fewer time points (Kerr and Churchill, 2001a). However the A-optimal design for a 14-variety experiment (2 treatments × 7 time points; Figure 4-2; adapted from output from the Experimental Design Tool: http://exgen.ma.umist.ac.uk/) may not be the most appropriate design for a time-course study. Under this design, temporally adjacent varieties (i.e., treated at time n and n+1) are not connected. Connectedness of varieties is preferred (i.e., varieties to be compared are assayed on the same microarray) to decrease technical variation within the comparison. Note that the A-optimal design only minimizes the average variance across all possible comparisons, not necessarily the ones of interest. Thus, if investigators are interested in comparing treated and vehicle samples from the same time point as well as adjacent time-points within a treatment class then the modified loop design is more appropriate with regard to decreasing the variability associated with each comparison.

Others have investigated the differences between the RD and LD (Dobbin and Simon, 2002; Simon, *et al.*, 2002; Vinciotti, *et al.*, 2004). It has been shown that the RD outperforms the LD when sample size is limited and class discovery is the primary goal **(Dobbin and Simon, 2002; Simon, *et al.*, 2002)**. However, the LD provides greater precision, and may be more appropriate when identifying differentially expressed genes (Vinciotti, *et al.*, 2004). Thus, the RD was not included in these studies since chemical classification, analogous to the sample classification problems, may be more

appropriately probed using RD, and that the LD is optimal for the identification of differentially expressed genes. Here we present an empirical comparison of the independent reference, loop, and modified loop designs, with respect to three temporal experiments engaged in identifying treatment responsive genes.

Materials and Methods

Microarray Study Design

A composite design (CD; Figure 4-1A) was employed that combined the IRD, LD, and MLD into a single, nonredundant representation of each dye-sample combination. This novel design limits data redundancy, reduces errors due to technical variation and minimizes confounding factors with run order that may arise if each design had been completed separately.

The CD was used to generate three datasets from three independent experiments. Experiment 1 (Exp-1) is an *in vitro* time course investigating gene expression changes following treatment with DMSO, a common vehicle typically used for *in vitro* experiments that is generally considered to be innocuous, compared to time-matched untreated (niave) controls. Experiment 2 (Exp-2) is a 17-β estradiol elicited *in vitro* gene expression time course study that includes time-matched vehicle (DMSO) treated cells. Experiment 3 (Exp-3), an *in vivo* experiment, that compares the temporal effects of 17-α ethynyl estradiol to a time-matched vehicle (sesame oil) control in murine liver tissue. All experiments are 7 time x 2 treatment (14 variety) experiments with three biological replicates.

Further details on microarray assay procedures are available at http://dbzach.fst.msu.edu. All animal husbandry and sample collection procedures were approved by the Michigan State University All University Committee on Animal Use and Care.

Data Normalization and Active Genes Filtering

Microarrays were scanned using an Affymetrix 428 scanner, and images were quantified using GenePix v5.0 or v5.1. Microarray data were normalized using a semiparametric normalization method that accounts for intensity-dependent effects (Eckel, *et al.*, 2004b) and active genes were identified using an empirical Bayes method (Eckel, *et al.*, 2004a). Active genes are defined as those with a posterior probability of being differentially expressed larger than 0.95. Normalization was performed in SAS v8.02; the empirical Bayes method was performed in SAS v8.02 and R v1.9.1.

Design Comparison Methods

Active gene lists were compared by creating tables of the active genes in SAS, and performing inner joins across the tables to identify overlapping active cDNAs.

Box-and-whisker plots were used to compare the standard error estimates from the model-based t-statistic of the General Linear Mixed Model (GLMM) across the three experimental designs and within each experiment. The GLMM is a linear effects model where the response variable, or normalized expression value, is modeled as a linear function of both fixed and random effects (e.g., microarray, treatment, date of hybridization, etc). The model-based t-statistic is the estimate, theta, from the GLMM divided by the standard error of theta. For the purposes of this analysis, only treatment

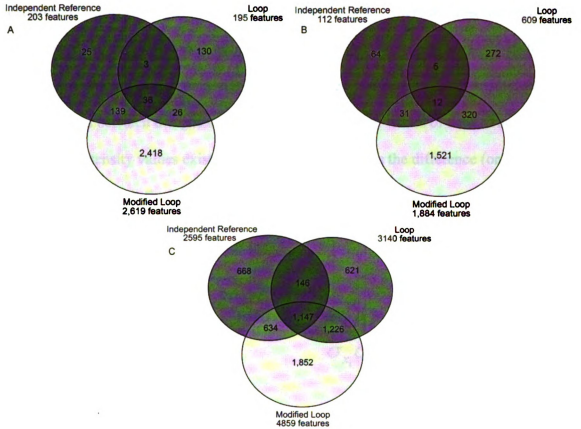

Figure 4-3: Comparison of Active Gene Lists. Active gene lists were generated and compared for each design within each experiment. The MLD consistently yields the largest active gene list. The degree of overlap between the lists is dependent upon the experiment, not the design. Comparisons are shown for the Exp-1 (A), Exp-2 (B), and Exp-3 (C) experiments.

and time-matched control comparisons are being made. The box represents the interquartile range (IQR), where the lower bound is the 25th percentile and the upper bound is the 75th percentile. The whiskers represent the fence, where the upper bound is the 75th percentile plus 1.5 times the IQR; the lower bound is the 25th percentile minus 1.5 times the IQR. Values outside of the fence are represented by asterisks in the plot.

Mean estimates were calculated as the arithmetic mean of the normalized feature intensity within a treatment variety per cDNA. Mean estimates were compared using 1) distributions of temporal correlations, 2) trajectory plots, and 3) 45-degree rotated

scatterplots, similar to the M vs A plot (Yang, *et al.*, 2002), where the abscissa represents the geometric mean of the mean estimates, and the ordinate axis represents the difference of the mean estimates between designs. These rotated scatterplots illustrate intensity relative biases between designs as deviations from ordinate values of zero. The normalized intensity values exist within $\log_2$ space and thus the difference (on the ordinate axis) reflects the $\log_2$ ratio of the mean estimates.

The temporal correlation (ρ) is defined per gene *i* as

$$\rho_i = \frac{\sigma_{x_i y_i}}{\sigma_{x_i}\sigma_{y_i}} = \frac{\varepsilon[(x_{i_t} - \mu_{x_i})(y_{i_t} - \mu_{y_i})]}{\sigma_{x_i}\sigma_{y_i}}$$

where $\sigma_{x_i y_i}$ represents the covariance for gene *i*; σ_{x_i} and σ_{y_i} represent the variances across time for designs x and y, respectively; μ_{x_i} and μ_{y_i} represent the arithmetic mean of the normalized gene expression value for the vectors x_i and y_i; x_{i_t} and y_{i_t} represent the mean estimates at time *t*. The $\varepsilon[Q]$ notation represents the arithmetic mean of the vector quantity Q. Histograms representing the distribution of correlations were generated for each pair of design comparisons (IRD vs LD, IRD vs MLD, LD vs MLD) for each of the three experiments.

Mean estimates (as calculated above) were also compared using trajectory plots. Principal components analysis (PCA) was performed on the treated and vehicle mean estimates together, per each design and experiment. Trajectory plots are three

dimensional scatterplots of eigenvalues from the first three principal components (PCs) of the PCA. Points within a treatment-variety are connected by line segments in temporal order, creating a treatment-variety surface (i.e., treatment or vehicle surfaces).

Table 4-1: Percent Overlap Between Active Gene Lists For Each Design and the Modified Loop Design

EXP-1	
IREF: % overlap with Modloop	86.2%
LOOP: % overlap with Modloop	31.8%
Exp-2	
IREF: % overlap with Modloop	38.3%
LOOP: % overlap with Modloop	54.5%
EXP-3	
IREF: % overlap with Modloop	68.6%
LOOP: % overlap with Modloop	75.6%

Visualizations and Statistical Analyses

All statistical analyses and scatterplots were performed/generated in SAS v8.02, unless otherwise noted within the referenced material. All other data visualizations were performed/generated in R v1.9.1.

Results

IRD, LD, and MLD Yield Different Active Gene Lists

Differences in variance and mean estimates of gene expression in each design can be attributed to differences in the active gene lists (Figure 4-3). The MLD consistently

Table 4-2: 99th Percentile of the Standard Error Distributions			
	Exp-1 (99th percentile)	*Exp-2* (99th percentile)	*Exp-3* (99th percentile)
IRD	0.345	1.056	0.216
LD	0.321	0.464	0.242
MLD	0.270	0.439	0.198

exhibits the largest active gene lists, with no clear pattern exhibited by the IRD and LD. The LD and the MLD exhibited the largest concordance in the Exp-2 and Exp-3 experiments, while the IRD and MLD exhibited the largest concordance in the Exp-1 experiment (Table 4-1).

Comparison of Standard Error Estimates

The standard error is used in the estimation of the model-based t-statistic, which is used to calculate a posterior probability for determination of the active gene list. A small standard error will inflate the t-statistic while a large standard error will deflate the t-statistic. In the Exp-1 and Exp-2 datasets, the MLD exhibits less variance than the IRD and LD (Figure 4-4A, B). However, in the Exp-3 dataset, the MLD and IRD exhibited similar degrees of variance, while the LD exhibited more variance based on comparisons of the interquartile ranges (Figure 4-4C). These relationships hold when comparing the 99th percentiles from the standard error value distributions (Table 4-2), with the exception that the MLD has less error than the IRD in Exp-3. The extreme values are not used in these comparisons as they are not representative of the majority of the data points. Thus, the MLD exhibits less variance than the LD, which is expected, as the two designs are directly related, with the exception that MLD harbors more technical replicates than the LD. The temporal confound exhibited by the IRD makes interpretation of the standard error differences difficult between the IRD, MLD, and LD.

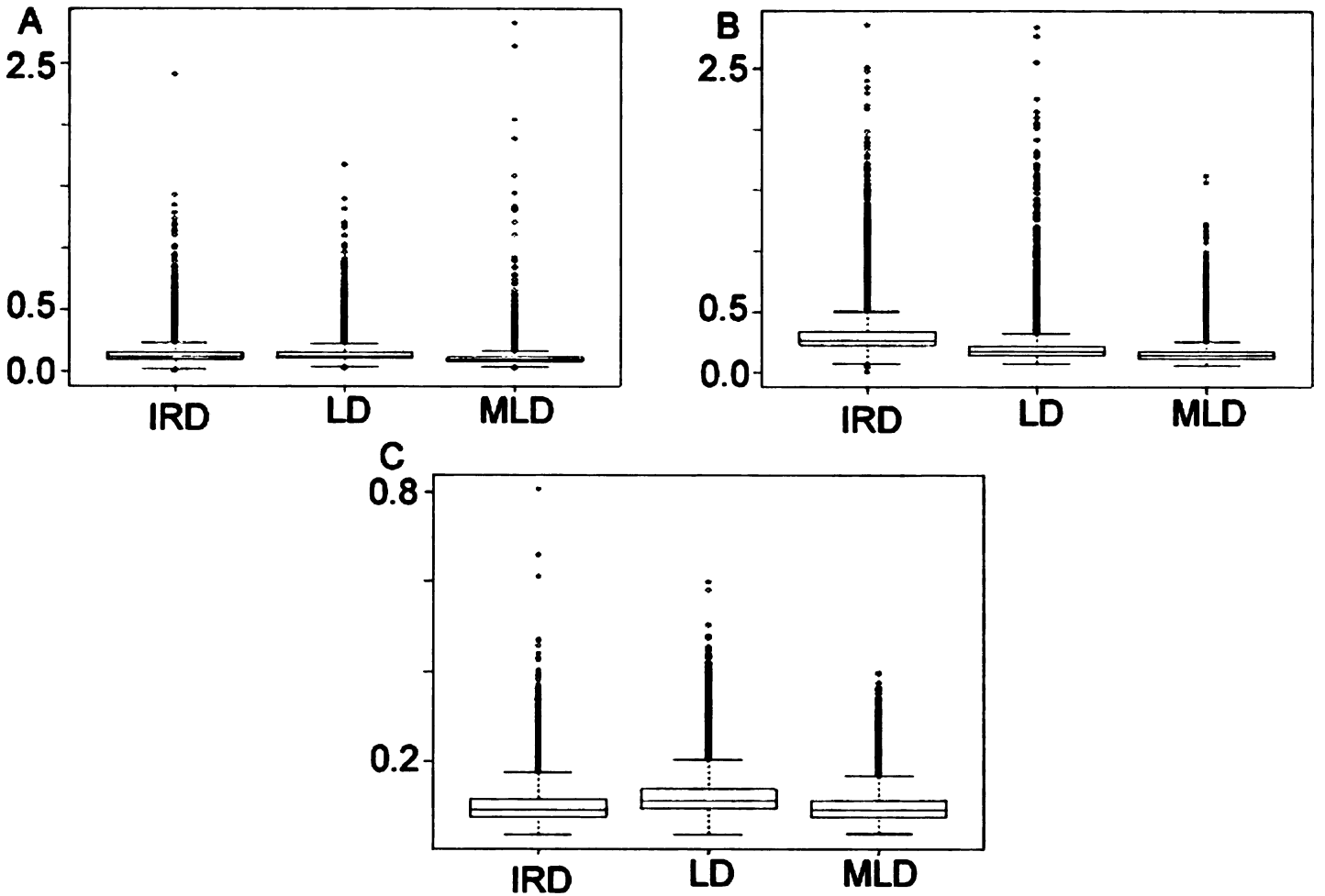

Figure 4-4: Comparison of global variance. The global variance is defined as the estimate of the standard error from the GLMM used in the calculation of the model-based t-value. The modified loop design provides the smallest standard errors for the Exp-1 (A), Exp-2 (B), Exp-3 (C) experiments.

Comparison of Mean Estimates

Mean estimates are used both in the calculation of the model-based t-statistic as well as calculating the treatment related fold change with respect to time-matched vehicle controls. The LD and MLD tend to yield more similar mean estimates based on their temporal correlation (Figure 4-5) in all three experiments. This is evidenced by the leftward skewed distributions, with correlation coefficients more skewed towards +1. Thus, the overall patterns, or trends, obtained from the data tend to be similar between the LD and MLD, but correlation says little about the absolute concordance of these estimates.

To examine the concordance of the estimates a 45 degree rotated scatterplot, similar to the M vs A plot is used (Figure 4-6). The x-axis represents the geometric mean of the mean estimates, while the y-axis represents the difference of the mean estimates from the two designs being compared. Differences were exhibited by all designs in all three experiments; however, the greatest differences occurred between the MLD and the IRD and LD in the Exp-3 experiment, with the largest difference exhibited between MLD and the IRD, where the average difference was approximately 2-fold (a difference of 1) at the lower mean estimates, which tapers back to an average of zero difference at the high mean estimates.

To further compare the means and the temporal relationships, a specialized PCA plot (aka trajectory plot), similar to those developed for metabonomics (Keun, *et al.*, 2004), was used. Trajectory analysis projects temporal microarray data into three dimensions, each representing a principal component from the PCA. Line segments are

used to connect each time-point within a treatment group, such that each join point is a treatment and time-point combination. For example, Figure 4-7C has the vertices labeled. The starting and end-points of the trajectories are of less interest compared to the overall shapes that are conveyed by the trajectories. The amount of variance explained by each principal component (PC), and the total amount of variance explained by the first three principal components is given in Table 4-3.

The trajectories in the IRD suggest the design is confounded with respect to time, as the treated and vehicle points tend to cluster closely based on time with congruent, or similarly shaped, surfaces (Figure 4-7A-i, B-i, C-i). Surfaces do not need to overlap or be superimposed on one another to be congruent, they simply need to convey similar shapes. The temporal congruency is lost in the MLD of all three of the experiments (Figure 4-7A-iii, B-iii, C-iii). The LD exhibits less temporal clustering than the IRD, and resembles an amalgamation of the IRD and MLD (Figure 4-7A-ii, B-ii, C-ii). Furthermore, the trajectory analysis further supports the notion that mean estimates from the designs differ greatly.

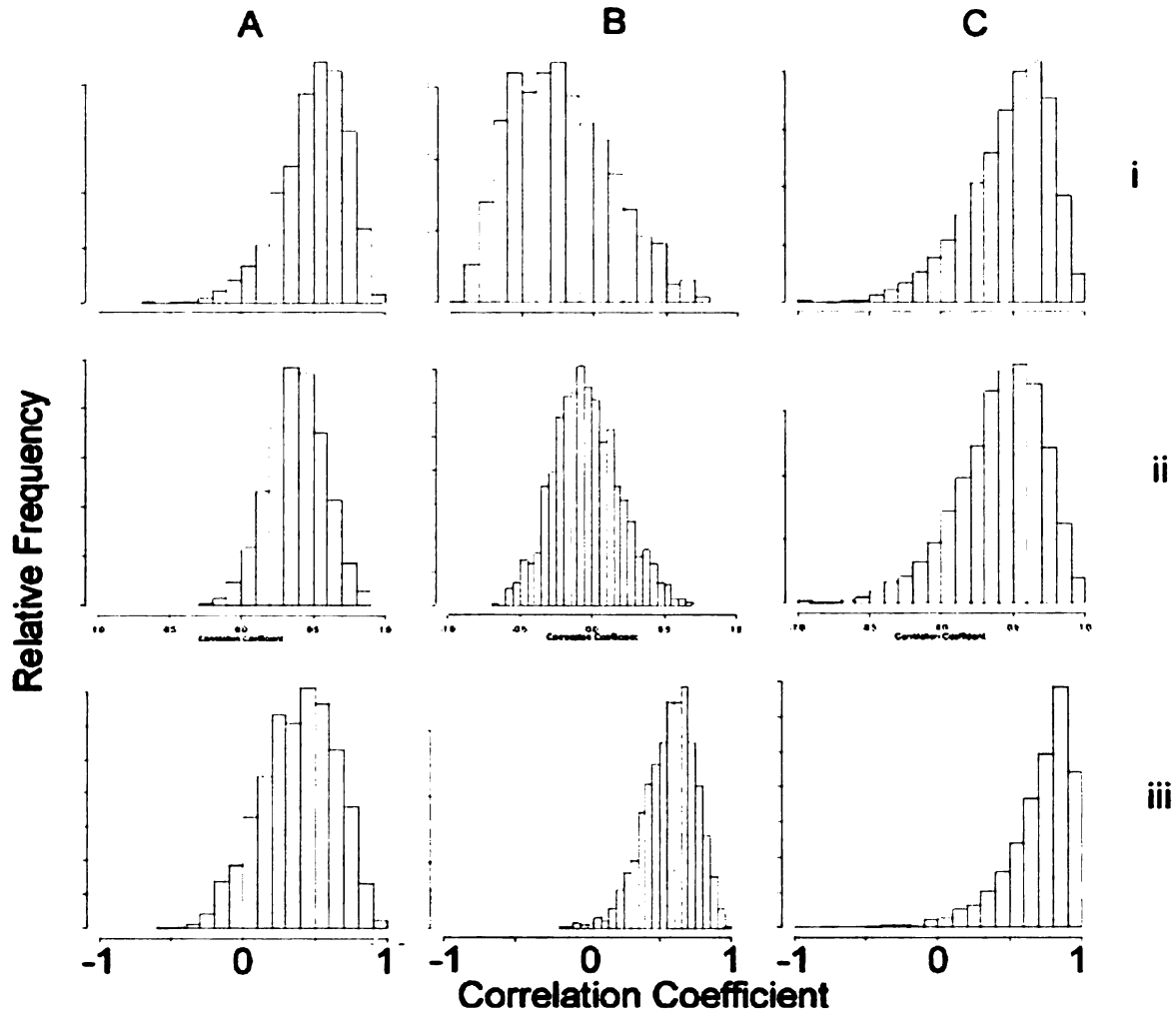

Figure 4-5: Correlations of Mean Estimates Across Design. Normalized mean estimates across designs were examined for temporal correlations within experiments. Correlation coefficients were calculated between designs plotted on histograms. Exp-1 (A), Exp-2 (B), and Exp-3 (C) experiments all show greater similarity between temporal correlations of MLD and LD (A-C-iii), than the other pairwise design comparisons. A-C-I represents the IRD and LD, A-C-ii represents the IRD and MLD, and A-C-iii represents the LD and MLD distributions of correlation coefficients.

Discussion

The choice of experimental design holds significant gravity over the results and their interpretation. Application of different designs results in different experimental interpretations, such as the identification of different biomarkers of exposure.

Three different designs were evaluated to identify their appropriateness for studying temporal changes in gene expression following exposure to a chemical or an appropriate control (i.e., time-matched vehicle). To compare the independent reference (IRD), loop, and modified loop designs (MLD), investigators used a unique design that combines all three (i.e., the composite design), ensuring that as many arrays as possible were shared between the designs, thus limiting the influence of technical (e.g., labeling reaction, microarray) and biological error (e.g., biological sample) on the comparison.

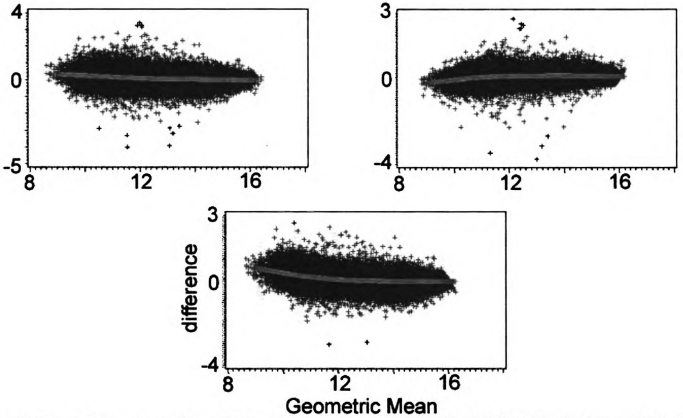

Figure 4-6: Rotated Scatterplot Comparing Mean Estimates Across Design. Plots of (A) IRD vs LD, (B) LD vs MLD, (C) IRD vs MLD illustrate gross differences between the designs. Depicted here are the plots for Exp-3. Each point represents a cDNA on the array, where the x-axis is the geometric mean of the mean estimates from the respective designs, while the y-axis is the $\log_2$ difference between the mean estimates, making it similar to the M vs A plot. If the designs yield the same estimates for a cDNA, then the difference would equal 0. The line represents the LOWESS fit to the data, and generalizes the overall pattern illustrating the trend across the estimates. This allows mean estimate biases inherence to the design to be identified. For example, the largest difference in mean estimates occurs between the IRD and MLD with mean estimates in the MLD approaching a 2-fold difference in normal space at lower estimate values.

Each design was examined using three independent composite datasets. Exp-1 and Exp-2 were performed in the murine Hepa1c1c7 cell line. Exp-2 featured a vehicle control at each time point, while the Exp-1 time course used untreated (niave) cells as the control. Exp-3 is an in vivo time course experiment examining the effects of ethynyl estradiol in the mouse liver.

The designs were compared by examining the temporal trajectories of the gene expression profiles and the active gene. Differences were examined further by identiying the amount and sources of variance and comparing estimated means across the designs.

The trajectory analysis confirmed that the IRD exhibited a temporal confound that compromised the detection of treatment effects. The overlap of treatment-variety surfaces (i.e., treatment and vehicle surfaces) indicated no difference was observed between the mean estimates. Treatment-variety surfaces that do not exhibit superposition, but are temporospatially congruent have treatment varieties that are confounded by time; that is, the temporal variance cannot be distinguished from the treatment variance. Although the mean estimates between treatment varieties may be different, the temporal confound renders time-point comparisons impossible due to the inability to separate temporal and treatment variances. Thus, hypotheses concerning treatment comparisons between times are not testable. This prohibits IRD as an appropriate design for probing temporal relationships of treatments, such as those required for kinetic modeling of a response. However, the IRD can be used to compare treatment and vehicle exposures that are independently considered at each time point.

Further insight is gained when interpreting the trajectory results from studies examining the temporal gene expression effects of DMSO, a common vehicle used in *in*

vitro studies. Trajectory analysis from the IRD suggests there is little difference in gene expression patterns across time between DMSO and untreated cells in culture, and that there is a great deal of temporal variance in both DMSO-treated and untreated cells. In contrast, MLD data suggests few temporal effects are exhibited in untreated cells, as depicted by the points being in relatively close-proximity to each other, while the effects due to DMSO-treatment are much greater, on the relative scale, with the cells returning to the "untreated state" by 48hrs. Given the temporal confound exhibited by the IRD, it is likely that the temporal interpretation from the MLD is more accurate, where untreated cells show significantly fewer temporal changes across time, and the DMSO treatment by itself generates a much different change in gene expression, with many of these effects being absent within 48hrs of treatment.

Large differences in the number and composition of active genes were observed between the three designs when the same P(1)t-value is chosen. The MLD yields the greatest number of active genes for all three experiments, which is not surprising given it includes more technical replication, and will generally exhibit less variance. However, the degree of overlap of active genes also differs greatly between experiments.

The three designs tend to yield differences in the mean estimates for active genes. Although these differences exist, the MLD and LD tend to yield estimates that are more closely correlated than the MLD and IRD. However, large degrees of scatter are still seen in the scatterplots, representing differences in the estimates. This difference in mean estimates also appears as differences in the temporal clustering between designs as evidenced by the distinctly different patterns in the trajectory analysis. For example, the

relative distance between time-points in the Exp-3 MLD and LD trajectories for the treated groups are quite different, with 12hr and 18hr relatively close in the MLD, and far

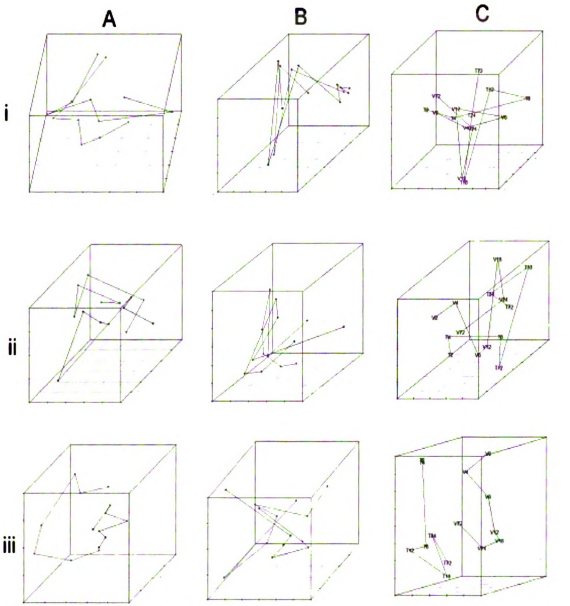

Figure 4-7: Trajectory Plots of Temporal Expression Changes. Eigenvalues were calculated from normalized mean estimates of gene expression across time to create 3D scatterplots, where the axes represent the three principal components that best represent the variance within the dataset. Design were compared within each experiment (i.e., Exp-1 (A), Exp-2 (B), and Exp-3 (C); IRD (i), LD (ii), MLD (iii)). A temporal confound is exhibited in the IRD across all of the experiments, exhibited as close proximity of treated and vehicle nodes for the same time-point, and similar shape patterns. The MLD illustrates greater treatment effects, relative to vehicle, as evidenced by the larger spatial distance compared to the IRD and LD. The labels in (C) represent the treatment variety and time-point, where T is treated and V is vehicle, and the number is the time-point in hours.

apart in the LD. This difference in the distances alone would skew the biological

interpretation with either relatively little treatment differences from 12-18hrs in MLD or larger differences in LD.

The primary difference between the MLD and LD is the number of technical replicates, where the MLD contains twice as many technical replicates. The MLD includes technical replicates of biological sample and dye interactions which decrease the amount of global variance in the MLD as compared to the LD. Increased technical replication also tends to increase the accuracy of mean estimates as illustrated by the different trajectory for each design. The MLD shows complete segregation of the treatment and vehicle spaces, while the LD still shows some similarities (i.e., close spatial proximity). The increased number of active genes at the same false positive rate is seen as a side effect of the decreased variance and increased accuracy. Thus, investigators gain the advantage of smaller false positive rates when limiting their active gene lists to a particular smaller size when using the P1(t)-value based cutoff method in the MLD as opposed to the LD.

However the MLD comes at a significant expense, both in terms of the number of microarrays required and the total amount of biological sample. Although advantageous compared to the LD and IRD, there are issues of practicality that may limit the use of the MLD. Technical replication allows an investigator to have more confidence in the expression level of a transcript within a particular sample, but biological, not technical, replication models the biological population's response to the treatment of interest. Therefore, it is best to sacrifice the technical replication for the sake of biological replication when sample or supply are limiting (Yang and Speed, 2002). To test this notion of the importance of biological vs technical replicates, an experiment comparing

Table 4-3: Percentage of Variance Explained By Each Principal Component

	Exp-1			*Exp-2*			*Exp-3*	
	IRD	**LD**	**MLD**	**IRD**	**LD**	**MLD**	**IRD**	**LD**
PC1	98.40%	98.80%	99.20%	96.50%	31.60%	33.00%	39.20%	52.60%
PC2	0.60%	0.30%	0.20%	1.90%	17.50%	16.70%	26.40%	14.20%
PC3	0.30%	0.20%	0.20%	0.50%	11.20%	10.70%	8.50%	7.50%
Total	99.20%	99.30%	99.60%	99.00%	60.30%	60.50%	74.00%	74.20%

the LD with an additional biological replicate to the MLD where the total number of microarrays is equivalent would be useful.

Conclusions

This analysis using three independent composite datasets illustrates that biological interpretation can be significantly influenced by experimental design. The design affected the estimated mean expression value, temporal clustering, and active gene lists which may lead to incorrect hypotheses regarding the temporal onset/occurrence of treatment-related effects. The MLD is the most appropriate design for temporal gene expression studies involving two treatment varieties, such as treatment and vehicle as it lacks the temporal confound exhibited by the IRD, and encompasses more technical replicates than the LD, ensuring more accurate normalized mean estimates. However, these advantages come at the cost of consumables and biological samples. Thus, the LD may be more appropriate when cost or the amount of biological sample is limiting. Advantages of the MLD are also overshadowed by the importance of biological replicates. Nevertheless, it is clear that technical replication does matter, both for estimation of standard error and the mean. Thus, experimental design considerations must include thoughtful analysis of the costs, sample requirements, and the underlying

biological questions to ensure data quality as ultimately, the experimental design choice may influence the biological interpretation of the data.

Acknowledgements

This work was supported by NIEHS grants ES 04911-12, ES 011271, and ES 011777.

TRZ is partially supported by the Michigan Agriculture Experiment Station.

References

Boverhof, D. R., Fertuck, K. C., Burgoon, L. D., Eckel, J. E., Gennings, C. and Zacharewski, T. R. (2004) Temporal and dose-dependent hepatic gene expression changes in immature ovariectomized mice following exposure to ethynyl estradiol. Carcinogenesis, **25**, 1277-91.

Cox, W. G., Beaudet, M. P., Agnew, J. Y. and Ruth, J. L. (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. Anal Biochem, **331**, 243-54.

Dobbin, K. and Simon, R. (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics, **18**, 1438-45.

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004a) Empirical bayes gene screening tool for time-course or dose-response microarray data. J Biopharm Stat, **14**, 647-70.

Eckel, J. E., Gennings, C., Therneau, T. M., Boverhof, D. R., Burgoon, L. D. and Zacharewski, T. R. (2004b) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, in press.

Fielden, M. R., Halgren, R. G., Fong, C. J., Staub, C., Johnson, L., Chou, K. and Zacharewski, T. R. (2002) Gestational and lactational exposure of male mice to diethylstilbestrol causes long-term effects on the testis, sperm fertilizing ability in vitro, and testicular gene expression. Endocrinology, **143**, 3044-59.

Irwin, R. D., Boorman, G. A., Cunningham, M. L., Heinloth, A. N., Malarkey, D. E. and Paules, R. S. (2004) Application of Toxicogenomics to Toxicology: Basic Concepts in the Analysis of Microarray Data. Toxicol Pathol, **32**, 72-83.

Kerr, M. K. and Churchill, G. A. (2001a) Experimental Design for Gene Expression Microarrays. Biostatistics, **2**, 183-201.

Kerr, M. K. and Churchill, G. A. (2001b) Statistical design and the analysis of gene expression microarray data. Genet Res, **77**, 123-8.

Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., Lindon, J. C. and Nicholson, J. K. (2004) Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. Chem Res Toxicol, **17**, 579-87.

Simon, R., Radmacher, M. D. and Dobbin, K. (2002) Design of studies using DNA microarrays. Genet Epidemiol, **23**, 21-36.

Vinciotti, V., Khanin, R., D'Alimonte, D., Liu, X., Cattini, N., Hotchkiss, G., Bucca, G., De Jesus, O., Rasaiyaah, J., Smith, C. P., Kellam, P. and Wit, E. (2004) An experimental evaluation of a loop versus a reference design for two-channel microarrays. Bioinformatics,

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, **30**, e15.

Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. Nat Rev Genet, **3**, 579-88.

CHAPTER FIVE

Summary and Conclusions

Sequencing of the human genome has ushered in the use of new large-scale technologies for the study of chemical effects in biological tissues and systems in the emerging field of toxicogenomics. A major hurdle in the adoption of these technologies is the implementation of cost effective data management schemes to manage the hordes data. However, to realize the full benefit of the technology, investigators must ensure the high quality of the data and that appropriate experimental designs are used.

The dbZach System, a combined database and analysis system, provides toxicogenomic data management capabilities for small laboratories, departments, and consortia. The relational database backend has been designed to faithfully and appropriately model biological relationships in a modular fashion. This decreases the time it takes a new biological investigator to become familiar with the system, and facilitates the incorporation of new technologies as they develop. The software capabilities of the system include upload, visualization, and mining of data.

The database is divided into several interconnected subsystems, or collections of tables. Each self-contained subsystem models a distinct biological concept or technology, such as cDNA clones, genes, microarrays, real-time PCR, etc, providing a modular database structure. As new technology develops, new modular subsystems can be integrated into dbZach without disruption of the current data management landscape. These new database back-end developments appear completely invisible to the user, allowing for seamless integration of nascent data types across time.

The development of the dbZach System has allowed for large-scale, multivariate analysis of trends within microarray data across experiments. Observations from these analyses lead to the development of high and low quality historical datasets, which were

instrumental in the development of novel quality assurance and control (QA/QC) protocols. These protocols have lead to improvements in investigator performance, and coupled with investigator experience, improved the results generated within the laboratory.

The current QA/QC protocol consists of three divisions, based on empirical observations from datasets of varying quality within dbZach. The Support Vector Machine (SVM), a statistical learning method for multivariate data classification, was used to generate a nonlinear mathematical model for identifying high from low quality microarrays. The method was improved by using a logistic regression to identify the most predictive variables prior to training the SVM. The most predictive variables included the Cy5 global feature to background intensity ratio, Cy3 and Cy5 global background intensity, and Cy3 global feature intensity. The Cy5 global feature to background intensity ratio was the most predictive variable, which is not surprising as Cy5 has been reported to be more sensitive than Cy3 to ozone (Fare, *et al.*, 2003).

The second division of the QA/QC protocol leverages a semiparametric normalization procedure (Eckel, *et al.*, 2005) to identify microarray subgrids which are misaligned during the automated feature identification. Misaligned subgrids are those where the software misannotates a region of the region, such as aligning an entire subgrid row one row off. Misaligned subgrids appear as diagonal lines in the modified MA plots generated as a result of the normalization procedure. This step is of the utmost importance as feature identification and microarray quantification software often fail to properly align microarrays, especially in the presence of high local background, and misaligned grids are often overlooked by the user.

The third division of the protocol analyzes the distribution of median feature intensities on a per array basis. The distribution of median feature intensities may influence the downstream statistics, including gene activity and the estimate of the mean expression. For these reasons, the distribution of feature intensities is controlled such that all experiments within the laboratory must follow the same distribution. This distributional standardization also facilitates comparisons across tissues and chemicals.

As the microarray technology has continued to mature so has the field of microarray experimental design; however, little guidance existed as to the most appropriate design under different circumstances. This prompted the comparison of three temporal experimental designs: the independent reference, loop, and modified loop designs with regards to time-course toxicology studies. These designs were compared using three independent experiments investigating the temporal response of 1) cells in culture to DMSO, a vehicle commonly used in *in vitro* experiments, to untreated cells, 2) cells in culture to 17β-estradiol compared to DMSO treated vehicle controls, and 3) mice to 17α-ethynylestradiol compared to sesame oil vehicle controls.

The experimental designs yielded different active gene lists, with varying degrees of overlap within each experiment. The modified loop design consistently exhibited the largest active gene list, likely due to the increased number technical replicates. The independent reference design exhibited a temporal confound, while the modified loop design exhibited a complete mixing of the samples. The loop design appeared as a mix of the independent reference and modified loop designs, with considerably less of a temporal confound.

Based on these results, investigators should consider using the loop and modified loop designs in lieu of the independent reference design. If only interested in comparing responses within time, and there is no interest in comparing the results across time, then the independent reference design would be appropriate, as the confounding of microarray, temporal, and treatment variances does not hinder the analysis. As the modified loop uses significantly more microarrays than the loop design, the potential benefits of the modified loop must be weighed against the economics of the loop design.

Generally, if the cost savings between the modified loop and loop designs are such that an additional biological replicate can be performed, then it would be advised to use the loop design. However, if the resources are available to perform the modified loop, and the cost difference is not enough to allow for an additional biological replicate, the increased accuracy afforded by the modified loop design may be justified.

Future Directions

The dbZach System is currently in a relative state of stability. The back-end database is well developed, and has proven capable of managing several microarray experiments. However, the present functionality of the system primarily resides within data upload of RT-PCR, microarray, and histopathology data, and minimal data interaction interfaces for unskilled users. Currently, it is being augmented by new microarray data mining tools, such as the Visualization Control Center (VCC).

The VCC provides data mining capabilities, such as plotting data in 2- and 3-D for visualization of data trends. It is being outfitted with pattern recognition algorithms, such as the k-means and agglomerative hierarchical clustering algorithms. These improvements will enhance investigator-centered data mining activities.

Future dbZach development will concentrate on the accommodation of new technologies, such as proteomics and metabolomics. New subsystems and upload interfaces will need to be created to manage these data. As these new data domains are captured, new cross domain data mining capabilities will also need to be developed for data mining of biological knowledge from combinations of genomic, proteomic and metabolomic data resident within the database.

The current state of the QA/QC protocol facilitates the identification of microarrays of questionable quality. However, future work should also focus on monitoring investigator performance which may include the use of Shewhart plots that illustrate the relationship between a weighted average, daily quality metric across time. The primary challenge will be the development of the quality metric; however, one example being the net number of high quality arrays produced per day. Investigator-based performance monitoring should lead to a net increase in laboratory data quality, as trends in investigator performance may be identified, and facilitate intervention, introspection, and further assay optimization when necessary.

A much larger goal is the formation of global quality metrics (Shi, *et al.*, 2004). The establishment of global quality metrics is important for performing comparisons of data from different laboratories using data within repositories. In the case of regulatory agencies, it is important as a basis of comparison of data from sponsors, and when performing risk assessments. For example, if a generic drug producer were to use microarray data to illustrate bioequivalence, the Office of Generic Drugs at the FDA may compare the signatures seen in the microarray results from the sponsor to results obtained from the initial patent holding sponsor, an independent third party, or from within FDA,

in an attempt to verify the bioequivalence and to identify putative signatures of toxicity. In these cases, the FDA would need to ensure all of the data used in the comparison were of high quality, or else the results and interpretation may become skewed.

The results from the comparisons of the experimental designs illustrate that although comparisons of interest should be tested on the same arrays, as asserted by others (Kerr and Churchill, 2001; Yang and Speed, 2002), this arrangement must be considered carefully to avoid confounding of the temporal effect. Thus, these results suggest the use of the loop and modified loop designs is superior to the independent reference design when the intention is to make comparisons of chemical effect across time. However, the next step is to further define the appropriate use of the loop and modified loop designs based on sample sizes and analysis of statistical power. Using the current datasets as examples, the statistical power with regards to the empirical Bayes method could be calculated using methods similar to those reported for other microarray datasets and tests (Tempelman, 2005; Tsai, *et al.*, 2005; Wei, *et al.*, 2004).

Furthermore, with the deluge of statistical mechanisms for normalization and identifying active genes, the loop and modified loop designs from these datasets could be used to perform comparisons of the methods. By anchoring these comparisons with the results from real-time PCR experiments, it may be possible to assess methods using different designs, and conditions which will provides further guidance regarding the appropriate analysis methods to use when confronted with a particular design.

Conclusions

In 1999 the BISTI report (Biomedical Information Science and Technology Initiative; http://www.nih.gov/about/director/060399.htm) recommended more concerted

integration of computational approaches into biological sciences, and the development of better software for data analysis and infrastructure for data management, with the fruits of to be shared with the greater community. These recommendations are especially true for the omic technologies, where massive datasets are generated and require novel data management, quality assurance, and experimental design considerations. Providing the toxicogenomics community with a modular data management product ensures its utility can continue to evolve as new technologies are developed. Besides the obvious benefit of having the data properly managed, databases also provide a mechanism for developing novel quality assurance methodologies, a framework for experimental design comparisons, and facilitate data sharing with repositories. Thus, these software and hardware development efforts, when combined with conventional toxicology, facilitates more comprehensive and predictive safety assessments.

Literature Cited

Eckel, J. E., Gennings, C., Therneau, T. M., Burgoon, L. D., Boverhof, D. R. and Zacharewski, T. R. (2005) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, **21**, 1078-83.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y. and Wang, Y. (2003) Effects of atmospheric ozone on microarray data quality. Anal Chem, **75**, 4672-5.

Kerr, M. K. and Churchill, G. A. (2001) Experimental Design for Gene Expression Microarrays. Biostatistics, **2**, 183-201.

Shi, L., Tong, W., Goodsaid, F., Frueh, F. W., Fang, H., Han, T., Fuscoe, J. C. and Casciano, D. A. (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. Expert Rev Mol Diagn, **4**, 761-77.

Tempelman, R. J. (2005) Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol, **105**, 175-86.

Tsai, C. A., Wang, S. J., Chen, D. T. and Chen, J. J. (2005) Sample size for gene expression microarray experiments. Bioinformatics, **21**, 1502-8.

Wei, C., Li, J. and Bumgarner, R. E. (2004) Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics, **5**, 87.

Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. Nat Rev Genet, **3**, 579-88.

LITERATURE CITED

LITERATURE CITED

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. and Hogue, C. W. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res, **33 Database Issue**, D418-24.

Bader, G. D. and Hogue, C. W. (2000) BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics, **16**, 465-77.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res, **33 Database Issue**, D154-9.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S. A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J. and Winegarden, N. (2004a) Submission of microarray data to public repositories. PLoS Biol, **2**, E317.

Ball, C. A., Sherlock, G. and Brazma, A. (2004b) Funding high-throughput data sharing. Nat Biotechnol, **22**, 1179-83.

Boverhof, D. R., Burgoon, L. D., Tashiro, C., Chittim, B., Harkema, J. R., Jump, D. B. and Zacharewski, T. R. (2005) Temporal and dose-dependent hepatic gene expression patterns in mice provide new insights into TCDD-mediated hepatotoxicity. Toxicol Sci.

Boverhof, D. R., Fertuck, K. C., Burgoon, L. D., Eckel, J. E., Gennings, C. and Zacharewski, T. R. (2004) Temporal and dose-dependent hepatic gene expression changes in immature ovariectomized mice following exposure to ethynyl estradiol. Carcinogenesis, **25**, 1277-91.

Boverhof, D. R., Tam, E., Harney, A. S., Crawford, R. B., Kaminski, N. E. and Zacharewski, T. R. (2004b) 2,3,7,8-Tetrachlorodibenzo-p-dioxin induces

suppressor of cytokine signaling 2 in murine B cells. Mol Pharmacol, **66**, 1662-70.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, **29**, 365-71.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. and Sansone, S. A. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, **31**, 68-71.

Breitkreutz, B. J., Stark, C. and Tyers, M. (2003) Osprey: a network visualization system. Genome Biol, **4**, R22.

Buck, M. J. and Lieb, J. D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics, **83**, 349-60.

Burt, J. W., Burgoon, L. D., Humes, D., Kwekel, J. C., Harney, A. S. and Zacharewski, T. R. (2005) Effects of estrogen on immature, ovariectomized mice: A multi-approach, tissue-by-tissue comparison. In preparation.

Bushel, P. R., Hamadeh, H., Bennett, L., Sieber, S., Martin, K., Nuwaysir, E. F., Johnson, K., Reynolds, K., Paules, R. S. and Afshari, C. A. (2001) MAPS: a microarray project system for gene expression experiment information and data validation. Bioinformatics, **17**, 564-5.

Cheadle, C., Vawter, M. P., Freed, W. J. and Becker, K. G. (2003) Analysis of microarray data using Z score transformation. J Mol Diagn, **5**, 73-81.

Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet, **32 Suppl**, 490-5.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Birney, E. (2003) Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res, **31**, 38-42.

Cox, W. G., Beaudet, M. P., Agnew, J. Y. and Ruth, J. L. (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. Anal Biochem, **331**, 243-54.

Cox, C. (1999) Nietzsche: Naturalism and Interpretation, University of California Press, Berkeley, California.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. and Clamp, M. (2004) The Ensembl automatic gene annotation system. Genome Res, **14**, 942-50.

Dobbin, K., Shih, J. H. and Simon, R. (2003) Statistical design of reverse dye microarrays. Bioinformatics, **19**, 803-10.

Dobbin, K. and Simon, R. (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics, **18**, 1438-45.

Dombkowski, A. A., Thibodeau, B. J., Starcevic, S. L. and Novak, R. F. (2004) Gene-specific dye bias in microarray reference designs. FEBS Lett, **560**, 120-4.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) Pattern Classification.

Dudoit, S., Yang, H. Y., Callow, M. J. and Speed, T. (2002) Statistical methods for identifying differntially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, **12**, 111-139.

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004a) Empirical bayes gene screening tool for time-course or dose-response microarray data. J Biopharm Stat, **14**, 647-70.

Eckel, J. E., Gennings, C., Therneau, T. M., Burgoon, L. D., Boverhof, D. R. and Zacharewski, T. R. (2005) Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, **21**, 1078-83.

Edgar, R., Domrachev, M. and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, **30**, 207-10.

Efron, B. and Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol, **23**, 70-86.

Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A., Anagnostopoulos, A., Baldarelli, R. M., Baya, M., Beal, J. S., Bello, S. M., Boddy, W. J., Bradt, D. W., Burkart, D. L., Butler, N. E., Campbell, J., Cassell, M. A., Corbani, L. E., Cousins, S. L., Dahmen, D. J., Dene, H., Diehl, A. D., Drabkin, H. J., Frazer, K. S., Frost, P., Glass, L. H., Goldsmith, C. W., Grant, P. L., Lennon-Pierce, M., Lewis, J., Lu, I., Maltais, L. J., McAndrews-Hill, M., McClellan, L., Miers, D. B., Miller, L. A., Ni, L., Ormsby, J. E., Qi, D., Reddy, T. B., Reed, D. J., Richards-Smith, B., Shaw, D. R., Sinclair, R., Smith, C. L., Szauter, P., Walker, M. B., Walton, D. O., Washburn, L. L., Witham, I. T. and Zhu, Y. (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res, **33**, D471-5.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y. and Wang, Y. (2003) Effects of atmospheric ozone on microarray data quality. Anal Chem, **75**, 4672-5.

Fielden, M. R., Halgren, R. G., Dere, E. and Zacharewski, T. R. (2002a) GP3: GenePix post-processing program for automated analysis of raw microarray data. Bioinformatics, **18**, 771-3.

Fielden, M. R., Halgren, R. G., Fong, C. J., Staub, C., Johnson, L., Chou, K. and Zacharewski, T. R. (2002) Gestational and lactational exposure of male mice to diethylstilbestrol causes long-term effects on the testis, sperm fertilizing ability in vitro, and testicular gene expression. Endocrinology, **143**, 3044-59.

Fisher, R. A. (1962) The place of the design of experiments in the logic of scientific inference. Colloq. Int. Cent. Nat. Recherche Scientifique, **110**, 13-19.

Fong, C. J., Burgoon, L. D., Gupta, G., Humes, D. G. and Zacharewski, T. R. (2005a) Temporal Gene Expression Analysis of Mouse Hepa-1c1c7 Cells Treated with 17beta-Estradiol by cDNA Microarray. In preparation.

Fong, C. J., Burgoon, L. D. and Zacharewski, T. R. (2005) Comparative Microarray Analysis of Basal Gene Expression in Mouse Hepa 1c1c7 Wild-type and Mutant Cell Lines. Toxicol Sci, in submission.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC, Boca Raton, FL.

Glas, A. M., Kersten, M. J., Delahaye, L. J., Witteveen, A. T., Kibbelaar, R. E., Velds, A., Wessels, L. F., Joosten, P., Kerkhoven, R. M., Bernards, R., van Krieken, J. H., Kluin, P. M., van't Veer, L. J. and de Jong, D. (2005) Gene expression

profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. Blood, **105**, 301-7.

Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, D. and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res, **31**, 94-6.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, **286**, 531-7.

Grant, G. R., Manduchi, E., Pizarro, A. and Stoeckert, C. J., Jr. (2003) Maintaining data integrity in microarray data management. Biotechnol Bioeng, **84**, 795-800.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V. A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, **30**, 52-5.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, **32**, D258-61.

Hautaniemi, S., Edgren, H., Vesanen, P., Wolf, M., Jarvinen, A. K., Yli-Harja, O., Astola, J., Kallioniemi, O. and Monni, O. (2003) A novel strategy for microarray quality control using Bayesian networks. Bioinformatics, **19**, 2031-8.

Hessner, M. J., Meyer, L., Tackes, J., Muheisen, S. and Wang, X. (2004) Immobilized probe and glass surface chemistry as variables in microarray fabrication. BMC Genomics, **5**, 53.

Hood, L. and Perlmutter, R. M. (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol, **22**, 1215-7.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H.,

Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Birney, E. (2005) Ensembl 2005. Nucleic Acids Res, **33 Database Issue**, D447-53.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet, **2**, 343-72.

Irwin, R. D., Boorman, G. A., Cunningham, M. L., Heinloth, A. N., Malarkey, D. E. and Paules, R. S. (2004) Application of Toxicogenomics to Toxicology: Basic Concepts in the Analysis of Microarray Data. Toxicol Pathol, **32**, 72-83.

Jarvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O. P. and Monni, O. (2004) Are data from different gene expression microarray platforms comparable? Genomics, **83**, 1164-8.

Joint Directors of Laboratories. (1991) Data Fusion Lexicon.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D. and Kent, W. J. (2003) The UCSC Genome Browser Database. Nucleic Acids Res, **31**, 51-4.

Kerr, M. K. and Churchill, G. A. (2001a) Experimental Design for Gene Expression Microarrays. Biostatistics, **2**, 183-201.

Kerr, M. K. and Churchill, G. A. (2001b) Statistical design and the analysis of gene expression microarray data. Genet Res, **77**, 123-8.

Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., Lindon, J. C. and Nicholson, J. K. (2004) Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. Chem Res Toxicol, **17**, 579-87.

Kristensen, V. N., Sorlie, T., Geisler, J., Langerod, A., Yoshimura, N., Karesen, R., Harada, N., Lonning, P. E. and Borresen-Dale, A. L. (2005) Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: clinical implications. Clin Cancer Res, **11**, 878s-83s.

Kwekel, J. C., Dalgleish, H. A., Burgoon, L. D., Harkema, J. R. and Zacharewski, T. R. (2005) Analysis of gene expression during uterine Induction and regression in immature, ovariectomized rats following treatment with ethynyl estradiol. In preparation,

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. and Davis, R. W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A, **94**, 13057-62.

Lindsay, M. A. (2003) Target discovery. Nat Rev Drug Discov, **2**, 831-8.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature, **431**, 308-12.

Luyendyk, J. P., Mattes, W. B., Burgoon, L. D., Zacharewski, T. R., Maddox, J. F., Cosma, G. N., Ganey, P. E. and Roth, R. A. (2004) Gene expression analysis points to hemostasis in livers of rats cotreated with lipopolysaccharide and ranitidine. Toxicol Sci, **80**, 203-13.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, **33 Database Issue**, D54-8.

Mattes, W. B., Pettit, S. D., Sansone, S. A., Bushel, P. R. and Waters, M. D. (2004) Database development in toxicogenomics: issues and efforts. Environ Health Perspect, **112**, 495-505.

McKusick, V. A. (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edition. Johns Hopkins University Press, Baltimore, MD.

Miles, M. F. (2001) Microarrays: lost in a storm of data? Nat Rev Neurosci, **2**, 441-3.

Mischel, P. S., Cloughesy, T. F. and Nelson, S. F. (2004) DNA-microarray analysis of brain cancer: molecular classification for therapy. Nat Rev Neurosci, **5**, 782-92.

Model, F., Konig, T., Piepenbrock, C. and Adorjan, P. (2002) Statistical process control for large scale microarray experiments. Bioinformatics, **18 Suppl 1**, S155-63.

Moggs, J. G., Tinwell, H., Spurway, T., Chang, H. S., Pate, I., Lim, F. L., Moore, D. J., Soames, A., Stuckey, R., Currie, R., Zhu, T., Kimber, I., Ashby, J. and Orphanides, G. (2004) Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth. Environ Health Perspect, **112**, 1589-606.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R. and

Zdobnov, E. M. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res, **31**, 315-8.

Nicholson, J. K. and Wilson, I. D. (2003) Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov, **2**, 668-76.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C. and Afshari, C. A. (1999) Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog, **24**, 153-9.

Petri, A., Fleckner, J. and Matthiessen, M. W. (2004) Array-A-Lizer: A serial DNA microarray quality analyzer. BMC Bioinformatics, **5**, 12.

Pruitt, K. D. and Maglott, D. R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res, **29**, 137-40.

Qin, L. X. and Kerr, K. F. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res, **32**, 5471-9.

Quackenbush, J. (2002) Microarray data normalization and transformation. Nat Genet, **32 Suppl**, 496-501.

Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Contrino, S., Vilo, J., Abeygunawardena, N., Mukherjee, G., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A. and Sansone, S. A. (2003) ArrayExpress: a public database of gene expression data at EBI. C R Biol, **326**, 1075-8.

Ross, J. S., Schenkein, D. P., Pietrusko, R., Rolfe, M., Linette, G. P., Stec, J., Stagliano, N. E., Ginsburg, G. S., Symmans, W. F., Pusztai, L. and Hortobagyi, G. N. (2004) Targeted therapies for cancer 2004. Am J Clin Pathol, **122**, 598-609.

Rouchka, E. C., Gish, W. and States, D. J. (2002) Comparison of whole genome assemblies of the human genome. Nucleic Acids Res, **30**, 5004-14.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, **270**, 467-70.

Selvanayagam, Z. E., Cheung, T. H., Wei, N., Vittal, R., Lo, K. W., Yeo, W., Kita, T., Ravatn, R., Chung, T. K., Wong, Y. F. and Chin, K. V. (2004) Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling. Cancer Genet Cytogenet, **154**, 63-6.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, **13**, 2498-504.

Shi, L., Tong, W., Goodsaid, F., Frueh, F. W., Fang, H., Han, T., Fuscoe, J. C. and Casciano, D. A. (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. Expert Rev Mol Diagn, **4**, 761-77.

Shih, J. H., Michalowska, A. M., Dobbin, K., Ye, Y., Qiu, T. H. and Green, J. E. (2004) Effects of pooling mRNA in microarray class comparisons. Bioinformatics, **20**, 3318-25.

Simon, R., Radmacher, M. D. and Dobbin, K. (2002) Design of studies using DNA microarrays. Genet Epidemiol, **23**, 21-36.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr. and Brazma, A. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol, **3**, RESEARCH0046.

Sun, Y. V., Boverhof, D. R., Burgoon, L. D., Fielden, M. R. and Zacharewski, T. R. (2004) Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. Nucleic Acids Res, **32**, 4512-23.

Tempelman, R. J. (2005) Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol, **105**, 175-86.

Tong, W., Cao, X., Harris, S., Sun, H., Fang, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., Perkins, R., Shi, L. and Casciano, D. (2003) ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. Environ Health Perspect, **111**, 1819-26.

Townsend, J. P. (2003) Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics, **4**, 41.

Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P. and Cho, K. W. (2002) Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic Acids Res, **30**, e54.

Tsai, C. A., Wang, S. J., Chen, D. T. and Chen, J. J. (2005) Sample size for gene expression microarray experiments. Bioinformatics, **21**, 1502-8.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, **98**, 5116-21.

Ulrich, R. and Friend, S. H. (2002) Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov, **1**, 84-8.

Ulrich, R. G., Rockett, J. C., Gibson, G. G. and Pettit, S. D. (2004) Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. Environ Health Perspect, **112**, 423-7.

Vinciotti, V., Khanin, R., D'Alimonte, D., Liu, X., Cattini, N., Hotchkiss, G., Bucca, G., De Jesus, O., Rasaiyaah, J., Smith, C. P., Kellam, P. and Wit, E. (2004) An experimental evaluation of a loop versus a reference design for two-channel microarrays. Bioinformatics,

Wang, X., Ghosh, S. and Guo, S. W. (2001) Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Res, **29**, E75-5.

Waters, M., Boorman, G., Bushel, P., Cunningham, M., Irwin, R., Merrick, A., Olden, K., Paules, R., Selkirk, J., Stasiewicz, S., Weis, B., Van Houten, B., Walker, N. and Tennant, R. (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics, **111**, 15-28.

Waters, M. D. and Fostel, J. M. (2004) Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet, **5**, 936-48.

Waters, M. D., Olden, K. and Tennant, R. W. (2003b) Toxicogenomic approach for assessing toxicant-related disease. Mutat Res, **544**, 415-24.

Wei, C., Li, J. and Bumgarner, R. E. (2004) Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics, **5**, 87.

Weinshilboum, R. and Wang, L. (2004) Pharmacogenomics: bench to bedside. Nat Rev Drug Discov, **3**, 739-48.

Wetmore, B. A. and Merrick, B. A. (2004) Toxicoproteomics: proteomics applied to toxicology and pathology. Toxicol Pathol, **32**, 619-42.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A. and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res, **32**, D35-40.

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 expression data via mixed models. J Comput Biol, **8**, 625-37.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H. H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol, **3**, research0048.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000) DIP: the database of interacting proteins. Nucleic Acids Res, **28**, 289-91.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J. and Quackenbush, J. (2002a) Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol, **3**, research0062.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, **30**, e15.

Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. Nat Rev Genet, **3**, 579-88.

Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J. and Golub, T. (2001) Molecular classification of multiple tumor types. Bioinformatics, **17 Suppl 1**, S316-22.