UNIVERSITY OF CALIFORNIA, SAN DIEGO

Comparative Functional Genomics of Energy Metabolism and Insulin Resistance in Mammalian Systems

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioengineering with Specialization in Bioinformatics

by

Albert Hsiao

Committee in charge:

Professor Shankar Subramanium, Chair
Professor Jerrold M. Olefsky, Co-chair
Professor Dorothy Sears
Professor Christopher K. Glass
Professor Trey Ideker

2005

The dissertation of Albert Hsiao is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

Chair

_____

Co-chair

University of California, San Diego

2005

Table of Contents

List of Symbols

m	sample mean
s^2	sample variance
μ	'true' mean
σ^2	'true' variance
Λ	matrix of variance parameters
λ	'slope' of the exponential prior
θ	set of prior parameters
ϕ	set of prior hyperparameters
π	posterior probability density
A	expression-dependent variance
B	expression-independent variance
ρ_A	correlation in expression-dependent variance
ρ_B	correlation in expression-independent variance

List of Figures

List of Tables

VITA

October 19, 1978	Born, Escondido, California
2000	B.S., California Institute of Technology
2003	M.S., University of California, San Diego
2005	Ph.D., University of California, San Diego

PUBLICATIONS

ì Variance-modeled posterior inference of microarray data: detecting gene expression changes in 3T3-L1 adipocytesî Bioinformatics, Vol. 20 issue 17, pp.3108-3127, November, 2004.

ì VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression dataî Nucleic Acids Research, 2005, *accepted*.

FIELDS OF STUDY

Major Field: Bioengineering with Specialization in Bioinformatics

Studies in Bioinformatics.
Professors Shankar Subramaniam and Trey Ideker

Studies in Biomechanics and Computational Biology.
Professor Andrew McCulloch

Studies in Biomedical Imaging.
Professors Thomas Liu and Thomas R. Nelson

Studies in Endocrinology and Metabolism.
Professors Jerrold M. Olefsky and Dorothy Sears

Studies in Engineering and Computer Science.
Professor James Arvo, California Institute of Technology

Studies in Biology and Cerebellar Electrophysiology.
Professor James M. Bower, California Institute of Technology

ABSTRACT OF THE DISSERTATION

Comparative Functional Genomics
of Energy Metabolism and Insulin Resistance
in Mammalian Systems

by

Albert Hsiao
Doctor of Philosophy in Bioengineering with Specialization in Bioinformatics

University of California, San Diego, 2005

Professor Shankar Subramaniam, Chair
Professor Jerrold M. Olefsky, Co-chair

Type 2 diabetes mellitus is a highly prevalent human disease often preceded by a period of hyperinsulinemia and insulin resistance. In both pathologic states, tissues involved in energy storage and metabolism demonstrate impaired physiological responses to insulin. Several cell and animal models have been developed to study the action of insulin and its effects in key metabolic tissues. In order to identify commonalities and differences between these model systems and relate these back to human disease, we have undertaken a unique functional genomics approach.

Gene expression microarrays allow the simultaneous measurement of thousands of transcripts. To demonstrate normal physiology in a variety of systems, we apply expression profiling to characterize the effects of insulin in the tissues of C57/BL6 mice, lean Zucker rats, and lean human subjects. We further compare these results with previously published data in 3T3-L1 adipocytes. Comparative analysis across several organisms provides novel insights into the transcriptional responses of three key tissues: skeletal muscle, adipose, and liver. In addition, we apply a similar approach to study models of insulin resistance. In 3T3-L1 adipocytes, we revisit TNF-α-induced resistance and compare this with several whole-organism models. In order to interpret the volumes of data collected from these experiments, however, we need a reliable and efficient tool to effectively manage this data.

Interpretation of high-throughput expression data is a bioinformatics problem requiring careful integration of statistics, mathematical modeling, computational infrastructure, and biology. As gene expression and promoter microarrays become increasingly prevalent, there is a greater need for precise and accessible analytical tools. These tools not only need to encapsulate mathematically robust methods, but also be simple enough so that statistical details can be sufficiently separated from biological interpretation. Utilizing a Bayesian framework, we have developed such an approach, dubbed VAMPIRE, and implemented it as a web-based, database-driven application. The tools presented here represent a unique solution to the interpretation of gene expression data. These advances have subsequently allowed us to piece together a comprehensive picture of the physiological responses to prolonged hyperinsulinemia, and to suggest novel targets for further study of type 2 diabetes.

CHAPTER I

Introduction

A. Type 2 Diabetes Mellitus and Insulin Resistance

Type 2 diabetes mellitus is a polygeneic, multifactorial disease of increasing prevalence in the United States. Although specific genetic defects that are strongly predictive for this disease have not yet been found, it has a considerable genetic component. Studies of monozygotic twins have shown concordance rates of 70-90%. In comparison, similar studies of type 1 diabetes have shown rates of only 35-50%[1]. A number of individual genes have been identified to contribute to this disease, but no single culprit has been found[2]. Environmental factors also play a significant role. Not only is obesity is present in approximately 80% of diabetic patients, but exercise and diet can greatly affect insulin-resistance underlying the disease. The pathogenesis of this heterogeneous disease thus involves a complex interplay between genetic susceptibility and environment.

Insulin resistance is generally believed to be the primary lesion in type 2 diabetes. Insulin normally acts on three major tissues to control blood glucose. In skeletal muscle, insulin promotes glucose uptake, glycogen production, and protein biosynthesis. In adipose tissue, enhanced uptake of glucose helps to power the synthesis of fatty acids and promotes the esterification of fatty acids in the form of triglycerides. In liver, insulin suppresses glucose production and instead promotes energy storage in the form of glycogen. In susceptible individuals however, insulin sensitivity in target tissues gradually wanes. Because of the loss of sensitivity, pancreatic beta cells are forced to produce increasing quantities of insulin. As target tissues become increasingly unresponsive, this compensatory mechanism eventually fails to suppress blood glucose. It is not until blood glucose reaches a diagnostic threshold, with or without physical symptoms of diabetes, that we are able to identify patients as diabetic.

Diagnostic criteria for diabetes have evolved substantially from its description by Galen and Areteus in ancient Greece as a disease of excess thirst and urine production[3]. In 1769, Cullen noted that contrast to diabetes insipidus, the urine of patients with diabetes mellitus was sweet, ìwith the smell, color, and flavor of honeyî. Since then, as clinical medicine has evolved, the etiology and pathogenesis of this heterogeneous disease have gradually been used to categorize patients into sub-groups. Type 1 diabetics demonstrate a primary lesion in insulin secretion, either through autoimmune or idiopathic causes. While type 2 diabetics may also show decreases in insulin secretion, they show substantial insulin resistance not observed in type 1. In addition, modern diagnostic criteria are used to identify an earlier state of ìpre-diabetesî or impaired glucose tolerance. While diabetics are diagnosed by observations of fasting plasma glucose levels (FPG) of 126 mg/dL or greater, patients with impaired glucose tolerance exhibit FPG below this threshold but exceeding 110 mg/dL. Normal subjects typically have fasting glucose levels at or below 100 mg/dL. With improved diagnostic criteria, physicians have been able to intervene at earlier stages in the progression of the disease.

A diverse set of treatment options are currently available for type 2 diabetes. While exercise and diet are ideal as a first-line treatment against insulin-resistance and diabetes, compliance for these treatments is often very weak. Furthermore, this form of treatment is often insufficient for proper glycemic control. Because of these problems, a variety of medical options are commonly used. α-glucosidase inhibitors such as acarbose and miglitol inhibit gastrointestinal glucose absorption. Sulfonylureas act as insulin secretagogues, promoting beta cell insulin secretion. Metformin suppresses hepatic glucose production, promotes glucose utilization, and may also improve insulin sensitivity. Finally, thiazolidinediones like rosiglitazone and pioglitazone appear to target insulin sensitivity specifically. Together, these drugs are fairly effective for a large fraction of the diabetic population. There exists however, a substantial proportion of

patients for whom these drugs are contraindicated, who suffer from various drug side effects, or who remain in poor glycemic control despite therapy. In order to develop novel therapies for type 2 diabetes then, it is of fundamental importance to better understand the pathogenesis of insulin resistance.

A number of animal models have been developed to study insulin resistance. In 1995, a novel secreted factor controlling satiety was discovered. This hormone, leptin, is secreted by adipocytes and sensed by hypothalamic neurons. One commonly used model to study insulin resistance then, is the Zucker fa/fa rat, deficient in the leptin receptor. These animals demonstrate elevations in fasting insulin, fasting free fatty acids, hepatic glucose production, and have impaired insulin-stimulated glucose uptake. In many ways, they share a similar to physiologic profile to insulin-resistant human patients. Alternatively, many animals develop insulin resistance simply when subjected to high fat diets. High fat fed rats and mice also show elevations in experimental measures of insulin resistance. These, and related models, have already provided much insight into the impact of diet and obesity on insulin sensitivity, as well as interactions between the target tissues. Even so, much of the cellular physiology of these model systems remains unexplored. In order to address this aspect of insulin resistance, we applied a unique functional genomics approach to tackle two fundamental issues: (1) what are the physiologic transcriptional targets of insulin in each target tissue, and (2) how might these transcriptional effects be altered in the insulin-resistant state? By answering these questions, we can better characterize the mechanisms of human insulin resistance, and suggest new strategies for further study and therapeutic intervention.

B. Analysis of High-Throughput Expression Data

Addressing the global transcriptional impact of insulin in each of the target tissues and model systems requires a modern bioinformatics approach. A number of microarray platforms have emerged over the past fifteen years to begin to probe the global responses

of cells and tissues to external stimuli. In the early 1990s, Fodor *et al* described a photolithographic technique that could be used to synthesize oligonucleotides on the surface of a glass slide to simultaneously probe thousands of nucleotide sequences in a given sample[4]. This manufacturing method has since been commercialized by Affymetrix (Santa Clara, CA). Shortly thereafter, the laboratory of Patrick O. Brown at Stanford described a protocol for the creation of robotic microarray spotters and cDNA microarrays, making this technology more accessible to academic labs. Numerous other commercial and academic entities have since developed improvements to these designs, providing increasing reliability to the manufacturing of individual microarrays. In general, a labeled nucleotide sample is washed over the surface of the array, hybridizing to complementary probes. Excess sample is washed away and the remaining sample is read as a fluorescent signal on a confocal microscope. Methods for labeling the input RNA and imaging the hybridized array have also advanced. As improvements in microarray platforms have progressed at rapid pace, so have analytical algorithms for interpreting the resulting data. Current methods for analysis of microarray data can be divided into several components: (1) image processing and summary measures of gene expression, (2) detection of differential-expression, (3) clustering methods, and (4) gene annotation.

The first stage of microarray analysis, image processing, involves the analysis of confocal microscopic images to obtain intensity measurements for each of the thousands of probes spotted or synthesized on the array surface. A variety of commercial software tools currently provide adequate treatment of these image-processing needs (Gene Chip Operating Software, Affymetrix, Santa Clara CA; Genepix, Union City, CA). These tools delineate the locations of each probe, and quantify the amount of signal present at each position to obtain a matrix of intensity measurements. If more than one probe is used to investigate a particular transcript, the intensities for these probes can be combined to provide more robust summary measures of gene expression. Li and Wong (2001)

published one of the first statistical methods for combining multiple probes into a single measurement[5]. Since then, several commercial and academic groups have devised comparable statistical measures[6, 7] (Affymetrix, Santa Clara CA). Once a summary measure of gene expression can obtained by one of these techniques, it becomes possible to test for differential-expression between two sets of RNA samples.

One of the most fundamental questions asked in gene expression analysis is whether a given gene is differentially-expressed by a particular stimulus. For example, an experimenter may want to determine what genes are induced by treatment with the growth factor, Wnt. Several RNA samples may be collected from cells that have either been treated with Wnt or with vehicle alone. The biological question can then be addressed by performing a statistical test to identify those genes that show the most change relative to measurement noise. Unfortunately, the massively parallel nature of microarray studies limits the effectiveness of traditional statistical tests in addressing this question. As there are typically on the order of 10^5 genes being probed by the array, multiple-testing correction must be used to limit the number of false positives detected by the test. For example, when applying a conventional *t*-test, scientists often choose type I error (α) of 0.05 as a significance threshold. Then, if only a single gene is tested, one can expect only a 1 in 20 chance that a ìsignificantî change in gene expression occurred by random chance alone. However, when 10^5 genes are simultaneously tested at this threshold, 500 false-positives are expected. There are two multiple-testing corrections commonly used for microarray analysis: Bonferroni-correction and empirical false-discovery rates. Application of these thresholds with conventional statistical tests frequently yields very few, and often no statistically significant changes, given the numbers of experimental replicates commonly performed. It is partially because of this result that statistical methods have not been rapidly adopted by the biological research community.

Instead of statistical methods, the vast majority of microarray studies have adopted a fold-change approach, or some variant thereof. It has been widely recognized in this kind of data that the signal-to-noise ratio is far worse at lower intensity measurements than at higher intensities. Because of this, microarray users often define an arbitrary threshold, statistical or otherwise, to remove low-reliability data prior to analysis. Then, an observed change of 2-fold or greater is deemed as ìsignificantî. While not a statistically robust approach, the fold-change method can give fairly reliable results, detecting many genes known to be regulated by their corresponding stimuli. Still, this approach is somewhat limited because of the need to specify arbitrary thresholds that are presumed to indicate significance. Furthermore, these thresholds are rarely, if ever, determined *a priori* and correlated with statistical concepts of significance. These thresholds therefore subject to the experimentersí bias to identify differentially-expressed genes.

A third level of analysis involves the identification of patterns in gene expression across individual samples or treatment conditions. To tackle this problem, several methods have been adapted from data mining and engineering. In clustering, each set of microarray measurements can be interpreted as a point in *n*-dimensional space. The goal of these methods is to subdivide the individual samples into several groups, minimizing the distance between members within a group while maximizing the distance between groups. This method can either be used to combine RNA samples that are most similar to each other or to identify genes that share the same patterns of expression. By finding genes that share patterns of expression across multiple treatment conditions, it is hypothesized that one can identify shared mechanisms of co-regulation. In addition to clustering, researchers have also devised variations of principle components analysis (PCA) to reduce the dimensionality of microarray data. The basic premise is that there exist abstract, functional groups of genes, which can be identified by applying singular value decomposition. This tool of linear algebra has been widely applied in a variety of

engineering disciplines to assist in filtering out ìnoiseî and to identify interactions. Combinations of PCA and clustering methods are now routinely used to either identify previously unknown relationships between genes or relationships between individual RNA samples.

The fourth and final level of analysis described here relates to integrating biological knowledge with observed gene expression data. While gene annotation databases are currently quite sparse, these databases may ultimately be powerful tools for reducing the dimensionality of expression data. Instead of relying on mathematical constructions alone, biological databases can be used to group genes that are known to have related functions, that belong to the same cellular pathways, or that share similar regulatory elements. Examples of these databases include Gene Ontology (GO)[8], Kyoto Encyclopedia of Genes and Genomes (KEGG)[9], and TRANSFAC[10]. There are two general approaches currently used in integrating expression data with systematic biological knowledge. Non-statistical methods simply provide an interface for accessing annotation information for a selected set of microarray features or genes, while statistical methods look for statistical enrichment of annotation terms by comparing the selected set with a background set. By identifying enrichment, statistical methods can identify specific pathways that are most relevant to the microarray experiment. A variety of such tools have recently been developed to address this problem[11-14], but few adequately integrate microarray expression and make this kind of analysis truly accessible.

C. Our approach

In order to further our understanding of insulin resistance and type 2 diabetes, we have undertaken a comparative functional genomics approach. Individual model systems for studying insulinís transcriptional action each have fundamental limitations. For example, mouse and rat models of insulin resistance are often limited by the sheer size of the tissues needed. In order to obtain sufficient RNA content from a particular tissue type,

nearly the entire tissue must be used. Because of this, there can be considerable differences in the cellular content of the sampled tissue, and further study cannot easily be performed on the same animal. In contrast, in the human model, a small biopsy of relatively inconsequential size is obtained from a living patient. Further studies can be performed on this individual at a later time. As animal to animal variability can be fairly substantial, analysis of different animals can negatively impact the detection of meaningful changes in gene expression. Human models of insulin-resistance are also limited. Skeletal muscle is not homogeneous, so biopsied tissue may not necessarily reflect the behavior of the entire muscle. Patient populations have considerable variations in age, genetic background, concurrent disease, and environment. These variations may confound the detection of differences between patient groups. Furthermore, many experimental procedures which are ethically tolerated in mice or rats are not equivalently tolerated in human subjects. Thus, under the hypothesis that general mechanisms of insulin action will be conserved across model systems, it is possible to average out these model-specific issues, and gain insight where each model alone is deficient. To further our understanding of the actions of insulin and the role of transcriptional networks in mediating its action, we examine three model systems. In mice, we explore the effect of hyperinsulinemic-euglycemic clamp at 90 and 240 minutes. This model, mimicking the transition from fasting to fed state, explores the effect of sustained increases of insulin in liver. In lean littermates of Zucker fa/fa rats, we examine the effect of the clamp on the transcriptional state of skeletal muscle and liver. In humans, a clamp of non-diabetic subjects identifies novel transcriptional targets in skeletal muscle.

For each of the model systems and tissues, we must not only be able to easily identify the individual genes undergoing significant transcriptional change, but we must also interpret these transcriptional events in a global context. In other words, our approach must be able to summarize changes in expression of hundreds of genes into a few

physiologically meaningful ideas and hypotheses that can be subsequently tested. Since thousands of cross-comparisons are performed across organisms, subject groups, tissues, and treatment conditions, the data accounting and analysis pipeline must also be equivalently simplified.

In order to accomplish the previously stated goals, then, we have undertaken the following steps:

1. Derive a novel, robust foundation for the interpretation of microarray data which can incorporate statistical concepts of significance with minimal user-intervention.
2. Derive and assess novel statistical tests for the analysis of microarray data. These tests must be able to adequately detect:
    a. Differences in expression in unpaired data
    b. Differences in expression in paired data
    c. Differences in expression changes between unpaired measurements
    d. Differences in expression changes between paired measurements
3. Design and implement a relational database-driven application for the analysis of expression data, which integrates these novel methods.
4. Design and implement a web-based application to integrate biological knowledge databases with expression analyses.
5. Apply the implemented statistical tools to identify the similarities and differences between models of insulin-stimulated gene expression.

D. Organization of the Text

We provide in this introductory chapter a brief description of type 2 diabetes mellitus and the experimental models used to understand this disease. We describe in further detail the clinical manifestations of diabetes, modern treatment options, and current knowledge of the underlying pathophysiology. We elaborate on several model systems

used to study insulin resistance, and explain the need to explore the transcriptional responses of key metabolic tissues. The manufacturing process used for the creation of gene expression microarrays is briefly overviewed. We end the chapter by describing the methods currently used to interpret microarray data, and address their limitations.

In Chapter 2, we describe the mathematical foundations of a novel statistical approach, VAMPIRE, which can be used to interpret high-throughput functional genomics data. This approach applies deterministic and stochastic modeling techniques to characterize the error structure of each array data set. We also derive a Bayesian statistical test, which is used to identify differentially-expressed genes between two groups of expression measurements. We subsequently compare the results of our method to conventional statistical approaches and show enhanced sensitivity in low-replicate data, with nominal losses in specificity. In addition, we use these studies to begin to elucidate the transcriptional effects of thiazolidinediones in 3T3-L1 murine adipocytes.

Chapter 3 further explores the implications of the VAMPIRE statistical framework. In this work, we extend both the modeling procedure and the Bayesian test to further increase statistical power in *paired* microarray data. Our novel paired method is dubbed Bivariate Microarray Analysis (BMA). We quantitatively assess the sensitivity and specificity of several statistical tests in a simulated data set. By doing so, we demonstrate the accuracy of *a priori* significance thresholds in assessing type I error, and show that increased sensitivity in BMA does not necessarily come at the cost of specificity. We further demonstrate the effectiveness of our method in two previously published data sets of reasonable importance to diabetes research: (1) PDGF-treated T98G cells, which serves as a model of growth factor-regulated gene expression, and (2) a time course of endotoxin-treated RAW264.7 macrophages to illustrate critical genes indicative of the innate immune response. Together, these data sets show that reliable statistical interpretation of gene expression data does not always require large sample numbers, and

suggest that it is possible to incrementally determine the number of necessary samples during a microarray experiment.

In Chapter 4, we introduce the VAMPIRE microarray analysis suite, a web-based platform for interpreting functional genomics data. We describe the design and implementation of this tool, and elaborate on its ability to handle incremental additions to microarray data sets. We also introduce a novel utility for functional interpretation of gene expression data, GOby. Integrated into the VAMPIRE suite, GOby takes advantage of existing annotation databases to identify biological functions that are enriched among differentially-regulated genes. Together, VAMPIRE and GOby represent one of the first efforts toward reliable integration of gene expression data and systematic biological knowledge.

Chapter 5 explores insulin-regulated gene expression across a variety of tissues and organisms. As many model organisms are used to better understand the mechanisms of action and physiological roles of insulin, we considered it essential to explore their transcriptional similarities between major target tissues of rats, mice, and humans. We suggest that insulin may mediate its role in regulating skeletal muscle energy metabolism by co-opting the differentiation program, and through this mechanism, incidentally suppress cell proliferation. In liver, we identify transcriptional regulators of gluconeogenesis and inflammation as the dominant conserved targets of insulin-regulated gene expression.

In Chapter 6, we review strengths and weaknesses of our statistical approach, and discuss directions of future research in microarray analysis. We also present here a summary of the findings obtained from our comparative functional genomics analysis. We suggest several areas for future research, for both interpreting gene expression data, and for understanding the basic causes of insulin resistance.

May 19, 2005

Albert Hsiao has my permission to include the following paper, of which I was a co-author, in his doctoral dissertation.

Hsiao, Albert; Worrall, Dorothy S.; Olefsky, Jerrold M.; Subramaniam, Shankar. ì Variance-modeled posterior inference of microarray data: Detecting gene-expression changes in 3T3-L1 adipocytesî , Bioinformatics, vol. 20 issue 17, 2004.

_____

Dorothy Sears (Worrall)

_____

Jerrold M. Olefsky

_____

Shankar Subramaniam

This chapter, in full, is a reprint of the material as it appears in Bioinformatics 2004, Hsiao, Albert; Worrall, Dorothy S.; Olefsky, Jerrold M.; Subramaniam, Shankar., Oxford University Press, 2005. The dissertation author was the primary investigator and single author of this paper.

CHAPTER II

Variance-Modeled Posterior Inference of Microarray Data

A. Abstract

Microarrays are becoming an increasingly common tool for observing changes in gene expression over a large cross section of the genome. This experimental tool is particularly valuable for understanding the genome-wide changes in gene transcription in response to thiazolidinedione (TZD) treatment. The TZD class of drugs is known to improve insulin-sensitivity in diabetic patients, and is clinically used in treatment regimens. In cells, TZDs bind to and activate the transcriptional activity of peroxisome proliferator-activated receptor gamma (PPAR-γ). Large-scale array analyses will provide some insight into the mechanisms of TZD-mediated insulin-sensitization. Unfortunately, a theoretical basis for analyzing array data has not kept pace with the rapid adoption of this tool. The methods that are commonly used, particularly the fold-change approach and the standard *t*-test, either lack statistical rigor or resort to generalized statistical models that do not accurately estimate variability at low replicate numbers.

We introduce a statistical framework that models the dependence of measurement variance on the level of gene expression in the context of a Bayesian hierarchical model. We compare several methods of parameter estimation and subsequently apply these to determine a set of genes in 3T3-L1 adipocytes that are differentially-regulated in response to thiazolidinedione treatment. When the number of experimental replicates is low (n=2-3), this approach appears to qualitatively preserve an equivalent degree of specificity, while vastly improving sensitivity over other comparable methods. In addition, the statistical framework developed here can be readily applied to understand the implicit assumptions made in traditional fold-change approaches to array analysis. Our statistical approach is implemented as a set of Java-based tools called VAMPIRE, accessible from our website at http://genome.ucsd.edu/VAMPIRE.

B. Introduction

Thiazolidinediones (TZDs) are a class of insulin-sensitizing agents commonly used to treat patients with Type II Diabetes Mellitus. This class of drugs acts by binding the PPAR-γ nuclear receptor[15], thereby modifying transcription of key glucoregulatory and lipid regulatory genes[16, 17]. Several individual downstream targets of PPAR-γ have recently been identified and confirmed by RT-PCR, but the mechanism of insulin-sensitization is still poorly understood. A fundamental understanding of the gene transcription events leading to the anti-diabetic action of TZDs would greatly further our understanding of insulin action, insulin resistance, and enable new, more focused efforts at drug discovery. To tackle this problem, we treated differentiated 3T3-L1 adipocytes with three different TZDs (Pioglitazone, Rosiglitazone, Troglitazone) and followed up with microarray analyses of RNA.

Microarrays have revolutionized modern cell biology by opening a broader window to the influence of physiological perturbations on gene regulation. Although a variety of statistical methods are available and used to analyze microarray data, the field of microarray statistics is still in its infancy, and is being actively developed. At the present time, array statistics can be subdivided into three principal levels: (1) providing summary statistics that indicate the degree of gene expression of each array feature from microscope imaging data, (2) deducing the measurement uncertainty between summary statistics of multiple experimental replicates to determine which genes are differentially expressed, and (3) determining patterns in gene expression in several tissues, or under various perturbations. For the first class of microarray statistics, the methods for summarizing gene expression are quite well-developed, and have been much improved by the work of Wong and coworkers[18]. The latter two classes of array statistics unfortunately still lag behind, and are a fundamental limitation to precise, reproducible interpretations of array data.

Measurement variance is an indication of the reproducibility of several measurements given the same experimental conditions. A number of potential sources of variability exist, which lie beyond the control of the experimenter, and many of these can influence individual genes differently. It has been widely observed that the degree of measurement uncertainty is dependent on the level of gene expression. In many cases, the variance decreases as a fraction of the total signal as the expression level increases. Among the sources of error that are independent of gene expression are the random noise introduced by the equipment used to gather image data and the variability of quality and fluorescence. In addition, the strength of hybridization and the absolute amount of RNA harvested from the sample are examples of factors that influence variance in an expression-level dependent fashion. Biological variability, the changes in gene expression that are variable between controlled measurements, may contribute to both classes of variance. All of these factors affect the level of reproducibility. In order to determine whether a particular gene measured on an array is differentially expressed between two physiological conditions then, it is necessary to account for uncontrollable sources of variation.

A fold-change cutoff remains the *de facto* standard for this level of array analysis, but several recent articles assert the need for more updated analytical methods[19]. Fold-change approaches qualitatively give a fair level of reproducibility, but are plagued by the lack of a solid statistical foundation, and are unable to handle expression-dependent variability. To perform this kind of analysis, practitioners compute the ratio of gene expression between two treatment conditions, where one is a ìtreatedî and one is a ìcontrolî experimental condition. When the ratio exceeds an arbitrary threshold, often a 2-fold increase or a 0.5-fold decrease, the gene associated with the array feature is considered to be ìdifferentially-expressedî. This method has been improved by the introduction of a minimum-expression cutoff to account for the increased fractional

variability at low-expression levels. Despite these improvements, this approach is still rather arbitrary because the best cutoffs are difficult to determine *a priori*, and it is difficult to correlate these cutoffs with traditional concepts of statistical significance. Several new methods have been developed to address these limitations.

Among the most popular advanced methods for this second tier of analysis currently include Significance Analysis of Microarrays (SAM)[20] and Cyber-T[21]. SAM utilizes a non-parametric bootstrap technique to estimate the degree of variability for each feature of the array. Cyber-T, in contrast, relies on a Bayesian modification of a traditional *t*-test. In order to better assess the uncertainty, Cyber-T examines the variance of features in a gene expression window near the sample mean of the feature of interest, and uses these as ìpseudo-replicatesî. Both methods are substantial improvements over fold-change approaches as they can ascertain the significance level of observed changes in gene expression.

It is our opinion that this tier of analysis can be further improved by more precisely modeling the relationship between variance and gene expression level. By incorporating this model into statistical analysis, the accuracy of calls for ìdifferentially-expressedî genes should be substantially increased. Noting that the number of replicates of microarray experiments tends to be small, often numbering between two and four per treatment condition, it is likely that the variance estimates derived from a global model will likely be more accurate than estimates obtained from individual features alone. We therefore choose to apply a global variance model related to that proposed by Rocke and Durbin (2001) to approximate the relationship[22]. Given the parameterized relationship between variance and mean expression, we apply a Bayesian hierarchical model to compute a likely prior density for the ìtrue meanî expression level of each feature on the array, and compute the corresponding posterior density after observation of each replicate. The net effect is the conversion of a few point measurements into an easily

interpretable likelihood density for the ìtrue meanî expression of each feature. This can be used to answer a wide assortment of analytical questions, including whether a gene is differentially expressed between two treatment conditions.

C. Methods and Statistical Framework

Only a handful of assumptions are needed to develop a fairly rigorous statistical method for the analysis of microarray data. It is assumed here that there is a ìtrueî expression level for each feature in the array, designated μ_i. Each measurement of a particular feature is considered to be a sampling from a distribution centered about μ_i, and whose variance is σ_i^2. We also assume that the likelihood of any measurement from this distribution is approximately normal. So far, these assumptions are typical of many statistical tests, including the *t*-test. We also assume that the variance can be expressed as a deterministic function of μ_i. In other words, there is no uncertainty in the amount of variance for a feature, if both (1) the featureís ìtrueî expression level is known, and (2) the relationship between and ìtrueî expression and variance is known. A last assumption used in this method is that in the absence of prior observations, the prior knowledge about the ìtrueî expression, μ_i, is given by the local distribution of sample means. If additional prior knowledge is available, this data can alternatively be used to construct prior densities for μ_i.

The above assumptions are encapsulated by a Bayesian statistical model (Fig. 2.1). By applying Bayesí rule to a given prior density, we can narrow the posterior density for each μ_i after observing several replicates of microarray data. The model that we choose belongs to the exponential conjugate family devised by Morris (1983)[23].

The likelihood function for the univariate exponential family can be expressed as

$$f(x \mid \theta) = a(x)\exp[x\phi(\theta) - \psi(\theta)] \qquad (1)$$

where θ is the unknown parameter and x is a single measurement. For the normal density with known variance,

$$a(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-x^2/2\sigma^2\right] \tag{2a}$$

$$\phi(\theta) = \theta/\sigma^2 \tag{2b}$$

$$\psi(\theta) = \theta^2/2\sigma^2 . \tag{2c}$$

where the variance of the normal density is represented as σ^2. Applied directly to the microarray problem, the unknown parameter θ is our ì true meanî , μ_i.

Given a conjugate prior of the form,

$$P(\theta \mid \alpha, \beta) \propto \exp[\alpha\phi(\theta) - \beta\psi(\theta)] . \tag{3}$$

the conjugate posterior is

$$P(\theta \mid X, \alpha, \beta) \propto \exp[\alpha_m \phi(\theta) - \beta_m \psi(\theta)] \tag{4}$$

where α, β are arbitrary prior parameters and the corresponding conjugate posterior parameters are given by

$$\alpha_m = m\bar{x} + \alpha , \tag{5}$$

$$\beta_m = m + \beta , \tag{6}$$

and m is the number of array replicates.

We can re-express the posterior in a more familiar form. Rearranging terms and completing the square, the conjugate posterior becomes:

$$P(\theta \mid X, \alpha, \beta) \propto \exp\left[-\frac{\beta_m}{2\sigma^2}\left(\theta - \frac{\alpha_m}{\beta_m}\right)^2\right] . \tag{7}$$

Therefore, the conjugate posterior is a normal density with

$$E[\theta] = \frac{\alpha_m}{\beta_m} \tag{8}$$

and

$$Var[\theta] = \frac{\sigma^2}{\beta_m}. \quad (9)$$

Conjugate families are useful in general because of their ability to simplify computations, but this choice has several valuable properties. In this model, the posterior density for the unknown parameter θ is normal. Normal random variables are easily summed, and therefore testing the equality of μ_i between two experimental conditions becomes a simple z-test. This test can be rapidly performed by evaluating an error function. Furthermore, since the ratio of two normal random variables is a Cauchy random variable, posterior densities for fold changes between experimental conditions can be easily assigned. This may be useful in determining the posterior density for a fold-change statistic.

Variance model

Measurements of microarray data are modeled with a two-component global variance model as previously described by Rocke and Durbin[22]. In their model, the variance is decomposed into expression-dependent variance, σ_η^2, and expression-independent variance, σ_ε^2, which are assumed to be constant parameters for a given microarray experiment. A similar model was also used by Ideker *et al* in determining the significance of gene expression changes in two-color microarrays[24]. It can easily be shown that the relationship between variance and true expression level can be written as

$$Var[X] = \sigma^2 = A\mu^2 + B \quad (10)$$

and when σ is small compared to μ, the log-transformed variance is related to the non-transformed variance by

$$Var[\log(X)] \approx \sigma^2 \mu^{-2} \quad (11)$$

and thus,

$$Var[\log(X)] \approx A + B\mu^{-2} \qquad (12)$$

where A is the expression-dependent variance, and B is expression-independent variance for the random variable X (Appendix). Furthermore, if X is normal, then in the limit that σ is very small compared to μ, log(X) will also be approximately normal.

Given these expressions for variance, there are several methods that we can use to estimate the components of variance directly from microarray data in either non-transformed or log-transformed coordinates. One approach is to use least-squares regression to approximate the curve that best-fits the observed mean squared deviations (MSD) at each expression level. The MSD are the maximum likelihood estimator for the variance of each array feature. Ideally, we would square the differences between the measurements and μ_i, but the MSD uses the sample mean, m_i, as a surrogate. We have observed empirically that this method gives better results when performed in log-transformed rather than in non-transformed coordinates. A second approach is to use a Markov Chain Monte Carlo (MCMC) method to compute estimates for the unknown parameters, A and B. We simulate the likelihood function for all of the measurements X, and use this to compute the mean and variance of the posterior densities. The variance gives us an indication of convergence of the MCMC simulation. The likelihood function for each measurement is a normal likelihood given by

$$X_{ij} \sim N\left(x_{ij} \mid \mu_i, A\mu_i^{\,2} + B\right) \propto \exp\left[-\frac{1}{2}\frac{\left(x_{ij}-\mu_i\right)^2}{A\mu_i^{\,2}+B}\right] \qquad (13)$$

where i indexes each array feature and j indexes each replicate measurement. As in the case of least-squares regression, μ_i is again unknown, so we use the sample mean, m_i, as a surrogate. While it would be ideal to integrate out μ_i, analytical integration of this equation is impractical.

According to the variance model, the sample mean, m_i, becomes an increasingly inaccurate estimator of μ_i at low expression levels, where fractional variance is largest.

This causes the MSD to dramatically underestimate the true variability. Qualitatively, we can understand this by noticing that μ_i does not necessarily lie at the center of observed measurements, but m_i always does. Since this can impair the accuracy of both approaches for estimating A and B, it becomes necessary to draw an expression-level cutoff for the purposes of estimating the unknown parameters. Only measurements whose sample means reside above the cutoff are used. The use of a cutoff is substantial departure from previous attempts to estimate the parameters of the two-component variance model. The issue of where a reasonable cutoff should reside will be addressed later, but once parameter estimates are available, they apply to all of the features on the array, regardless of where the cutoff is ultimately drawn.

Prior density for μ_i

The prior density describes the relative likelihood that ìtrueî gene expression, μ_i, is any particular value, before any measurements are observed. It summarizes our prior knowledge about μ_i. Since we are using the exponential conjugate model, any prior of the form

$$P(\theta \mid \alpha, \beta) \propto \exp[\alpha\phi(\theta) - \beta\psi(\theta)] \qquad (14)$$

could theoretically be used. If previous measurements are available, then the corresponding posterior densities can be parameterized by the above equation and used as a prior. In most cases though, no obvious prior exists. One reasonable choice then, is to compute a prior for μ_i by simulating the likelihood density for prior parameters with MCMC. To do this, we consider each gene expression measurement as a sampling from the hierarchical model whose prior is parameterized by an exponential distribution. We can then find an estimate of the unknown hyperparameter after observing all of the measurements. It is by systematically estimating the hyperparameter, λ, rather than

arbitrarily choosing a value, that we make this model truly hierarchical. The exponential prior can be expressed as

$$P(\theta \mid \lambda) \propto \exp[-\lambda\theta] \tag{15}$$

An added complication arises because the likelihood function of our Bayesian model assumes that variance is constant, which is not globally true given our variance model. Locally however, within a given ì expression regionî, variances are relatively constant. It is therefore possible to compute the parameters of an exponential prior in the vicinity of each expression level. The likelihood of observing a set of data is described by

$$f(X_i \mid \lambda) = \int_a^b f(X_i \mid \mu_i, \lambda) f(\mu_i \mid \lambda) d\mu_i \tag{16}$$

where the data likelihood function is

$$f(X_i \mid \mu_i, \lambda) = \left[2\pi\sigma^2\right]^{-r/2} \exp\left[-\frac{\sum_{j=1}^{r}(X_{ij} - \mu_i)^2}{2\sigma^2}\right] \tag{17}$$

and the prior density for μ_i is

$$f(\mu_i \mid \lambda) = \frac{\lambda e^{-\lambda\mu_i}}{e^{-\lambda a} - e^{-\lambda b}}. \tag{18}$$

In these equations, X_{ij} represents the measurements of features found in the expression region, i indexes each feature, j indexes each of the r replicates, [a, b] gives the boundaries of the local vicinity, and σ^2 is the corresponding variance. The method we devised to perform segmentation of the array features may be found in the appendix. Integrating out μ_i, we find

$$f(X_i \mid \lambda) = \frac{\lambda\left[2\pi\sigma^2\right]^{-(r-1)/2}}{\sqrt{2r}} \frac{g(b) - g(a)}{e^{-\lambda a} - e^{-\lambda b}} \exp\left[\frac{\lambda^2\sigma^2}{2r} - \frac{(r-1)s_i^2}{2\sigma^2} - \lambda\bar{x}_i\right] \tag{19}$$

where

$$g(y) = erf\left[\frac{ry + \lambda\sigma^2 - r\bar{x}_i}{\sqrt{2r\sigma^2}}\right] \quad (20)$$

r is the number of array replicates, *a* and *b* are the segment boundaries, and $\bar{x}_i$ and s_i^2 are the mean and sample variance of the *i*ith feature, respectively.

Posterior density for μ_i

The posterior density describes our knowledge about the likely values for the true gene expression, μ_i, after observing data. In our Bayesian model, this density is normal, and the parameters of the posterior are readily interpreted. They are

$$E[\mu_i] = \overline{x_i} - \frac{\lambda_i \sigma_i^2}{m} \quad (21)$$

$$Var[\mu_i] = \frac{\sigma_i^2}{m} \quad (22)$$

The mean of the posterior is just the sample mean modified by a ìposterior shiftî, and the variance of the posterior is simply the experimental variance scaled down by the number of replicates, *m*. These parameters can be understood qualitatively. As the slope of the prior, $-\lambda_i$, grows more negative, the posterior density is shifted further to the left. With a greater λ_i, there are many more features on the array whose ìtrue meanî expression lies to the left, that it is more likely that μ_i for the current feature also lies to the left. This change is mediated by the magnitude of uncertainty in our measurement of the sample mean, σ^2/m. As the number of replicates grows, the uncertainty in ìtrueî gene expression decreases, and the magnitude of the posterior shift also decreases.

Statistical tests

One of the most common tasks in the analysis of microarray data is to determine what features are differentially expressed between two experimental conditions. Recognizing that the posterior densities for μ_i in each treatment are normal, the summary statistic $Z_i = \mu_i^e - \mu_i^c$ is the normal density $Z_i \sim N(\mu_z, \sigma_z^2)$ with parameters

$$\mu_z = E[\mu_i^e] - E[\mu_i^c] \quad (23a)$$

$$\sigma_z^2 = Var\left[\mu_i^e\right] + Var\left[\mu_i^c\right] \tag{23b}$$

The superscripts *e* and *c* designate the posteriors of the experimental and control treatments, respectively. Thus, we can simply test the null hypothesis

$$H_0 : \left|\mu_i^e - \mu_i^c\right| = 0 \text{ against } H_1 : \left|\mu_i^e - \mu_i^c\right| \neq 0 \tag{24}$$

to determine whether the *i*íth feature is differentially expressed.

When the difference between the posterior densities of μ_i for two treatments is sufficiently large, then the change in expression is considered to be ì significantî. This is qualitatively similar to the *t*-test, as well as traditional fold-change approaches to array analysis. Unlike a *t*-test though, where the observed means are compared to the observed sample variance, we instead compute the likelihood that the statistic *Z* overlaps with 0. The corresponding z-score depends on the uncertainty in our knowledge of μ_i, and not on any particular observation of sample variance. In order to provide a meaningful indication of statistical significance that can be readily understood by the biological sciences community, we define a *p*-value as

$$p_i = \begin{cases} \int_{-\infty}^{0} f(z_i)dz, & E[Z] \geq 0 \\ \int_{0}^{\infty} f(z_i)dz, & E[Z] < 0 \end{cases} \tag{25}$$

where $f(z_i)$ is the density function for Z_i. It is worthwhile to note that the *p*-value of a *t*-test is defined in a similar manner. In addition, we define statistically significant differential-expression of the *i*'th feature as fulfillment of the inequality

$$p_i < \alpha \tag{26}$$

where α is the Bonferroni-corrected significance threshold, as described in Baldi and Long (2001). Since the number of features on each array tends to be large, it is necessary to correct for the number of statistical tests. To achieve an experiment-wide false-positive rate of α_{Bonf},

$$\alpha = 1 - \exp\left[\frac{\ln\left(1-\alpha_{Bonf}\right)}{N}\right] \tag{27}$$

where N is the number of features on the array.

Another useful statistic easily obtained from this model is the posterior density of a fold-change in expression level. Since the ratio of normal densities is Cauchy, the fold-change statistic is the Cauchy density given by

$$f(u) = \frac{1}{\pi} \cdot \frac{\sigma_i^e / \sigma_i^c}{u^2 + \left(\sigma_i^e / \sigma_i^c\right)^2}, \tag{28}$$

where $u=\mu_i^e/\mu_i^c$ is the posterior fold change for the $i$ith array feature and σ_i^e and σ_i^c are the posterior density standard deviations for experimental and control conditions, respectively.

Implementation

The statistical methods and parameter estimation procedures described in this paper are implemented in a collection of command-line Java tools using standard object-oriented design techniques. This collection of tools, known as VAMPIRE (Variance-Modeled Posterior Inference with Regional Exponentials), handles estimation of the variance model parameters, calculates the maximum likelihood estimates for the exponential prior, computes the posterior densities for each feature in the array, and computes p-values for comparison of multiple experimental treatments to a set of controls. The MCMC framework used in VAMPIRE is a freely available Java package known as Hydra, written by Gregory Warnes. In all cases, we applied the *CustomMetropolisHastingsSampler* for simulation with a *NormalMetropolisComponentProposal* suggesting potential states. To estimate variance parameters, proposal states were generated for log(A) and log(B) with variances of 0.5. For the hyperparameter, we used (1) in log-transformed coordinates, a start state of 1.0

and variance of 1.0 propose new states and (2) in non-transformed coordinates, a start state of 0.0 and the proposal had a variance of 10^{-5}. Additionally, the open source library COLT is used for statistical computations and random number generation.

Analysis with SAM and Cyber-T

Both Cyber-T and SAM were used to complement the analysis performed by VAMPIRE, and to demonstrate the relative effectiveness of each method. Cyber-T was performed with the default parameters provided on the web site with the exception of setting the ìconfidenceî in the Bayesian estimate, ν, to the suggested value, 10. ìSignificantî features were defined as the subset of array features whose *p*-values resided below α=$2.00 \cdot 10^{-6}$ (a two-sided test with $\alpha_{Bonf} = 0.05$). Similarly, we used the two-class, unpaired data test in SAM while adjusting the δ parameter to achieve a median fraction of false significant genes of 0.05.

Array analysis of 3T3-L1 adipocytes

The previously described statistical methods were used to examine gene expression changes in 3T3-L1 murine adipocytes in response to treatment with several different thiazolidinediones (TZDs).

Cell Culture. 3T3-L1 fibroblasts were grown in DMEM-high glucose (25mM glucose, 1.8mM $CaCl_2$) medium containing 10% calf serum. 2 days postconfluency fibroblasts were differentiated into adipocytes in DMEM-high glucose containing 1ug/mL insulin, 0.1 ug/mL dexamethasone, and 112 ug/mL isobutylmethylxanthine to the medium. The differentiation medium was removed after 4 days and replaced with DMEM-high glucose containing 10% fetal calf serum, Glutamax, and 1% penicillin-streptomycin. The medium was changed every third day.

RNA preparation and microarray analysis. Twelve days post-differentiation, 3T3-L1 adipocytes were incubated for 24 hours with one of the following thiazolidinediones or with vehicle (0.1% DMSO): 20uM troglitazone, 20uM pioglitazone, 1uM rosiglitazone.

After treatment, total RNA was prepared from cells using Trizol (Invitrogen). Double stranded cDNA was synthesized with the SuperScript Choice system (Invitrogen) using a T7-$(dT)_{24}$ oligomer (Ambion) and the RNA from each treatment condition. Biotin labeled probes from the cDNAs were subsequently obtained by in vitro transcription using the BioArray High Yield RNA transcript Labeling Kit (Enzo). The probes were then hybridized to U74Aver2 Affymetrix GeneChips. The GeneChip microarrays were hybridized for 16 hours at 45^{o}C and 60 rpm in a rotisserie box. GeneChips were further processed using the standard Affymetrix protocols for eukaryotic GeneChips. Average Difference Scores for each gene feature were determined using the scanned GeneChip image and Affymetrix MAS 5.0 software. The microarray data described in this study has been deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database, and can be accessed at http://www.ncbi.nlm.nih.gov/geo (Accession: GSE1458).

D. Results

Variance model parameter estimation

Accurate estimation of the parameters of the variance model is critical for the interpretation of observed changes in gene expression. In order to account for the poor accuracy of m_i as an estimator of μ_i at low expression levels, we introduced the concept of an expression-level cutoff. Only when the sample mean of a particular microarray feature lies above the cutoff are the measurements from this feature used in parameter estimation. We expect that cutoffs set too low would give poor variance estimates. If the expression level cutoffs are set too high, we anticipate that the parameter estimates would become increasingly unstable as the number data points decreases. Between these two extremes is where a reasonable cutoff should lie. Two methods were employed to estimate variance parameters in each of the 3T3-L1 treatment conditions, and both give comparable results across a broad spectrum of expression-level cutoffs.

Similar behavior is observed for all of the treatment conditions with respect to the choice of expression-level cutoff. Estimates for expression-dependent variance, *A*, decrease with expression level cutoff until estimates by both methods level off. In contrast, estimates of expression-level independent variance, *B*, gradually increase as the cutoff increases, reach a plateau, and then become increasingly instable. When cutoffs are drawn at even higher levels, expression-dependent variance dominates until expression-independent variance is virtually undetectable by either method of estimation.

The plateau represents a broad region where the parameter estimates are stable against the cutoff choice. Within this region lie the most reliable point estimates for *A* and *B*. When no cutoff is used, we anticipated that variance estimates could be rather inaccurate since the MSD tend to underestimate the true variability of low-expression features. Since expression-independent variance typically dominates measurements of low-expression features, it is clear that the total variance for these features is drastically underestimated when no cutoff is used. For high-expression features, the accuracy of MSD is not as much of a problem, and the model compensates for the differential behavior of MSD by overestimating *A* and underestimating *B*. The opposing behavior between *A* and *B* continues as expression level cutoffs increase, until a plateau is reached.

In order to choose the best cutoff then, it is necessary to choose a point in the region where estimates of *A* and *B* plateau, prior to becoming unstable. MCMC, being a more robust statistical method, is much less sensitive to the choice of cutoff than the regression approach, and gives stable estimates for a broader range of cutoffs. An additional advantage is that its estimates are bound to physically meaningful values, where regression can potentially give negative estimates of variance parameters. It is therefore reasonable to choose MCMC as the final estimator of the parameters of variance given a choice of a cutoff. The benefits of MCMC simulation unfortunately come with increased computational cost. In order to strike a balance between accuracy and computation time,

the automated cutoff choice in VAMPIRE uses regression to ìtestî a spectrum of potential cutoffs. The automated cutoff is then determined as the mean of the broadest range of cutoffs for which there is less than 10% variation in $\sqrt{B}$, the expression-independent error. The cutoff choice is then supplied to the MCMC simulator for final parameter estimation. With this cutoff choice, the estimated variance is essentially the same across each of the treatment sets (Fig. 2.2). Illustrative of this point, when the parameter estimates for the control treatment are applied in place of the correct estimates for any of the TZD treatments, the number and identity of differentially expressed genes only changes by about 10-15 percent (results not shown). This variation is smaller than that observed between the differentially expressed genes obtained from each of the thiazolidinedione treatments.

In log-transformed coordinates, both methods of parameter estimation provide qualitatively good fits with mean squared deviations (MSD) beyond the cutoff where they are trained (Fig. 2.3). Below this cutoff however, MCMC-simulated parameter estimates appear to fit the MSD better than regression estimates. This is particularly noticeable in non-transformed coordinates. There are a couple of potential reasons for the superior quality of MCMC-simulated parameters. Least-squares regression is performed in log-transformed coordinates rather than the non-transformed coordinates that MCMC uses. Since the log-transformed equation of variance (12) is inherently inaccurate at low expression levels, this introduces some error into the regression parameter estimates. Applying the regression approach in non-transformed coordinates is even worse; it frequently gives negative variance parameters. This is because this method is a purely mathematical approach, and is not constrained to physically meaningful values. Since the curve is non-linear, least-squares regression is not expected to give the maximum-likelihood estimate of the parameters. Instead, as it tries to minimize square-distances, it

fits steep-sloped portions of the curve more tightly than milder-sloped portions, giving biased estimates. MCMC simulation is not hindered by these limitations.

These results also demonstrate the behavior of mean squared deviations at very low expression levels, where MCMC-derived estimates of variance no longer fit the observed MSD. As mentioned previously, this is a limitation of MSD as an estimator of σ_i^2, because of the use of m_i as a surrogate for μ_i. The downward bias introduced by employing this surrogate can be understood by examining the behavior of MSD. Given a mean expression $\bar{x}$, there is a fundamental limit to the maximum value that the mean square deviation can attain. This upper bound is given by

$$MSD_{\max} = (n-1)\cdot \bar{x}^2 \qquad (29)$$

where n is the number of replicate measurements (Appendix). When the variance is large compared to the signal, as is the case at low expression levels, MSD are depressed (Fig. 2.3b), and should not be used to estimate variance.

It is worthwhile to note that the use of a parameterized variance model is the major difference between current global and local variance model approaches. When using a global model, it is assumed that there is a strong correlation between variance and each independent variable, whereas a local model makes no such assumptions. It is clear that by plotting MSD against sample means that such a relationship exists, and furthermore can be well-parameterized by the model we have chosen. When the sample sizes are small, the global model of variance has several advantages over the local one. In Cyber-T, SAM, and other methods that do not take advantage of modeling variance across the array, some variant of MSD is used as the indicator of measurement uncertainty. This can be remarkably inaccurate. MSD, for the most part, can be considered as independent samplings from a density centered about the ìtrueî variance, approximated by the global model. When replicate numbers are low, these samplings may dramatically underestimate

or overestimate the underlying variance. At low expression-levels, these methods will theoretically detect too many genes to be differentially expressed as the variance is systematically underestimated. In most cases though, since the replicate number is taken into account, these methods will be overly restrictive for most array features, and as a whole, very few gene expression changes will be detected as statistically significant, due to the inherent uncertainty in small sample sizes.

The variance estimates for each of the treatment sets were very similar. This result is not altogether surprising because the sources of experiment-to-experiment variability are very similar. The same methods were used to derive cell lines, treat with TZD or DMSO, and the same protocol was used to purify mRNA, hybridize, and measure signals from each of the chips. The variance parameter estimates then, can be a useful indicator of experimental consistency. A dramatic increase in variance between experimental treatments could either indicate a fundamental impact of treatment on biological variability, or simply a flaw in the design or execution of the experiments themselves.

Prior parameter estimation and posterior shift

In order to find a local prior at each expression level, array features are first segmented into several non-overlapping expression regions. Within each region, a local exponential density is used to approximate the observed density of sample means (Fig. 2.4). As this exponential is parameterized by a single hyperparameter, λ, a probable prior can be computed by MCMC simulation of the marginal likelihood function (15). A point estimate for λ is thus computed for each expression region. Because of the curvature of the variance function though, and since there are far fewer array features that exhibit high expression levels, it is necessary to model high expression levels in log-transformed coordinates. Otherwise, there are simply too few features per region to obtain meaningful point estimates. In the log-transformed coordinate system, variance stabilizes as expression-levels increase, allowing for broader ranges of expression levels to be

included in each region. With this change, it is then possible to estimate the hyperparameter λ for an exponential prior in each high-expression region.

Although only a single point estimate of λ is computed for each expression region, there are a number of ways to extrapolate λ_i for each feature (Fig. 2.5). The most straightforward approach is to determine which region the *i*'th feature belongs to, and set λ_i equal to the corresponding estimate. With this method, the ìshiftî in the mean of the posterior becomes a discontinuous function. These discontinuities are likely an artifact of regional segmentation, and do not represent any fundamental behavior of the data. In order to adjust for this, a second approach was taken to extrapolate λ_i. The ìposterior shiftî from the latter term of (21), was computed at the mean of each expression region, and we applied piecewise continuous finite elements to approximate the best-fit line between these points. In addition, we constrained the finite elements to 2 fewer degrees of freedom than the number of data points we had to approximate with, and weighted each data point by the inverse variance of the posterior density for λ. This allowed the finite elements to fit the posterior shifts related to more accurate estimates of λ more tightly, while allowing freedom around vague estimates. This more sophisticated method of interpolation smoothes out the discontinuities between expression regions, but at a slightly greater computational cost. Surprisingly, the number of significant gene expression changes did not change by more than 5% when either method was applied, and the identity of significant changes was essentially the same (not shown). In our data, this may be because the posterior shift was small compared to measured expression levels. Therefore, although theoretically superior, finite element extrapolation of λ_i may not always be necessary to obtain useful results. In other data sets however, this prior may have a more marked effect. Therefore, both methods have been implemented in the VAMPIRE package.

Posterior inference

Once a local prior is calculated, and the posterior density for the ìtrue meanî expression is computed for each feature on the array, it becomes possible to make inferences on whether observed changes in gene expression are statistically significant. Based on the posterior densities for μ_i, a p-value for each array feature was computed, comparing each treatment to the control. As expected, when we analyzed the 3T3-L1 data, only a subset of the largest fold changes was determined to be statistically significant. The fold-change threshold for significance was much higher for low-expression features than for high-expression features. This dependence of the significance threshold on expression level mirrors the behavior of the log-transformed equation of variance (12).

The strong correspondence between our global variance modeling approach and traditional fold change approaches is clearly evident (Fig. 2.6). All of the features with statistically significant changes in gene expression can also be found by applying a standard two-fold/half-fold cutoff. Unfortunately, if we chose to use a fold-change cutoff alone, we would find so many spurious results at low expression levels that they would vastly outnumber those that are truly significant. Since expression-independent variance dominates at low-expression levels, and can constitute a large fraction of the observed signal, dramatic fold-changes in gene expression can often occur by random chance alone. To address this concern, many researchers also institute minimal expression-cutoffs, but optimal thresholds for maximizing sensitivity and specificity are difficult to determine *a priori*. Furthermore, the level of statistical significance of the chosen threshold is difficult to determine. Variance-modeled posterior inference inherently addresses these issues. Once an experiment-wide false positive rate is specified, the fold change thresholds required for statistical significance are implicitly computed across all

expression levels. There is little need for additional filtering of low-intensity measurements prior to analysis.

We employed two additional methods for the analysis of our 3T3-L1 microarray data. Both SAM, which uses a non-parametric bootstrap method, and Cyber-T, which applies Bayesian statistics for a modified *t*-test, have become popular tools. The first set of analysis was performed by pooling TZD treated data into a single ìtreatedî data set (n=8), and compared to DMSO controls (n=3). With an experiment-wide false-positive rate of 0.05 for Cyber-T and VAMPIRE, and a false-discovery rate of 0.05 for SAM, all three methods showed substantial overlap in the sets of genes considered to be differentially expressed (Fig. 2.7). Nearly three out of four genes detected by Cyber-T (default parameters, ν=10) were also detected by VAMPIRE, while over half of the ìsignificantî features in SAM were also ìsignificantî in VAMPIRE. The considerable overlap is strongly suggestive that the genes detected by each of these methods are indeed differentially expressed. Moreover, since the number of detected features are very similar, it is likely that all three methods share a comparable level of sensitivity under these conditions.

At an estimated experiment-wide false-positive rate of 0.05, each analysis tool also produced a set of genes that were not detected by either of the other methods. Among the 23 genes that were only identified by VAMPIRE, 4 had known function in glucose metabolism, lipid metabolism or gene regulation by Gene Ontology. These included phosphoenolpyruvate carboxykinase 1 (Pck1), acetyl-CoA acetyltransferase 2 (Acat2), fatty acid CoA ligase 2 (Facl2), and growth arrest and DNA-damage-inducible 45 gamma (Gadd45g). At the same time, of the 24 genes that only SAM considered statistically significant, only 1 had known function in metabolism or gene regulation. This sole gene was Cbp/p300-interacting transactivator 2 (Cited2). Cyber-T detected very few additional genes in pooled TZD data, and none of these had related functions. The identities of the

remaining genes are listed in tables 2.1, 2.2, and 2.3. While the differences in numbers of Gene Ontology annotations are probably not statistically significant, these results do demonstrate that when the number of replicates is large, and the desired false positive rate is low (0.05), all of these methods also show a similar level of specificity.

In most microarray experiments, investigators rarely use ample replication to assess gene expression at the greatest level of precision. In most cases, only two or three replicates of each treatment condition are performed. Similarly, we compared individual TZD treatments to controls to observe the effect of this low level of replication. Neither SAM nor Cyber-T was able to reproduce a similar set of differentially regulated genes when analyzing unpooled, individually-treated data. In both cases, only a handful of genes were found for each TZD treatment, and there was minimal overlap between the ìsignificantî genes. This contradicts our intuitive understanding that PPAR-γ ligands should influence the expression of the same set of genes. In contrast, VAMPIRE found an average of 50 differentially-expressed array features in each of the individual treatments. These sets had substantial overlap with each other, as well as with the pooled TZD analysis, sharing a total of 23 genes (Fig. 2.8). Furthermore, of the 24 genes that were selected by all three statistical methods on the pooled data, 12 of these were also found by VAMPIRE to be ìsignificant changesî in all three individual treatments. It is therefore apparent that even when the number of experimental replicates is low, our method retains ample sensitivity and specificity, which is typically lost by other approaches.

E. Discussion

The detection of differentially-expressed genes is a common starting point for the interpretation of microarray data. At high replicate numbers, VAMPIRE, Cyber-T, and SAM all perform similarly. The results suggest that our global model of variance is suitable for understanding the degree of experimental variability for each feature on the array. At low replicate numbers, our model, adapted from Rocke and Durbin (2001),

takes advantage of a known relationship between variance and gene expression level to extrapolate an estimate of variance beyond what is available in the few experimental replicates. Without using such a model, the measurement variability is essentially left undetermined, and a feature-specific test can identify very few significant expression changes. Therefore, both SAM and Cyber-T are only able to identify very few ìstatistically-significantî genes from smaller data sets.

Although it is still ideal to collect a greater number of replicates for an optimally-controlled microarray experiment, our method attempts to maximize the value of the available data set. In comparison to current approaches, VAMPIRE appears to be much more sensitive at low replicate numbers for identifying these changes while preserving a considerable level of specificity. In addition, the relationships between the choice of fold-change thresholds, expression-level cutoffs, and the levels of statistical significance are not entirely clear. The correspondence between our method and fold-change methods can be qualitatively understood by examining ìsignificantî fold-changes at each expression level. At very high expression levels, the results of a two-fold cutoff are very similar to results from VAMPIRE. At low expression levels however, where features are typically removed in the fold-approach by a minimum-expression cutoff, VAMPIRE detects significance by implicitly adjusting the fold-change threshold. The statistical framework used in our method can clearly serve as a basis for quantitatively understanding traditional ideas of microarray analysis.

The empirical methods that we employ in our modeling procedure may be sophisticated, but they are grounded in conceptually simple ideas. VAMPIRE uses various estimation techniques to obtain robust parameter estimates directly from microarray data, and uses the observed data to narrow the posterior to a normal distribution. The parameters of the variance model are computed by ìfittingî the mean-square deviations, while accounting for the uncertainty involved in using the sample

mean as a surrogate for the true expression mean. To obtain optimal estimates, our method takes advantage of commonly available computational power by applying Markov Chain Monte Carlo simulation. The prior likelihood density can also be readily understood, although a familiarity with the Bayesian concept of prior and posterior likelihoods is necessary. When a prior density based on prior knowledge is not available, we rely on the distribution of gene expression across the array as a prior for the ìtrue meanî expression. When a greater number of features exhibit a given expression level, the prior probability at that expression level increases. Based on this concept, an estimate for the prior parameter, λ, is determined by employing an assortment of computational methods.

Besides its utility for inference in differential gene expression, the parameter estimates can provide additional information about the nature of the data we collect. In the case of the two-component variance model, the expression-dependent variance, A, and expression-independent variance, B, provide insight into the magnitude of variance at each level of gene expression. In our 3T3-L1 adipocyte data, we observed a relatively consistent set of parameter estimates, despite varying treatment conditions. This suggests that the sources of variance are relatively consistent across our experiments. On the other hand, if substantially different estimates are obtained for two similarly-treated experiments, this may be a sign that there are inconsistencies in experimental procedures, or sources of uncontrolled variability. Furthermore, if the variance estimates are reliably different despite rigorous controls, then this suggests that the biological variability may be significantly different due to the treatments themselves.

The current model adapted into VAMPIRE currently accounts for two-classes of variance in a global model, but can be extended to provide more detailed modeling of feature-specific variability. While there are undoubtedly other components of measurement variance that are dependent on each individual gene, it is not immediately

clear what the corresponding relationships are. For example, it is possible that the variance may be dependent on the length or the nucleotides used in the probes for each feature. If such a relationship can be characterized, this information can then be incorporated into an improved model to obtain improved variance estimates. As with any model however, increasing refinement will also increase the number of unknown parameters, which become increasingly vague without additional sources of data. Because of this, a more refined model may or may not show increased predictive ability.

A number of Bayesian models have recently been described in the literature, which have further developed the approach taken by Baldi and Long (2001). The empirical Bayesmethod devised by Smyth relies on array data to stably estimate the hyperparameters of a hierarchical model, which explicitly incorporates a linear model to describe the relationships between gene expression in different treatment conditions[25]. Tests may then be performed on the ìcontrast estimatorsî to determine whether observed expression changes are statistically significant. The moderated *t*-statistic used by Smyth is conceptually equivalent to the statistic Z_i that we describe here. We believe that the global variance approach that we apply is likely to be more robust when the number of replicate measurements is low, for the same reasons that we have described previously. It is not clear however, whether this continues to be the case when large numbers of array replicates are available, and the sample variances become more accurate estimators of feature-specific variability.

The power of microarray analysis presumably lies far beyond that of determining differential-expression alone. These snapshots of gene expression can be interpreted across a variety of experimental conditions by a third tier of array analysis, which often involves a clustering approach. Current clustering methods commonly rely on deterministic fold-changes of gene expression for computing ìdistancesî between expression patterns of various genes. These methods are inherently limited by the same

problems that limit a fold-change approach for detecting differentially-expressed genes. The experimental variability of low-expression genes is large enough that large fold-changes may not represent true changes in gene regulation. It is therefore essential to incorporate an estimate of variability into clustering, and adapt a stochastic interpretation of a fold-change. For clustering methods that rely on fold-changes, we can apply the posterior density of the fold-change, which our model predicts is a Cauchy density. In place of a deterministic distance metric, we can instead apply the average distance between the expression patterns of each pair of genes, integrating over the likelihood densities. This ìaverage distanceî metric still satisfies necessary requirements for a mathematical concept of distance: it is symmetric, positive definite with equality when the distance between a profile and itself is computed, and obeys the triangle inequality (although there may be multiple expression patterns that have zero-distance between them). If an absolute measure of gene expression is desired for clustering instead of a fold-change measure, the corresponding posterior density is even more convenient; it is normal. Again, an ìaverage distanceî metric can be applied. The use of these posterior densities and a stochastic concept for measurements of gene expression should improve the reliability of clustering methods, and more generally, enhance the knowledge that we harvest from our microarray experiments.

F. Acknowledgements

This work was supported by grants from the National Institutes of Health, NIDDK RO1-DK33651 (JMO), NIDDK KO1-DK62025 (DSW), and NIGMS K54 GM62114 (SS), a grant from LSI and Pfizer La Jolla, it103-10124 (JMO and SS), and a grant from the Hillblom Foundation (JMO and SS). AH is a graduate student in the UCSD Medical Scientist Training Program and is supported by a Fellowship from the Whitaker Foundation and the UCSD MSTP Training Grant T35-GM07198.

G. Appendix

Properties of log-transformed data.

Let $X_1, X_2, \ddot{O}, X_n$ be I.I.D. random variables with known mean μ, and variance σ^2.

$$E[X] = \mu \tag{A1}$$

$$Var[X] = \sigma^2 \tag{A2}$$

Let $Z \equiv \log(X)$.

$$E[Z] = E_X[\log(X)] \tag{A3}$$

$$Var[Z] = E_X\left[\log^2(X)\right] - E_X^2[\log(X)] \tag{A4}$$

Expanding a $\log(x)$ and $\log^2(x)$ about μ,

$$\log(x) = \log(\mu) + \frac{x-\mu}{\mu} - \frac{1}{2}\frac{(x-\mu)^2}{\mu^2} + \frac{1}{3}\frac{(x-\mu)^3}{\mu^3} + \ldots \tag{A5}$$

$$\log^2(x) = \log^2(\mu) + 2\cdot\log(\mu)\frac{x-\mu}{\mu} + (1-\log(\mu))\frac{(x-\mu)^2}{\mu^2} - \left(1-\frac{2}{3}\log(\mu)\right)\frac{(x-\mu)^3}{\mu^3} + \ldots \tag{A6}$$

Correspondingly,

$$E[\log(X)] = \log(\mu) + E\left[\frac{x-\mu}{\mu}\right] - \frac{1}{2}E\left[\frac{(x-\mu)^2}{\mu^2}\right] + \frac{1}{3}E\left[\frac{(x-\mu)^3}{\mu^3}\right] + \ldots \tag{A7}$$

$$E\left[\log^2(X)\right] = \log^2(\mu) + 2\cdot\log(\mu)\cdot E\left[\frac{x-\mu}{\mu}\right] + (1-\log(\mu))\cdot E\left[\frac{(x-\mu)^2}{\mu^2}\right] - \left(1-\frac{2}{3}\log(\mu)\right)E\left[\frac{(x-\mu)^3}{\mu^3}\right] + \ldots \tag{A8}$$

Applying the mean of X,

$$E[\log(X)] = \log(\mu) - \frac{1}{2}\frac{\sigma^2}{\mu^2} + \frac{1}{3}E\left[\frac{(x-\mu)^3}{\mu^3}\right] + \ldots \tag{A9}$$

$$E\left[\log^2(X)\right] = \log^2(\mu) + (1-\log(\mu))\cdot\frac{\sigma^2}{\mu^2}\cdot + \ldots \tag{A10}$$

Assuming that $\sigma << \mu$, then on average, $x - \mu << \mu$:

$$E[\log(X)] \approx \log(\mu) - \frac{1}{2}\frac{\sigma^2}{\mu^2} \tag{A11}$$

$$E^2[\log(X)] \approx \log^2(\mu) - \log(\mu)\frac{\sigma^2}{\mu^2} + \frac{1}{4}\frac{\sigma^4}{\mu^4} \tag{A12}$$

$$E\left[\log^2(X)\right] \approx \log^2(\mu) + (1 - \log(\mu)) \cdot \frac{\sigma^2}{\mu^2} \tag{A13}$$

After some algebra, we obtain the following simple expressions for the properties of $Z \equiv \log(X)$ when $\sigma << \mu$:

$$E[Z] \approx \log(\mu) - \frac{1}{2}\frac{\sigma^2}{\mu^2} \approx \log(\mu) \tag{A14}$$

$$Var[Z] \approx \frac{\sigma^2}{\mu^2} - \frac{1}{4}\frac{\sigma^4}{\mu^4} \approx \frac{\sigma^2}{\mu^2} \tag{A15}$$

Expression region segmentation.

For the sake of reducing computational cost, microarray features are segmented into non-overlapping regions that have relatively constant variance. At low expression levels, these regions are computed in non-transformed coordinates where the slope of the variance curve is flattest (Fig. 2.9). Here, the hyperparameter is only calculated once at the mean of each region. Based on the equation of variance (10), the boundaries between non-transformed expression regions are determined according to

$$\mu_1^n = \sqrt{\frac{\rho B}{A}} \quad \mu_{i+1}^n = \sqrt{(2+\rho)\mu_i}\,, \tag{A16}$$

where ρ is the maximum fractional variation allowed in non-transformed variance for any region and μ_i^n are the successive boundaries of non-transformed regions.

At high expression levels, the width of non-transformed expression regions becomes increasingly narrow, until very few array features lie within any individual segment. Since it is not possible to make accurate assessments of λ with so few data points, it is necessary to compute high expression regions in transformed coordinates. Regions are chosen so that log-transformed variance can be maintained nearly constant. The boundaries of log-transformed regions are obtained from the transformed equation of variance (12) by

$$\mu_1^l = \sqrt{\frac{B}{\tau A}} \quad \mu_{i+1}^l = \sqrt{\frac{\mu_i}{2+\tau}}\,, \tag{A17}$$

where τ is the maximum fractional variation allowed in log-transformed variance for any region.

We have chosen the transition to hyperparameter modeling log-transformed coordinates to lie at the point where the log-transformed variance equation is sufficiently accurate (Fig. 2.10). As previously mentioned, this equation is most accurate at high expression levels, where the standard deviation σ is small compared to μ. The error in the log-transformed variance equation is approximated by

$$\left|\frac{\sigma^2}{\mu^2} - Var[\log(x)]\right| \approx \frac{1}{4}\frac{\sigma^4}{\mu^4}, \tag{A18}$$

and from this, we derive the following heuristic

$$\mu > \sqrt{\frac{B}{4\gamma - A}} \tag{A19}$$

where γ is the maximum acceptable fractional error in the log-transformed variance.

Boundaries between expression regions.

The variance model can be written as

$$Var[X] = A\mu^2 + B \tag{A20}$$

where A is the expression-dependent and B is the expression-independent variance. Defining ρ as the maximal accepted fractional change in variance, then all μ that satisfy the inequality

$$\frac{A\mu^2}{B} \leq \rho \tag{A21}$$

are included in the first expression region. The upper bound for this region is therefore

$$\mu_1^n = \sqrt{\frac{\rho B}{A}}. \tag{A22}$$

Successive regions can be iteratively found by computing the incremental change in variance over the previous regional boundary. Each region includes all μ that satisfy

$$\frac{A\mu^2 - A\mu_i^2}{A\mu_i^2 + B} \le \rho. \qquad \text{(A23)}$$

Solving for the boundary gives

$$\mu_{i+1}^n = \sqrt{(2+\rho)\mu_i}. \qquad \text{(A24)}$$

A similar procedure can easily be performed to determine the boundaries of each log-transformed expression region.

Comments on the posterior shift and λ_i.

In order to reduce computational time, λ is only estimated by MCMC at the mean of each non-overlapping ì expression regionî. Given that each feature on the array should have its own local prior, there are a number of ways to decide what value of λ_i to use. The simplest method is to use the same λ_i for all features whose sample means lie within a given region. This requires very little additional computation, but could potentially give inaccurate results because of discontinuities between regions. One alternative is to approximate values for λ_i between computed locations by applying piecewise continuous and differentiable finite elements. We implement both methods in our analysis.

While it is probably effective to take the maximum likelihood λ_i as a point estimate, it is only marginally more difficult to apply the entire simulated likelihood density. The posterior density of μ_i is normal under our model, and the linear combination of normal variates is also normal. In addition, the same linear combination gives us the mean of this combined density. If we instead choose the mean of posterior estimate, we effectively integrate the posteriors with respect to the likelihood of each prior.

Upper limit on MSD.

The mean square deviation is computed as

$$MSD = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2 \tag{A25}$$

or alternatively,

$$MSD = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - \bar{x}^2 \cdot \tag{A26}$$

Maximizing MSD with respect to $\bar{x}$,

$$MSD_{\max} = \frac{1}{n}\max\left(\sum_{j=1}^{n}x_j^2\right) - \bar{x}^2 \cdot \tag{A27}$$

It can be easily shown that when $x_j \geq 0$ for all j

$$\max\left(\sum_{j=1}^{n}x_j^2\right) = n^2\bar{x}^2 \cdot \tag{A28}$$

Therefore, the upper bound is given by

$$MSD_{\max} = (n-1)\cdot\bar{x}^2 \tag{A29}$$

where n is the number of replicate measurements.

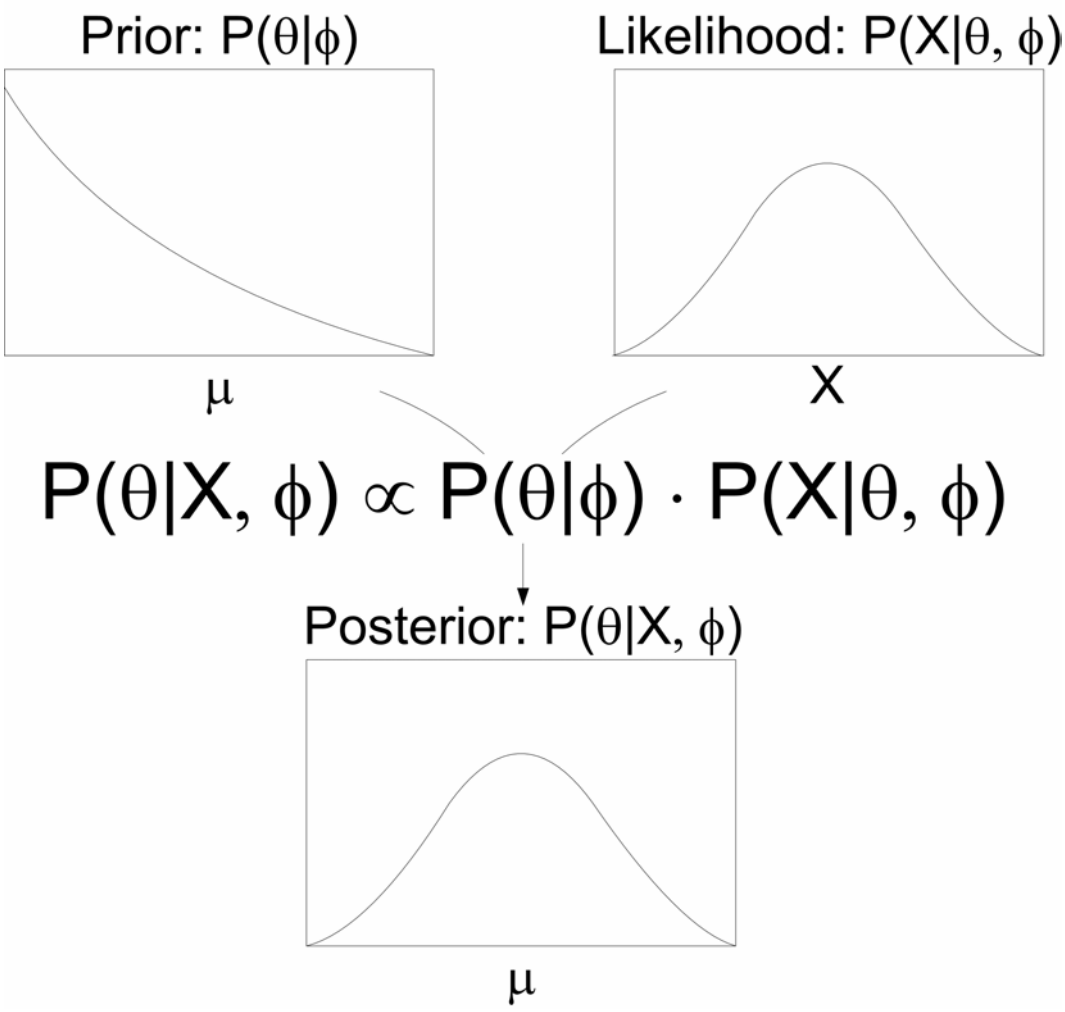


FIGURE 2.1

Schematic of the Bayesian model. An exponential prior density based on the local distribution of gene expression is narrowed down to a normally-distributed posterior density as individual measurements are observed. The variable X represents a vector of observed measurements, while θ is the vector of model parameters and ϕ is the vector of model hyperparameters.

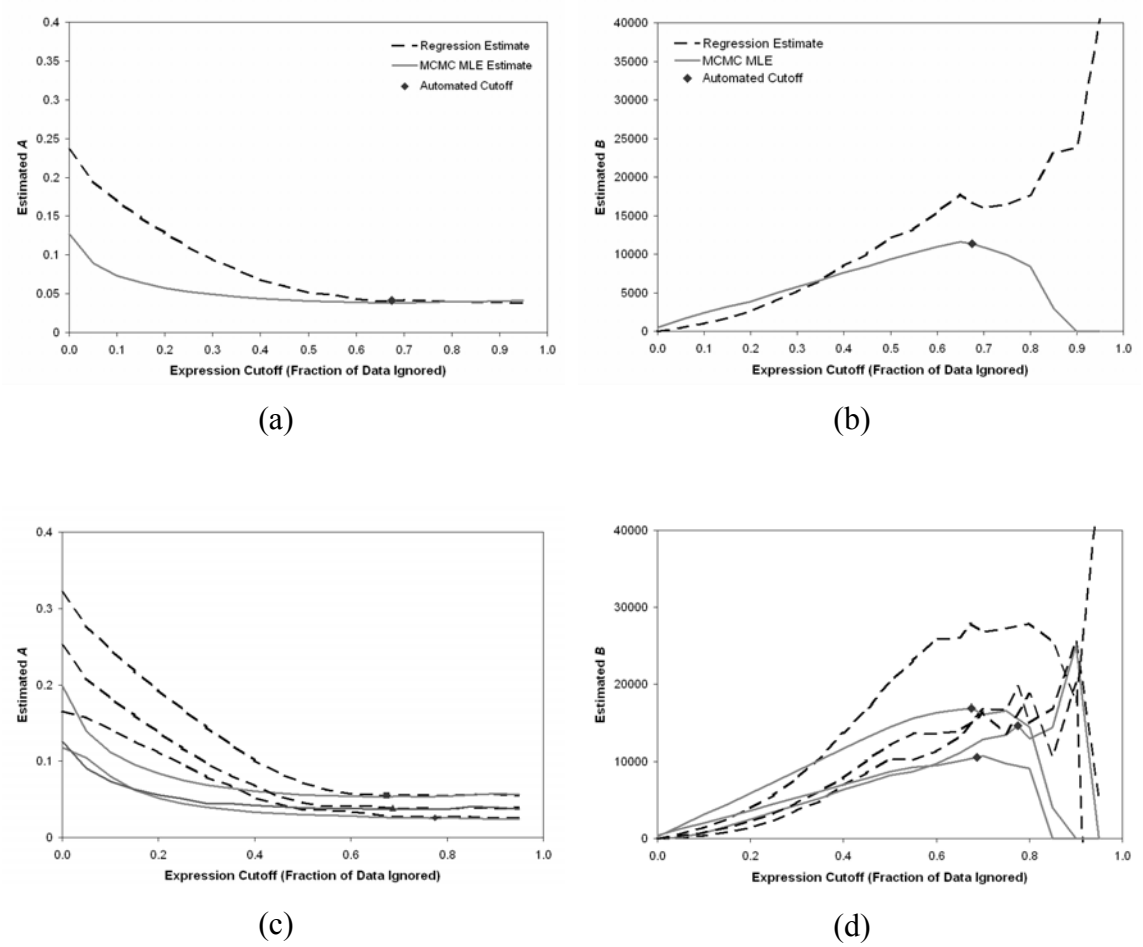
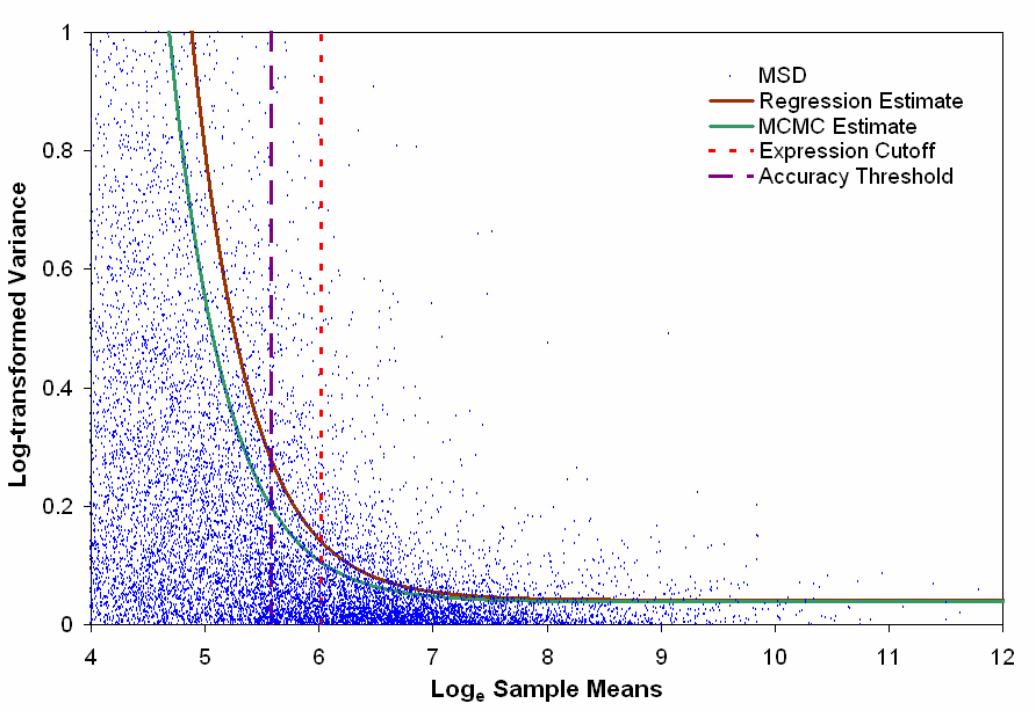


FIGURE 2.2

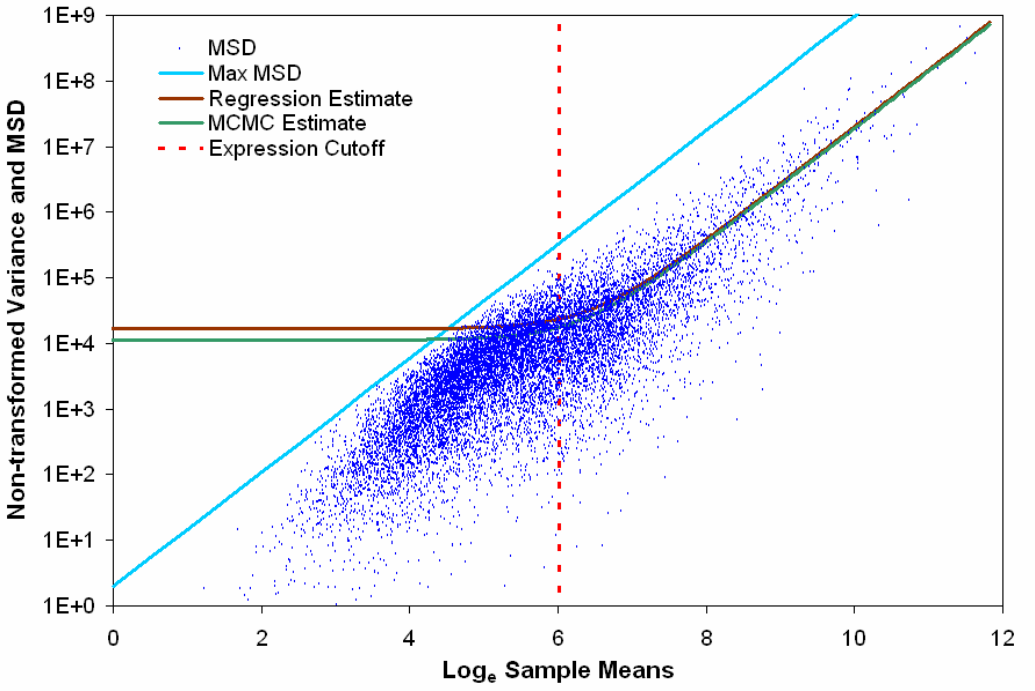
Variance model parameter estimation. Effect of minimum expression cutoff on the estimated variance parameters in DMSO-treated cells (a, b). In this data set, only the upper 32.5 percentile of array features was used by the automated cutoff choice. This corresponded to an average difference score of 410.8. Similar behavior was observed for each of the thiazolidinedione-treated data sets (c, d). The solid lines demonstrate the relative stability of MCMC-derived parameter estimates when compared to regression estimates given by dashed lines.

FIGURE 2.3

Variance model parameter fit. Comparison of methods of variance model parameter estimation for DMSO-treated 3T3-L1 (n=3) data in (a) log-transformed and (b) non-transformed coordinates. Both regression and MCMC models are trained on data beyond the minimum expression level cutoff. In (a), the log-transformed equation of variance is only expected to attain the desired level of accuracy above the threshold shown.


MSD
Regression Estimate
MCMC Estimate
Expression Cutoff
Accuracy Threshold
Log-transformed Variance
Log$_e$ Sample Means
0
0.2
0.4
0.6
0.8
1
4
5
6
7
8
9
10
11
12

(a)


MSD
Max MSD
Regression Estimate
MCMC Estimate
Expression Cutoff
Non-transformed Variance and MSD
Log$_e$ Sample Means
1E+0
1E+1
1E+2
1E+3
1E+4
1E+5
1E+6
1E+7
1E+8
1E+9
0
2
4
6
8
10
12

(b)

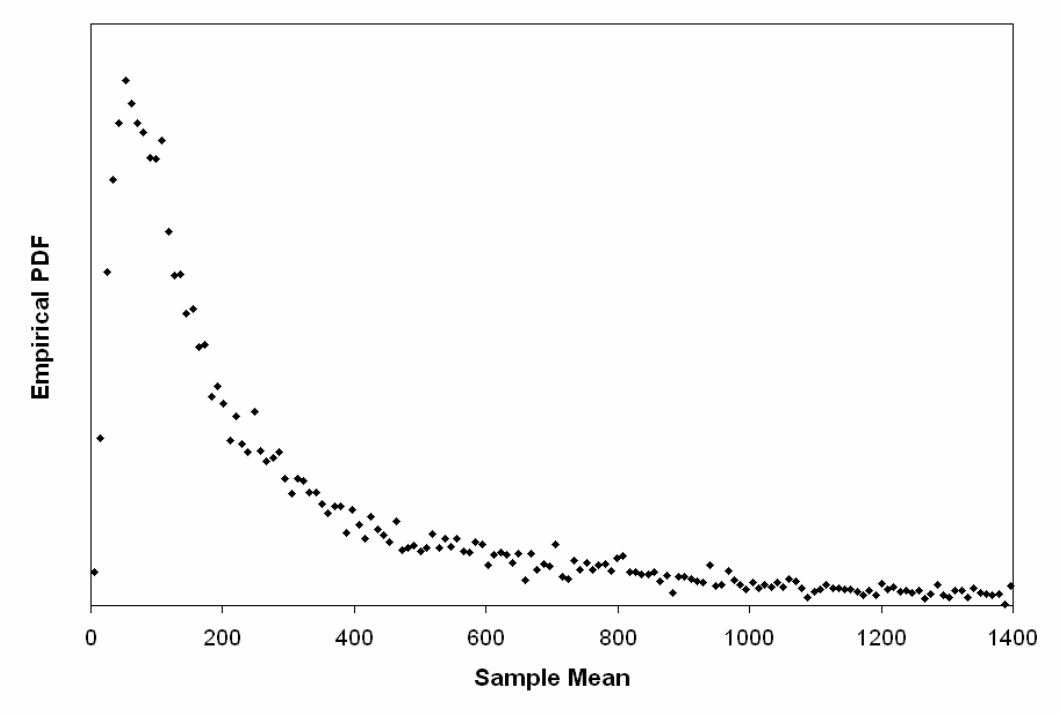


FIGURE 2.4

Observed prior density. The observed density of DMSO-control sample means was obtained by binning. This density is decomposed into piecewise exponentials when local priors are computed for each expression region.

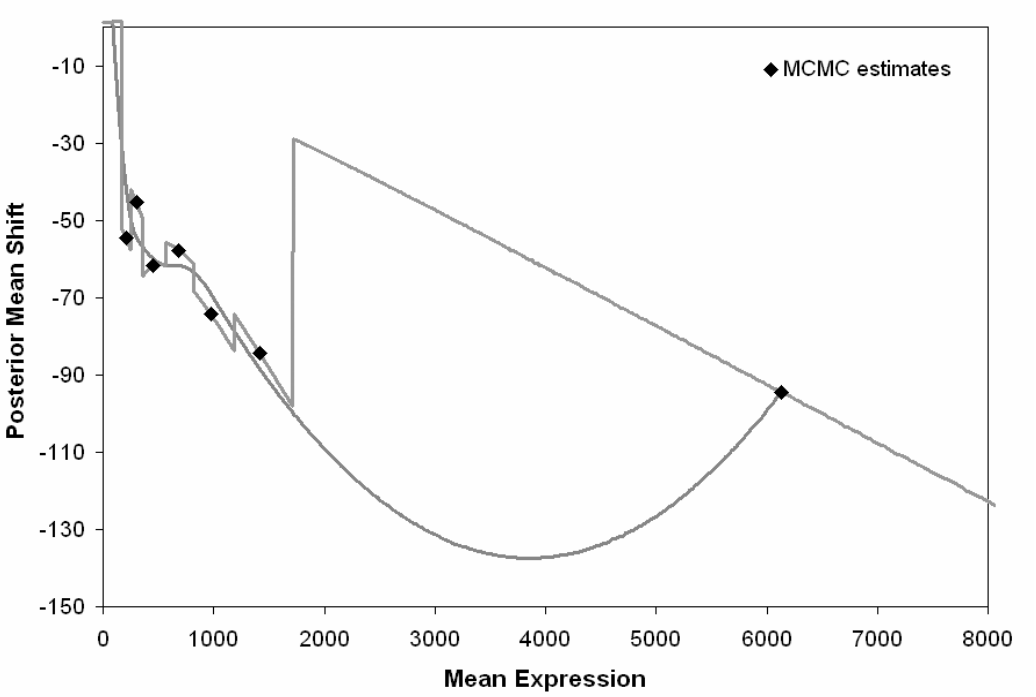


FIGURE 2.5

Posterior shift. The posterior ìshiftî was computed for each expression level of DMSO-treated 3T3-L1 adipocytes. The jagged line represents the shift when MCMC-estimated lambda are used for entire expression regions. The curved line is computed by finite-element interpolation of the shift between MCMC estimates.

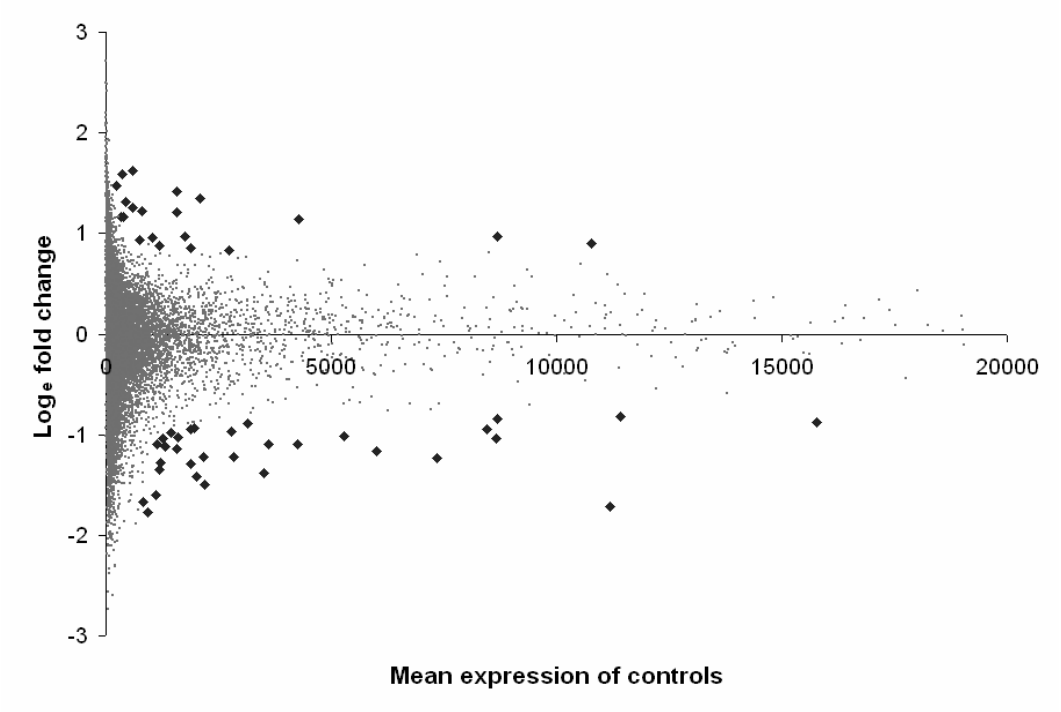


FIGURE 2.6

Significant expression changes. Expression-level changes in pioglitazone-treated cells (*n*=2) compared to control (*n=3*). The larger, highlighted spots were considered to be significant changes by VAMPIRE.

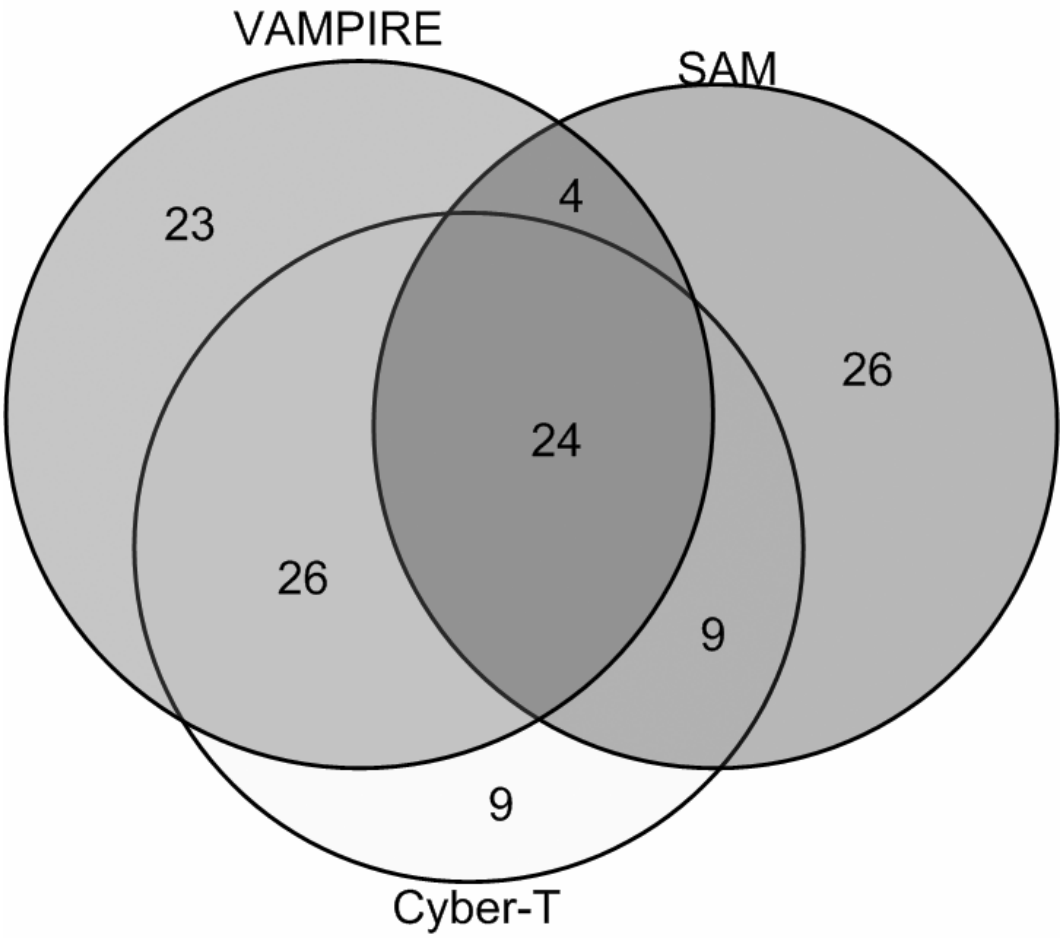


FIGURE 2.7

Pooled TZD comparison. Three statistical methods were compared on pooled TZD data. When comparing pooled TZD data ($n=8$) to DMSO controls ($n=3$), many of the same genes are recovered by all three approaches ($\alpha_{Bonf}=0.05$).

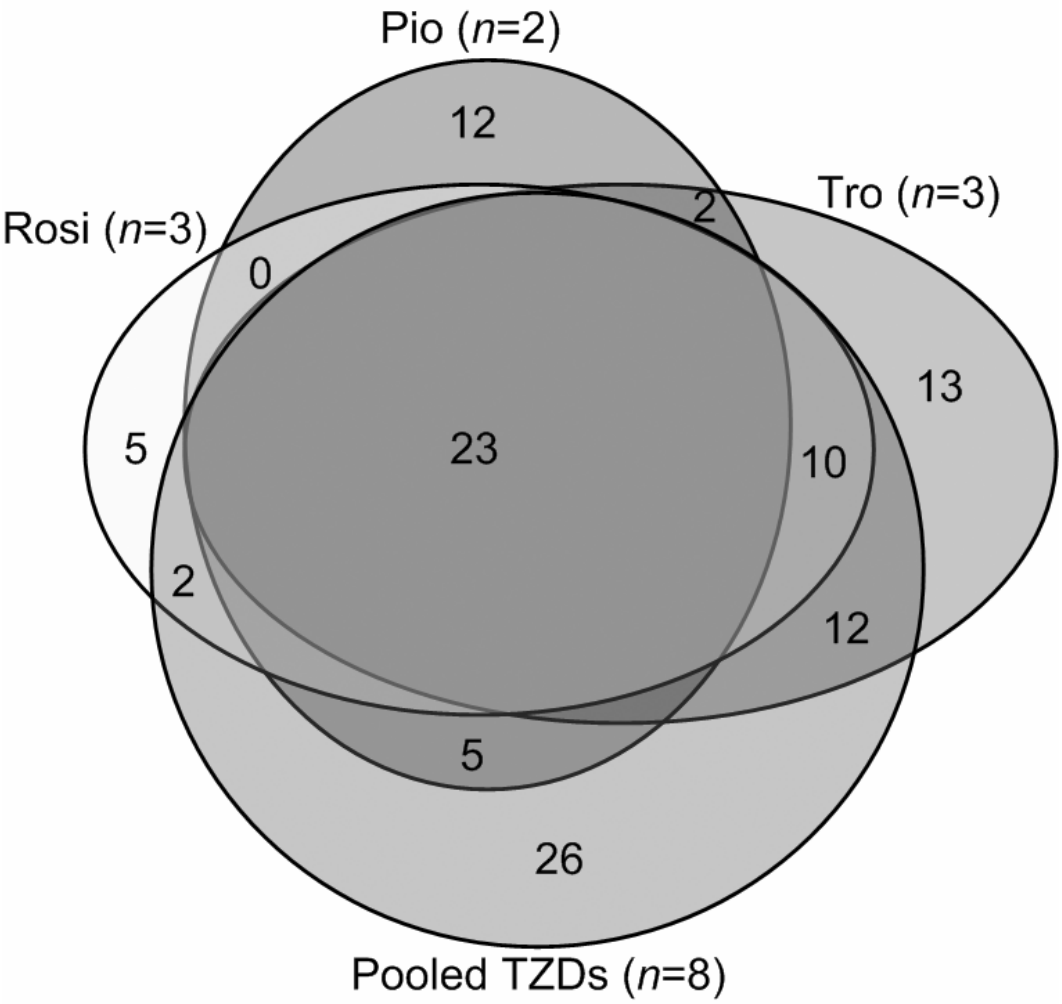


FIGURE 2.8

Individual TZD VAMPIRE analysis. The results of individual analysis by VAMPIRE are displayed above. The sets of ìstatistically significantî gene expression changes for each TZD treatment are well correlated with the same analysis on pooled data (α_{Bonf}=0.05).

FIGURE 2.9

Shape of the variance equation (non-transformed). Modeling in non-transformed coordinates occurs to the left of an accuracy threshold indicated by the vertical line.

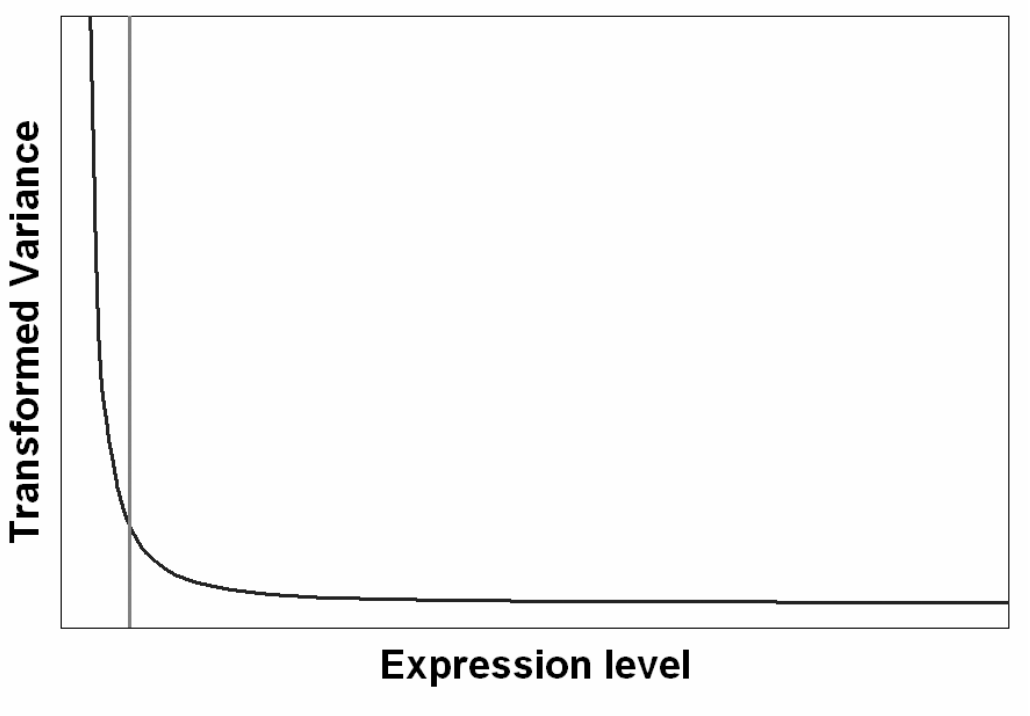


FIGURE 2.10

Shape of the variance equation (log-transformed). Features that lie to the right of the accuracy threshold are modeled in log-transformed coordinates.

TABLE 2.1

Features detected as significantly increased by VAMPIRE in response to thiazolidinedione treatment. The numbers listed under VAMPIRE are the fold-ratios between the means of the posteriors. The columns on the right show several standard statistics for this data set. The column labeled ìCî lists the mean average-difference scores from MAS 5.0 of DMSO-treated 3T3-L1s. The rightmost columns show the mean and standard error of fold-change, computed from all pairwise combinations of TZD-treated to control measurements. Legend: C, vehicle control; P, pioglitazone; R, rosiglitazone; T, troglitazone; X, pooled TZD treatments.

		SAM				Cyber-T				VAMPIRE			
probe	gene	P	R	T	X	P	R	T	X	P	R	T	X
Z37107	Ephx2				U				U	**2.6**		**2.7**	**2.5**
AF006688	Acox1				U				U	**2.5**		**2.7**	**2.6**
M61737	Fsp27				U				U	**2.5**			**2.2**
AW122615	---		U	U	U				U				**2.0**
M80206	Pvrl2			U	U				U				**2.1**
AI326963	---	U			U				U				**2.1**
AB017112	Slc25a20				U				U				**2.2**
AA755260	Dp1l1				U				U				**2.2**
AA797604	Angptl4				U				U				**2.2**
AW125480	1620401E04Rik						U	U	U	**5.5**	**5.8**	**9.3**	**7.1**
AJ001418	Pdk4							U	U	**3.6**	**3.2**	**4.3**	**3.7**
M14044	Anxa2						U		U	**3.2**	**3.8**	**3.7**	**3.6**
AW012588	MGC29978					U			U	**3.9**	**3.8**	**3.8**	**3.8**
U48403	Gyk								U	**4.0**	**4.8**	**3.4**	**4.1**
M93275	Adfp								U	**2.7**		**2.9**	**2.5**
AI324801	Cbr3								U	**2.6**		**2.7**	**2.6**
AF007267	Pmm1								U	**5.3**		**5.8**	**4.7**
U08439	Cox6a2			U					U			**2.8**	**2.4**
D26107	Pvrl2								U			**2.6**	**2.5**
AI841389	---								U			**3.2**	**2.7**
M22998	Slc2a1								U			**4.1**	**3.1**
AA871791	1300003D03Rik								U				**1.9**
AI838015	Prss15								U				**1.9**
AW047688	0610039N19Rik								U				**2.1**
AW047630	Ccrn4l								U				**2.3**
AI530403	Acaa1								U				**2.3**
AF064635	Hsd17b12								U				**2.3**
AF036164	Serpinf1								U				**2.4**
K03235	Plf								U				**2.6**
AI255972	4931406C07Rik								U				**3.5**
AV361373	Pdc				U					**3.7**	**3.8**	**3.0**	**3.5**
AI837130	Itpk1			U								**2.5**	**2.1**
AA798624	Ero1l											**3.2**	**1.9**
AW123751	2310056P07Rik											**2.4**	**2.1**
Z12604	Mmp11											**3.1**	**2.2**
M31658	Ghrh											**3.0**	**2.3**
X51829	Myd116											**4.4**	**2.7**
AV298880	Rbm9											**4.1**	**3.4**
Y12577	Arl4										**2.9**		**2.1**
U01310						U				**3.5**			**2.1**
AF009605	Pck1					U				**4.3**			**2.4**
X01756	---									**2.8**			**2.7**
AA823381	Ndufa5												**1.9**
U15977	Facl2												**2.0**
AW049373	2310016A09Rik												**2.0**
AI844355	2310042G06Rik												**2.1**
AF055638	Gadd45g												**2.2**
AV250694	---												**2.2**
AV347674	---												**2.3**
M35797	Acat2												**2.3**
AI839083	Pus1												**2.3**
AI255972	4931406C07Rik												**2.4**
AF109905	---												**3.1**
AA764261	Pim1											**2.4**	
AF003348	Npc1											**2.5**	
K02236	Mt2											**2.5**	
AW120614	---											**2.8**	
X67083	Ddit3											**2.9**	
X56824	Hmox1									**2.4**			
AI837948	E130012A19Rik									**3.5**			

TABLE 2.2

Features detected as significantly decreased by VAMPIRE in response to thiazolidinedione treatment. The numbers listed under VAMPIRE are the fold-ratios between the means of the posteriors. The columns on the right show several standard statistics for this data set. The column labeled ìCî lists the mean average-difference scores from MAS 5.0 of DMSO-treated 3T3-L1s. The rightmost columns show the mean and standard error of fold-change, computed from all pairwise combinations of TZD-treated to control measurements. Legend: C, vehicle control; P, pioglitazone; R, rosiglitazone; T, troglitazone; X, pooled TZD treatments.

		SAM				Cyber-T				VAMPIRE			
probe	**gene**	**P**	**R**	**T**	**X**	**P**	**R**	**T**	**X**	**P**	**R**	**T**	**X**
U49430	Cp			D	D			D	D	**0.3**	**0.3**	**0.2**	**0.3**
M96827	Hp				D			D	D	**0.3**	**0.3**	**0.2**	**0.3**
X95279	Thrsp				D			D	D	**0.4**	**0.4**	**0.3**	**0.3**
AW122933	Enpp2	D	D	D	D	D	D		D	**0.2**	**0.2**	**0.3**	**0.2**
L07924	Ralgds				D	D	D		D	**0.3**	**0.3**	**0.3**	**0.3**
U10374	Pparg				D	D			D	**0.2**	**0.4**	**0.2**	**0.3**
D86344	Pdcd4			D	D				D	**0.4**	**0.2**	**0.2**	**0.3**
AI854522	Pcdhga12			D	D				D	**0.3**	**0.3**	**0.4**	**0.4**
X78445	Cyp1b1				D				D	**0.2**	**0.2**	**0.3**	**0.3**
U66166	Sparcl1				D				D	**0.2**	**0.2**	**0.3**	**0.3**
U35374	Pnp				D				D	**0.3**	**0.2**	**0.3**	**0.3**
AI854020	Cdo1				D				D	**0.4**	**0.3**	**0.3**	**0.4**
M27008	Orm1				D				D	**0.3**		**0.3**	**0.4**
X13135	Fasn				D				D	**0.4**		**0.3**	**0.4**
U38196	Mpp1				D				D		**0.4**		**0.4**
AF023919	Itih4								D	**0.3**	**0.2**	**0.4**	**0.3**
J04953	Gsn								D	**0.2**	**0.2**	**0.3**	**0.3**
AJ002390	Anxa8					D		D	D	**0.3**		**0.3**	**0.4**
M27008	Orm1								D	**0.4**		**0.3**	**0.4**
AI838274	Slc29a1								D			**0.1**	**0.2**
X72862	Adrb3				D					**0.3**	**0.3**	**0.3**	**0.3**
AI840158	Angptl2				D					**0.3**	**0.3**	**0.2**	**0.3**
U67187	Rgs2				D					**0.2**		**0.3**	**0.4**
U55060	Lgals9									**0.2**	**0.2**	**0.3**	**0.3**
AW124890	Epb4.9									**0.2**			**0.3**
U76253	Itm2b			D	D				D		**0.4**		
X04653	Ly6a				D				D		**0.4**		
X66449	S100a6								D			**0.4**	
AA657044	2310047C17Rik								D		**0.4**		
AW049897	1110008K06Rik								D	**0.4**			
M77497	Cyp2f2				D					**0.3**			
AI852098	Elovl5				D					**0.4**			
X81323	Tpp2									**0.2**		**0.3**	
U03419	Col1a1									**0.4**		**0.4**	
AB023957	AB023957											**0.2**	
AA718169	Retn											**0.3**	
AI019679	1100001G20Rik											**0.3**	
AW121294	Arhu											**0.4**	
X52046	---											**0.4**	
U00978	Impdh1											**0.4**	
X04480	Igf1											**0.4**	
Z72486	Pold2										**0.3**		
AW125390	1110004C05Rik										**0.3**		
AB008180	Pcdha6									**0.3**			
U87948	Emp3									**0.3**			
M23362	Fgfr2									**0.3**			
D14572	Cbfb									**0.4**			
AI854103	4933436C10Rik									**0.4**			
M36579	S100a4									**0.4**			
X53928	Bgn									**0.4**			

TABLE 2.3

Features detected as significantly changed by other methods, but not by VAMPIRE in response to thiazolidinedione treatment. The columns on the right show several standard statistics for this data set. The column labeled ìCî lists the mean average-difference scores from MAS 5.0 of DMSO-treated 3T3-L1s. The rightmost columns show the mean and standard error of fold-change, computed from all pairwise combinations of TZD-treated to control measurements. Legend: C, vehicle control; P, pioglitazone; R, rosiglitazone; T, troglitazone; X, pooled TZD treatments.

		SAM				Cyber-T			
probe	**gene**	**P**	**R**	**T**	**X**	**P**	**R**	**T**	**X**
U85259	Esrra			U	U				U
L31777	---				U				U
AI854606	D14Wsu89e								U
AI846017	1810045K17Rik								U
U22519	---								U
L35599	---								U
U23095	---								U
X87685	---		U		U				
AI846102	Sla2				U				
AI848770	AI648866				U				
AI853918	2310061I04Rik				U				
AI843401	2300002G02Rik				U				
AI841400	1200015F23Rik				U				
AA408156	Psg19						U		
AW050255	Mlycd			U					
AI843119	Gsto1			U					
AB020239	Ak4			U					
M32599				U					
L13732	Slc11a1		U						
AV084684	---		U						
AA795062	Lsm3	U							
AI846851	Fdps	U							
AI851810	Clgn	U							
AV090583	---	U							
AI553024	---	U							
C85523	---	U							
Z38110	Pmp22				D				D
M18194	Fn1				D				D
D86232	Ly6c				D				D
AI843895	Srebf1				D				D
X04017	Sparc				D				D
K02782	C3								D
AA822296	Phka2		D		D				
AI836367	---				D				
Y15163	Cited2				D				
Z12171	Dlk1				D				
AF078705	Aoc3				D				
S38219	---				D				
AW125299	D14Wsu89e				D				
J03952	Gstm1				D				
AI836140	D15Ertd366e				D				
AA727291	2300003P22Rik				D				
AI837012	Sdbcag84				D				
X98471	Emp1				D				
X75129	Xdh				D				
AB030483	Eppb9				D				
AI843396	Gng10				D				
AI839697	Atp1a2				D				
U16818	Ugt1a1				D				
X04673	Adn				D				
AV355798	---					D			
L48687	Scn1b			D					
X13297	Acta2			D					
AF090334	Fkbp9			D					
Y15910	Diap2			D					
U94479	Ilk		D						
AJ006972	Tom1		D						
U05837	Hexa		D						
AA873956	5830404H04Rik	D							
AA824102	Hsd3b7	D							
AA960603	---	D							
AW123316	Mccc1	D							

May 19, 2005

Albert Hsiao has my permission to include the following paper, of which I was a co-author, in his doctoral dissertation.

Hsiao, Albert; Subramaniam, Shankar. ì Statistical interpretation of two-channel functional genomics data ñ the bivariate microarray analysisî , PLOS Computational Biology, 2005 (submitted).

_____

Shankar Subramaniam

This chapter, in full, has been submitted for publication in PLOS Computational Biology 2005, Hsiao, Albert; Subramaniam, Shankar., Oxford University Press, 2005. The dissertation author was the primary investigator and single author of this paper.

CHAPTER III

Statistical Interpretation of Two-Channel Data

A. Abstract

Conventional statistical methods for interpreting microarray data require large numbers of replicates in order to provide sufficient levels of sensitivity. These statistical tests require many samples ($n \geq 6$) to adequately assess variability for each array feature. However, the error structure of high-throughput microarray data can itself be a reliable measure of noise, and may be used to devise tests with improved accuracy.

We present here a novel approach for the analysis of microarray data that involves statistical modeling of paired, two-channel expression data in the context of the Bayesian framework. We further describe a novel statistical test that can be used to identify differentially-expressed genes. When large numbers of samples are available (n=6-8), our method, Bivariate Microarray Analysis (BMA), demonstrates comparable results with conventional paired statistical tests. When replicate numbers are reduced, conventional approaches typically suffer great losses in sensitivity for detecting differences in gene expression. In contrast, we show in validated microarray data that despite reduced replication, BMA retains substantial statistical power.

Although statistical interpretation of expression data conventionally requires larger replicate numbers ($n \geq 6$), it is possible with BMA to detect statistically significant changes in gene expression with minimal losses in specificity. Further, we show that combining results from BMA with Gene Ontology annotation yields consistent, reproducible results, even when only two replicates are used. The approach presented has been integrated into a web-based application, available at http://genome.ucsd.edu/microarray.

B. Introduction

Microarrays are invaluable tools for measuring transcriptional responses of cells and tissues. The most common and perhaps the most fundamental question typically asked is whether observed differences in gene expression are statistically significant. Traditionally, fold-change cutoffs have been used as indicators of significance. Although a fold-change can be a consistent measure at high signal intensities, it is not very reliable when the signals are low. Low intensities are therefore filtered out prior to applying a fold change threshold, leaving these expression changes poorly interpreted. Furthermore, it has been difficult to correlate these arbitrary thresholds with *a priori* indicators of statistical significance. While this approach has historically been used to produce meaningful biological results, more rigorous statistical techniques are needed to overcome these limitations.

In order to address the non-uniform variability of array data, a number of groups have systematically modeled the variance structure[5, 22, 24]. This variance structure can be decomposed into an expression-dependent and an expression-independent component. The approaches generally taken however, have been to normalize the microarray data to remove inhomogeneity, and process the resultant data with current statistical tests[6, 7, 26]. We have taken a fundamentally different approach to this problem. Recognizing that sample variance is a poor estimator of noise at low replicate numbers (n=2-3), we hypothesized that the variance structure itself could be used as a more precise estimator of variance[27]. We subsequently devised a modeling procedure to identify the maximum likelihood estimates of expression-dependent and expression-independent variance. This model was then incorporated directly into a Bayesian statistical test.

In addition to normalization strategies, several statistical methods have been described to interpret gene expression data. Non-parametric methods such as Significance Analysis of Microarrays (SAM)[20] are widely applicable because they do not rely on

explicit assumptions about the error structure. They can therefore be easily adapted to a variety of experimental designs without undertaking challenging mathematical derivations. These non-parametric methods, however, lack sensitivity in the absence of high levels of replication. On the other hand, parametric methods such as the *t*-test and Bayesian variants like Cyber-T[21] can be powerful because of their ability to extrapolate variability based on prior assumptions about the error structure. While parameterization limits broad applicability, these methods can reduce the number of samples needed to identify significant changes in gene expression. Current statistical methods however, still require many replicates to achieve sufficient sensitivity for biological interpretation.

We present a novel statistical approach for the interpretation of microarray data that can be performed on as few as two measurements per experimental condition (n=2). Our method, the BMA (Bivariate Microarray Analysis) is founded on the Bayesian framework developed in Hsiao *et al*, 2004[27] for the interpretation of one-channel array data. Our method effectively correlates conventional statistical thresholds with fold-change thresholds at each level of gene expression, by modeling the variance structure of each data set. In addition to modeling the relationship between signal intensity and variance, BMA also models the covariation between two color channels. By modeling covariation, we are able to overcome high spotting variability inherent in current two-channel platforms. We demonstrate first in simulated data that BMA dramatically increases sensitivity over other statistical approaches, particularly at low replicate numbers (n=2). This remains true independent of whether we choose an array-wide false-positive rate (α_{Bonf}) or a false discovery rate (FDR) as a significance threshold. In PDGF-treated T98 cells, we demonstrate that BMA is highly sensitive for gene expression differences even when only one dye-swap pair is applied (n=2). Lastly, we show that when BMA is applied to a time course study of lipopolysaccharide (LPS)-treated macrophages, the differential gene expression pattern and the Gene Ontology enrichment found with three

dye-swap pairs (n=6) is only marginally better than that detected by a single dye-swapped pair.

C. Statistics and Modeling

Bayesian framework

The mean of the signal intensities is typically used as a point estimate for gene expression, and the standard deviation is often used as an indicator of variability. Our method relies on the notion that a model that more accurately fits the error structure of microarray data will be a better estimator of variability. By integrating this variance model into an extension of our Bayesian framework, which is discussed more thoroughly in the supporting information, we transform the mean intensity into a probability density for gene expression. Instead of a single point estimate for gene expression, we obtain a probability density that describes the likely values for ì true expression levelî (μ). We can then apply statistical tests on this density to determine which genes are differentially-expressed between two experimental conditions. The foundation of our approach, then, lies at accurately modeling the amount of noise in microarray data.

The error structure of microarray data

Microarray gene expression data displays greater fractional error at low signal intensities than at high intensities. When the same RNA samples are hybridized on multiple chips, this relationship is often still present, indicating that some of this variation comes from the microarray platform itself. Platform-specific factors, such as the resolution of the microscope and dye system, fundamentally limit the reliability of low-intensity measurements. We refer to this type of noise as expression-independent variance (B). At greater signal intensities, this physical limit becomes less and less important. Here, expression-dependent variance (A) begins to dominate, but both kinds of error are small compared to the signal. When biological replicates are used, biological variability

can contribute to both components of variance. This simple linear error model has been widely used to explain the variance behavior in each single channel[5, 7, 22, 24, 27].

In a two-channel microarray, measurements are obtained from two RNA samples, each labeled with different dyes. These samples are subsequently hybridized with microarray probes on a single array[28]. Typically, one RNA sample represents a control and the other represents some kind of experimental treatment. In this system, we refer to each array as a single replicate (n=1). Parallel RNA samples from the control and experimental group can be subsequently dyed in reverse and hybridized to a second array, creating a dye-swap replicate. These may either be aliquots of the previous RNA samples to produce ìarrayî replicates or samples from parallel experiments to produce ìbiologicalî replicates. Either way, we refer to these two replicates as a dye-swap pair. Since probes are often spotted on the array surface, there can be considerable variability between arrays in the amount of probe at each location. This variation in ìprobe intensityî alters the intensity measurements for both color channels, and leads to covariation of the paired signals. We therefore modeled this covariation by introducing correlation coefficients for expression-dependent and expression-independent variance. The final variance model contains a total of six parameters. These parameters are related to the treatment variances (σ_1, σ_2) and correlation (ρ) through:

$$\sigma_1^2 = \mu_1^2 A_1 + B_1 \quad (1)$$

$$\sigma_2^2 = \mu_2^2 A_2 + B_2 \quad (2)$$

$$\rho\sigma_1\sigma_2 = \rho_A \mu_1 \mu_2 \sqrt{A_1 A_2} + \rho_B \sqrt{B_1 B_2} \quad (3)$$

where (A_1, A_2) are the expression-dependent variances, (B_1, B_2) are the expression-independent variances, ρA is the expression-dependent correlation, ρB is the expression-independent correlation, and μ_1, μ_2 are the ìtrueî expression level of each treatment condition.

Parameter estimation

Parameters for the variance model can be computed by Markov Chain Monte Carlo (MCMC) simulation of the bivariate normal density. This simulation computes the maximum likelihood estimates for each of the six variance parameters. Since the ìtrueî gene expression is not known prior to observing array measurements, the observed sample mean is used as a surrogate. This choice, although necessary, can be problematic at low intensities. With the sample mean as a surrogate for true expression, poor signal-to-noise ratios can cause underestimation of true variability. Therefore, we devised an estimation technique that identifies parameters which are stable against an expression-level cutoff.

Model parameters should be the same regardless of how much data is used to estimate them. Thus, we may progressively discard increasing amounts of low-intensity features to avoid the downward bias caused in areas of low signal-to-noise. In figure 3.1, we demonstrate the effect of these cutoffs on estimates of expression-independent error. This parameter has two regions that are relatively stable against successive quantile cutoffs. When smaller cutoffs are used, and less data is ignored, expression-independent error tends to be underestimated. When larger cutoffs are used, the parameters become unstable. In between, we identify a stable region of parameter estimates, which we later use in the statistical test. The details of the modeling procedure are further discussed in the supporting information. It is also important to note that the expression-level cutoff defined here is only used to improve the quality of parameter estimates. The resulting variance model is applied across the entire array. By combining this model with the Bayesian framework described earlier, we can use a statistical test to determine the significance of observed differences in gene expression.

Statistical test

In order to make BMA compatible with conventional concepts widely understood by the biological research community, we devised a statistical test that computes a *p*-value from the posterior densities. This p-value may be used to control type I error, which is the rate at which the null hypothesis is incorrectly rejected. In BMA, we define the null hypothesis (H_0) to be that the ìtrueî gene expression does not change across two experimental conditions. The alternative hypothesis (H_1) is accepted when the null hypothesis is rejected. In other words, we test

$$H_0: |\mu_2 - \mu_1| = 0 \text{ against } H_1: |\mu_2 - \mu_1| \neq 0 \qquad (4)$$

where μ_1 and μ_2 represent the ìtrueî expression of the two treatment conditions.

The *p*-value of BMA is defined in a way similar to that of a *t*-test. In a *t*-test, the *p*-value is defined as an integral of a *t*-distribution. When the ìtail areaî of the *t*-distribution is sufficiently small, we consider experimental results to be statistically significant. In BMA, we define the *p*-value as a two-dimensional integral of a bivariate normal density. In other words, we determine whether the ìtail volumeî of this distribution is sufficiently small. If so, we consider the gene to be differentially-expressed. A depiction of the bivariate integral is shown in figure 3.2. Explicitly,

$$p_i = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\mu_2} \pi_i(\mu_1,\mu_2)\, d\mu_1 \cdot d\mu_2, & E[\mu_1] > E[\mu_2] \\ \int_{-\infty}^{\infty}\int_{\mu_2}^{\infty} \pi_i(\mu_1,\mu_2)\, d\mu_1 \cdot d\mu_2, & E[\mu_2] > E[\mu_1] \end{cases} \qquad (5)$$

where π_i is the joint posterior density for the ìtrueî expression levels of the i'th feature.

A solution to the integration of the bivariate normal is given in the supporting information. When the p-value is less than the significance threshold, we reject the null hypothesis that $\mu_1=\mu_2$, and consider the difference in gene expression to be statistically

significant. The results of this test, when applied to one time point in LPS-treated macrophage data, are shown in figure 3.3.

Specifying a significance threshold

There are two commonly used methods for determining a significance threshold (α) when analyzing microarray data. Significance thresholds must be corrected for the multiple parallel tests that are being simultaneously performed. If we are interested in maintaining an array-wide false positive rate, we may set a Bonferroni-corrected threshold (αBonf). This is equivalent to the number of features we expect to appear as significant by random chance alone. If we set α_{Bonf}=0.05, then we expect only 1 feature to be significant when analyzing 20 batches of microarray experiments. Needless to say, this is a tremendously strict significance threshold, which would minimize false-positives. When the number of features is large, the *p*-value threshold for a two-sided test is

$$\alpha = \alpha_{Bonf} / 2n , \qquad (6)$$

where *n* is the number of array features.

In some cases, the false discovery rate (*FDR*) may be preferred. It is typically less stringent, but can still provide meaningful results. Assuming that the *p*-value defined here gives a reasonable estimate for the type I error, the false discovery rate is related to the array-wide false positive rate by

$$FDR = \frac{2n \cdot \alpha}{i} = \frac{\alpha_{Bonf}}{i} , \qquad (7)$$

where *i* is the number of significant features.

D. Results

Quantitative assessment

Statistical tests for interpreting microarray data are often difficult to compare, particularly because complete lists of ìrealî differences in gene expression are often not well-characterized. It is therefore necessary to assess the accuracy of these methods by performing tests on data sets, for which all expression differences are known. In such simulations, it is important to produce test data that reflects the behavior of real data with as much accuracy as possible. We therefore chose to test the accuracy of several statistical methods on simulated data sets derived from a two-color data set recently published by Rome *et al*[29]. In their work, the authors analyzed the RNA content of human skeletal muscle biopsies collected before and after application of a hyperinsulinemic-euglycemic clamp. Sample measurements for our simulated set were obtained from lognormal distributions centered on the pre-clamp mean intensities. The ìtreatedî experimental condition was obtained by ìspikingî 10% of the features. The sample means of these randomly selected ìspikedî features were multiplied or divided by a uniform random number between 1 and 20. Since the signal-to-noise ratio (SNR) is much greater at higher signal intensities, it is likely that subtle changes will be more readily picked up at high levels of gene expression. This model problem therefore contains both subtle and dramatic gene expression changes at a variety of intensity levels.

Prior to identifying the ìspikedî features, we tested the accuracy of two different modeling procedures. An equivalent model for the variance structure of two-color microarray data was previously proposed by Ideker *et al* [24]. We applied their modeling approach, VERA (Variability and Error Assessment) along with our own, to compare parameter accuracy (Fig. 3.4). Even with two replicates, BMA showed striking accuracy in its estimate of expression-independent variance. VERA tended to underestimate this parameter even with large replicate numbers (n=16). We attribute the accuracy of BMA

to the cutoff procedure, which reduces the effect of poor-quality measurements at low intensities. Since expression-independent variance dominates at low intensities, this precision is essential for interpreting faintly-expressed genes. In addition, both approaches had a tendency to underestimate expression-dependent variance when only two replicates were used, but improved with more replicates.

With these variance parameters, we then compared the accuracy of four different methods in identifying ìdifferentially-expressedî genes. This comparison was performed using three different thresholds of statistical significance. The quality of the predictions is displayed in table 3.1. Each of the statistical tests demonstrates improved sensitivity as significance thresholds were loosened and as larger numbers of experimental replicates were used. As one would expect, the FDR significance threshold was often predictive of the true false-positive rate. There is only one clear case where the false-positive rate substantially exceeded the desired FDR. At n=2 and a relatively loose FDR of 0.05, BMA shows a mild loss in specificity. This is likely due to an underestimation of expression-dependent variance at n=2. This loss of specificity however, was not observed with α_{Bonf}=0.05 or FDR=0.001. Therefore, as long as stringent significance thresholds are applied when replicate numbers are few, our method will likely improve sensitivity without sacrificing specificity.

These results show that BMA is likely to improve prediction of differentially-expressed genes. The variance-modeled approach is particularly beneficial at low replicate numbers where other statistical methods fail. In these conditions, the variance model estimates variability more accurately than traditional statistics. Applying this model in a Bayesian framework translates into improved sensitivity for subtle changes in gene expression. For further validation of these observations, we applied BMA to two biological data sets.

PDGF-treatment of T98G glioblastoma cells

Tullai *et al* recently published a study of T98G cells treated with platelet-derived growth factor (PDGF) for 30 min after 72 hour serum starvation[30]. The data set published under the Gene Expression Omnibus (GEO) accession GDS896 consists of 8 microarrays, each hybridized with an RNA sample from (a) PDGF-treated T98G cells and (b) their control counterparts. Their experiment was dye-swapped, resulting in 4 dye-swapped pairs. BMA, SAM, and Cyber-T were each used to identify differentially-regulated features at two different significance thresholds. Since Bonferroni multiple-testing correction is not available for SAM, we first applied a shared false discovery rate (*FDR*=0.005). All methods showed strong agreement at this level of replication (Fig. 3.5a), with few genes detected by any single method alone. In fact, over 90% of features were detected by at least two methods. In this particular data set, all features detected by Cyber-T were also detected by BMA, and over 85% of features detected by SAM were also detected by BMA. When we applied more stringent Bonferroni-corrected thresholds, we achieved similar results (Fig. 3.5b). This highlights the comparable sensitivity and specificity of these methods, when relatively large numbers of replicates are available.

While each method identified approximately 100 array features with all four dye-swap pairs, they showed considerable differences with smaller sample sizes. With individual dye-swap pairs, other methods require substantial adjustments to their significance thresholds to detect changes in gene expression. However, even at fairly loose false discovery rates, SAM could no longer identify differentially-expression with so little data (n=2, *FDR*=0.05). Cyber-T found between 4 and 8 features (α_{Bonf}=0.05), only one of which was mutually detected in all dye-swapped pairs (FOS). In contrast, BMA identified 40-70 features in each dye-swap pair analysis (α_{Bonf}=0.05). There was considerable agreement between each of these analyses, sharing a total of 30 significant changes (Fig. 3.5c). In addition, less than one-third of features identified were not otherwise detected by

the ìglobalî analysis of all four pairs. These results suggest that with BMA, smaller sample sizes may still be valuable for detecting reproducible gene expression changes at fixed significance thresholds.

When we examined individual analyses for enrichment in Gene Ontology annotations, each of the dye-swap pair analyses identified *MAP kinase phosphatase activity* as the most substantially enriched (all statistically significant at α_{Bonf}=0.05), and found features for DUSP1 (MKP-1) and DUSP5 to be upregulated by PDGF. MKP-1 has been previously noted to be transiently upregulated by PDGF[31]. Along with other dual-specificity phosphatases, it is believed that MKP-1 feedback modulates MAP kinase signaling components and growth factor signaling. DUSP5 has not been as extensively studied, but shares 45.2% sequence identity with DUSP1. Analysis of all three dye-swap pairs additionally revealed upregulation of DUSP16 (MKP-7). MKP-7 seems to preferentially dephosphorylate JNK and p38[32]. Although the precise mechanisms and interactions of these dual-specificity phosphatases are not yet clear, the consistent enrichment of this Gene Ontology term suggests that these genes are truly PDGF-regulated, and may play an essential role in PDGF action. More importantly for our analysis however, these results demonstrate that paired analysis by BMA can be extremely valuable, even when the number of microarray samples is very small. Gene Ontology post-analysis points to the same functional categories whether a single dye-swap pair is used or all four are used. While other statistical methods typically fail in this domain, BMA retains substantial sensitivity for real differences in gene expression with relative preservation of specificity.

Time-course study of LPS-treated macrophages

We also examined the effectiveness of several methods on two-channel microarray data available from the Alliance for Cellular Signaling (AfCS). [Reference: www.signaling-gateway.org]. In this data set, three pairs of dye-swapped measurements

were obtained for each of six time points during lipopolysaccharide (LPS) treatment of RAW 264.7 macrophages, for a total of 36 arrays. Data sets such as this are valuable for understanding the dynamics of gene expression in response to ligand. In addition, much is already known about the behavior of macrophages in the presence of LPS. This is an ideal system then, to determine whether it is possible to interpret gene expression responses while using fewer replicates.

When all three dye-swap replicates are available, similar sets of differentially-expressed genes can be obtained by any of the previously described methods. At the 8 hour time point, for example, 1492 genes were detected as differentially-expressed by BMA, SAM, and Cyber-T (FDR=0.001). In contrast, only 300 were detected solely by BMA, 279 solely by SAM, and 105 solely by Cyber-T. The results differ tremendously with fewer replicates, however. When SAM is given fewer than three dye-swap pairs, no delta can be found that satisfies the desired false-discovery rate. With Cyber-T, similar to what is observed in simulated data, far fewer numbers of significant gene expression changes are detected when either two dye-swap pairs (n=4) a single dye-swap pair (n=2) are used (Fig. 3.6). In contrast, our method detects similar numbers of gene changes regardless of how many replicates are used in analysis. This is further confirmed by comparing the sets of differentially expressed features (Fig. 3.7, 3.8). Despite using only the information contained in a single dye-swap pair, BMA identifies equivalent sets of significant changes. Furthermore, we found that similar Gene Ontology (GO) terms were statistically enriched whether we used a single dye pair or all of the available data (Table 3.2). In particular, genes involved in cell death and proliferation were strongly regulated by LPS. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were also similarly enriched (Table 3.5). Therefore, although increased sensitivity can be gained by increasing replicate number, BMA allows us to make consistent biological interpretations

of two-channel functional genomics data with as little as a single dye-swap pair, where previous methods required many more.

We further examined the 985 features that were *uniquely* found by BMA with only two replicates at the 8 hour time point (α_{Bonf}=0.05). Most of these were previously detected by Cyber-T with all six replicates. Enriched among these, were 17 features associated with components of the toll-like receptor signaling pathway as defined by KEGG. This list includes prominent figures such as toll-like receptor 2 (Tlr2) and toll-like receptor 4 (Tlr4). Tlr4 is believed to be required for LPS-induced signaling[33], while expression of Tlr2 and Tlr4 have previously been confirmed to be induced in murine alveolar macrophages by LPS[34]. Furthermore, it has long been known that LPS stimulation induces macrophage expansion *in vivo*[35]. This effect on gene expression is striking at the 8 hour time point, as BMA found 115 additional features involved in cell proliferation. As macrophages play a major role in antigen presentation of exogenous peptides, it is also interesting to note that seven additional features involved in MHC class I and one additional feature involved in MHC class II antigen presentation were also upregulated. This included several HLA loci and β_2-microglobulin. Class II MHCs are commonly thought to present exogenous antigens, but recent evidence has shown that immunity against *Mycobacterium tuberculosis* requires presentation on class I MHCs in a ëdetourí pathway [36]. The consistency of these results with known macrophage physiology is strongly suggestive of the quality of these genes detected by BMA. These 985 features would have been entirely missed by other methods if only two replicates were used. While previous statistical methods were capable of identifying these changes with additional data, we have shown that additional replication is not always necessary with BMA.

E. Discussion

We demonstrated here the effectiveness of the variance-modeled Bayesian approach on paired microarray data. The present method provides a reliable tool to identifying a set of differentially-expressed features with an *a priori*-specified significance threshold, and has a deep relationship to the commonly-used fold-change method. While most microarray users continue to use a fold cutoff with an expression minimum as the *de facto* standard, the fold-change method by itself lacks statistical rigor. Each platform and set of biological data may have considerably different patterns of noise, which makes it difficult to identify appropriate thresholds *a priori*. It has also previously been difficult to correlate these to multiple-testing thresholds used by the statistics community. BMA addresses these issues by connecting these opposing points of view. Each statistical threshold can be connected to a fold-change threshold at each level of gene expression. BMA requires only the most minimal information needed to interpret two-channel microarray data. Given measurements from only four RNA samples hybridized on two, two-channel arrays (dye-swapped), BMA identifies differentially-expressed microarray features with surprising accuracy. This is even more remarkable given that feature-specific variability is only assessed in the context of the global model. If an individual feature experiences an exceptional amount of noise, it still only acts as one voice among many in determining the local amount of error. The results we present clearly demonstrate improved detection in both simulated and real data. Thus, although conventional methods focus heavily on feature-specific variability, the global variance approach may be more effective for low-replicate, high-throughput data.

In our analysis of present statistical methods, BMA showed improved sensitivity at the significance thresholds and replicate numbers tested. As long as stringent significance thresholds were applied, specificity was well-preserved. In LPS-treated macrophage data, the significant features detected with three dye-swap replicates were not substantially

different from the features detected with a single dye-swap replicate. We observed similar, albeit not as extreme results when examining PDGF-treated T98G cells. In addition, since no additional experiments need to be performed to compute this threshold, our method is immediately applicable to existing data sets. It is important however, to interpret the analysis of the macrophage data with caution. It is not likely that all data sets will require so few replicates to obtain such resolution of transcriptional changes. The response of macrophages to endotoxin is an ancient and extremely robust response fundamental to innate immunity. Transcriptional responses to most other stimuli are likely to be more subtle and require larger numbers of biological replicates to achieve maximal resolution. For example, in PDGF-treated T98G cells, we observed considerable increases in sensitivity with additional replicate data. Nevertheless, for situations where such resolution is not required and resources are limiting, BMA can be an effective tool for accurately identifying transcriptional changes.

BMA also appears to preserve biological interpretation. When we examined a data set published by Tullai, et al consisting of PDGF-treated T98G cells, analysis of each individual dye-swap pair was able to identify significant enrichment of *MAP kinase phosphatase activity*. In fact, this was the single most-enriched Gene Ontology term in every BMA analysis we performed, and in some cases, the *only* statistically enriched term. We suggested that dual-specificity phosphatases may be essential to understanding the transient transcriptional responses of PDGF. Further studies will be required to detail the roles of these phosphatases. It is worth noting that this Gene Ontology term was also the most enriched in the analysis of all four dye-swap pairs. Similar results were obtained from enrichment analysis of LPS-treated macrophages. In both analyses, added replication incrementally improved the most enriched Gene Ontology terms by adding additional genes that did not reach statistical significance at lower replicate numbers. The

top-enriched terms remained the same. Thus, with BMA, large numbers of replicates are not always necessary for meaningful biological interpretation.

Microarray expression profile studies have, up until now, been limited by the costs of producing sufficient numbers of arrays to accommodate present statistical methods. BMA approaches this fundamental limitation by modeling the relationships between variability and gene expression, and applying this array-wide model as a more accurate indicator of error. It is a limiting assumption that is most appropriate with large numbers of probes (10000+). While hypothetically, gene-specific or feature-specific variability could introduce additional false positives and false negatives, we demonstrated that BMA still produces high-quality results in two data sets where this might be an issue. Transient 30 minute stimulation of T98G cells with PDGF and a time course of endotoxin-treated macrophages represent a fair spectrum of biological responses typically studied by expression profiling. In both cases, we have demonstrated that BMA is at least as reliable as other statistical methods at larger replicate numbers (n=6-8). Thus, in these data sets, feature-specific variations do not appear to substantially increase the false-positive rate of our test. As microarray manufacturers increase feature density, probe consistency, and apply multiple probe spots to produce more robust indicators of gene expression, platform-specific contributors to feature-specific variability may further diminish.

Although some previous attempts at modeling the sources of noise have been described[5, 22, 24], much of the literature is devoted to normalizing array data to reduce variability. While these forms of normalization are not necessarily incompatible with BMA, we believe they can introduce additional artifacts, particularly when underlying assumptions are too strong. For example, quantile and lowess normalization fail to account for the poor signal-to-noise ratios at low intensities. Artifacts can result from over-fitting of noisy, low-intensity data. When we modeled variance parameters without these forms of normalization (supplemental information), we found that the expression-

independent correlation coefficient (ρ_B) converged on a more realistic value, and was vastly more stable against quantile cutoffs. This suggests to us that the underlying variance structure at low intensities can become clouded by these forms of normalization. While other statistical tests will continue to require these techniques to optimize sensitivity, caution should be exercised in using these normalization procedures prior to applying our method.

Unlike non-parametric statistical tests like SAM, parametric tests like the *t*-test and BMA can only be applied when their parametric assumptions are satisfied. In general, additional parameterization increases sensitivity, at the cost of broad applicability. In the case of a *t*-test, the assumption is that repeated sample measurements are distributed normally. In BMA, we assume that the measurement noise for a given microarray probe can be estimated from a global paired variance model. All of these tests can be used to directly compare expression measurements obtained from different muscle biopsies, albeit with variable degrees of success. For data that is further processed, by log-transformation or certain forms normalization, only non-parametric tests can be immediately applied. Parametric tests must be appropriately modified with new assumptions regarding the underlying data. Thus, BMA achieves its improved accuracy by confining itself to a limited subset of statistical problems. Additional statistical tests derived from this framework may achieve similar improvements over conventional, more generalizable, statistical tests.

We believe that conventional statistical tests will ultimately be limited by the levels of replication needed to demonstrate statistical significance. Sample numbers demanded by these methods do not frequently exist because of financial and labor constraints. They are even less likely to exist in multiple-ligand and time-course studies. While large sample sets are probably still necessary to interpret data collected from diverse populations, we demonstrate here that with BMA, meaningful biological interpretations can still be made

from smaller sample sets. BMA takes advantage of the parallel nature of microarrays to model variability and reduce replicate number requirements. Two channel arrays have tremendous untapped potential for large-scale studies of functional genomics. We anticipate that because of BMAís improved sensitivity at low replicate numbers, it will provide (1) more consistent interpretations of paired array data and (2) an avenue for implementing more sophisticated experimental designs at greatly reduced cost.

F. Methods

In order to assess the effectiveness of several current statistical approaches, we created a simulated data set based on the error structure of the previously published data set of Rome *et al*, 2003[29]. We computed variance parameters from this data set and used this model to generate random paired-samplings from a lognormal density centered on the sample means. The simulated data set contains 20000 features. 10% of these features were randomly chosen to be spiked by multiplying the sample mean by e^s, where s is a sampling from a uniform random variate with support on $[-\log_e(20), \log_e(20)]$.

In our analysis with VERA and SAM[24], we used a likelihood ratio cutoff of λ_c=23.8 for statistical significance, which is approximately equivalent to a false-discovery rate of 0.001. For Cyber-T, we performed paired analysis with ν=10, and detected significance in two ways: using α_{Bonf}=0.05 and manually adjusting α_{Bonf} to achieve the desired false-discovery rates of 0.001 and 0.05. False-discovery rates in SAM were adjusted by adjusting Δ until the desired FDR were achieved. For BMA, we implemented an automated procedure that identifies the appropriate α_{Bonf} for the desired FDR, based on golden section root finding. A lower bound for α_{Bonf} was set at the minimum achievable FDR. The range of α_{Bonf} was subsequently adjusted using equation (7) until α_{Bonf} could be narrowed to within 10^{-5}. To determine the minimum achievable FDR, we similarly used a golden section minimization strategy.

To demonstrate the effectiveness of each method on biological data, we obtained two data sets. The PDGF-treated T98G cell data set was obtained from Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/, signals were background subtracted, and were normalized by setting the median intensities of each array to 1000. Array features marked as ìbadî in the raw data were masked, while control features were only masked prior during modeling. No additional masking of low-intensity measurements was performed. The AfCS LPS-treated RAW 264.7 Agilent inkjet-deposited oligo data set was obtained from the AfCS web site at http://www.signaling-gateway.org/. In order to give SAM and Cyber-T a fair chance, the processed signal intensities were normalized with intensity-dependent lowess normalization to standardize signals between arrays, although similar results were obtained by BMA without this normalization. In order to identify a stable set of variance parameters in LPS-treated data, a single variance model was obtained from a pooled data set containing all six time points. This model was then applied to all six time points to determine differential expression. This was step was necessary to overcome information loss caused by lowess normalization.

Statistical enrichment of Gene Ontology terms and KEGG pathways was computed by comparing the number of ìsignificantî features annotated with a particular GO term to the background. For the background, we used all of the features on the Agilent array. Exact likelihoods (*p*-values) were computed directly from the hypergeometric distribution.

The modeling procedure and the bivariate microarray analysis were originally implemented as a set of command-line Java tools. These tools have since been integrated into the VAMPIRE microarray analysis suite, which is available at http://genome.ucsd.edu/microarray.

G. Acknowledgements

We thank Yi-Xiong Zhou and Trey Ideker for our valuable discussions, and Wendy Ching, Dorothy Sears, Jason Papin, John Mongan, Jeff Gold, and Paul Insel for their comments on the manuscript. This work was supported by a grant from the National Institutes of Health, NIGMS K54 GM62114, a Life Sciences Informatics grant from the State of California and Pfizer La Jolla, and a grant from the Hillblom Foundation**.** AH is a graduate student in the UCSD Medical Scientist Training Program and is supported by a Fellowship from the Whitaker Foundation and the UCSD MSTP Training Grant T35-GM07198.

H. Statistical Framework

The Bayesian Model

In Hsiao, *et al* (2004), we previously described the univariate conjugate-exponential Bayesian model, which we may extend to the multidimensional case. The multivariate likelihood function may be written as:

$$f(\vec{x} \mid \vec{\theta}) = a(\vec{x}) \exp\left[\vec{x}^T \vec{\phi}(\vec{\theta}) - \psi(\vec{\theta})\right] \qquad (1)$$

where

$\vec{x} = (x_1 \quad x_2 \quad ...)^T$ is a vector of data,

$\vec{\theta} = (\theta_1 \quad \theta_2 \quad ...)^T$ is a vector of parameters of the function,

$a(\vec{x})$, $\vec{\phi}(\vec{\theta})$, $\psi(\vec{\theta})$ characterize the shape of the likelihood function.

In our paired, two-channel array problem, $\vec{x}$ represents a pair of measurements from a single replicate of one array feature and $\vec{\theta}$ is the ìtrue gene expressionî for both treatment conditions.

With the likelihood function in this form, the exponential prior may be expressed as:

$$P(\vec{\theta}) \propto \exp\left[\vec{\alpha}^T \vec{\phi}(\vec{\theta}) - \beta\psi(\vec{\theta})\right] \quad (2)$$

where

$\vec{\alpha} = (\alpha_1 \quad \alpha_2 \quad ...)^T$ and β characterize the shape of the prior

Combining the multivariate prior and the likelihood function, it can be easily shown that the multivariate posterior therefore takes a similar form:

$$\pi(\vec{\theta} \mid X) \propto \exp\left[\vec{\alpha}_m{}^T \vec{\phi}(\vec{\theta}) - \beta_m\psi(\vec{\theta})\right] \quad (3)$$

where

X is a matrix of data for a given feature,

$\vec{\alpha}_m = m\bar{x} + \vec{\alpha}$,

$\beta_m = m + \beta$

$\bar{x} = (\bar{x}_1 \quad \bar{x}_2 \quad ...)^T$

The Likelihood Function

Given that we have chosen the univariate normal density as our likelihood function in the single-channel case, the bivariate normal likelihood density is the logical choice for two-channel arrays. We can carry this extrapolation further, using the multivariate normal for multiple-channels, and simply use matrix notation to express the relevant equation.

For the random vector, $\vec{x} \sim N(\vec{\mu}, \Sigma)$, the multivariate normal may be expressed as

$$f(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \quad (4)$$

The covariance matrix Σ is computed by maximum likelihood MCMC simulation of the bivariate normal. This procedure was similarly performed in Ideker, et al 2000. The remaining undetermined parameters of the likelihood function are thus

$\vec{\theta} = \vec{\mu} = (\mu_1 \quad \mu_2 \quad ...)^T$

Rewriting the multivariate likelihood in the form of eq. 1,

$$a(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{x}\right)$$

$$\vec{\phi}(\vec{\theta}) = \Sigma^{-1}\vec{\mu}$$

$$\psi(\vec{\theta}) = \frac{1}{2}\vec{\mu}^T \Sigma^{-1}\vec{\mu}$$

The Prior

There are several ways to parameterize of the multivariate prior density, but we choose first to apply the simplest approach. Let the prior densities for the μ_i of each treatment condition be independent and let $\vec{\lambda} = (\lambda_1 \quad \lambda_2 \quad ...)$ be the vector of hyperparameters. Our exponential prior then takes the form

$$P(\vec{\theta}) \propto \exp(-\vec{\mu}^T \vec{\lambda}) \qquad (5)$$

Rewriting the prior in the form of eq. 2,

$$\vec{\alpha} = -\Sigma\vec{\lambda},$$

$$\beta = 0$$

By setting the vector $\vec{\lambda} = \vec{0}$, the prior becomes non-informative. Fully Bayesian treatment of this statistical model involves modeling of the hyperparameter $\vec{\lambda}$, but we have observed that typical estimated values for this hyperparameter do not substantially change the identities of significantly changed genes. Since the computational cost is tremendous, we have chosen not to implement regional exponential priors in BMA.

The Posterior

In previous work, we demonstrated that given our choices for our prior and likelihood function, we could obtain a normally distributed posterior. We may do the same with the multivariate version of the Bayesian model. The parameters of the posterior are

$$\vec{\alpha}_m = m\bar{x} - \Sigma\vec{\lambda},$$

$$\beta_m = m$$

It is easy to demonstrate that the posterior density is normal with the following moments:

$$E(\vec{\theta}) = \frac{\alpha_m}{\beta_m} = \bar{x} - \frac{\Sigma\vec{\lambda}}{m} \quad Var(\vec{\theta}) = \frac{\Sigma}{\beta_m} = \frac{\Sigma}{m}$$

Variance model parameter estimation

The procedure that we have implemented to compute variance model parameters involves identifying a region on a two-dimensional grid where parameter estimates are stable against expression-quantile cutoffs. At extremely low expression levels, where sample means lie below the expression-level independent variance (B), sample means are extremely inaccurate as indicators of true ìexpressionî. Furthermore, sample variances at these levels of gene expression underestimate true variability. These data points therefore introduce downward bias in B. To overcome this obstacle, we progressively eliminate the lowest quantiles from a training data set that is used to fit the variance model. When model parameters are sufficiently stable, where estimates of expression-independent error ($\sqrt{\hat{B}_1}$, $\sqrt{\hat{B}_2}$) stay within 10%, we assume that they reflect the ìtrueî parameters of the variance model.

We have implemented a variety of approaches for estimating model parameters, which rely on the concepts described above. Each produce similar results on simulated data, but can vary in real data, depending on the quality of the signal and the type of normalization used. In our experience, the widest stable intervals are obtained when minimal amounts of normalization are used. Parameter estimates can either be computed by linear regression or Markov Chain Monte Carlo (MCMC) simulation of the likelihood function. MCMC simulation is frequently much more stable, but suffers from tremendous computational cost. Thus, it is frequently impractical to compute the entire two-dimensional grid using MCMC without sacrificing grid resolution or without reducing the number of MCMC iterations. Alternatively, parameter estimates can be performed on

each color channel independently. In many cases, sampling only the *x* and *y*-axes of the grid is sufficient for finding stable parameters.

The stability of each variance parameter is displayed in Fig. 3.8. In these diagrams, MCMC simulation was performed for each pair of quantile cutoffs. The broadest stable region for sqrt(B_1) and sqrt(B_2) lies in the control condition quantile cutoffs of 0.6 to 0.8. As these parameters are typically the most sensitive to cutoff, we use their stability as an indicator of the stability of the entire model. It is apparent that expression-dependent variance parameters (A_1, A_2) are also stable in these regions, as well as their associated correlation coefficient (ρ_A). It is not immediately clear to us why the expression-independent correlation ρ_B has a tendency to approach 1 in lowess-normalized data, but it does not appear to have a severe effect on the results.

As shown in Fig. 3.9, similar results can be obtained from median-normalized data, but the variance parameters are typically much more stable. In addition, the expression-independent correlation (ρ_B) converges on a value less than 1. In our hands, slightly fewer differentially-expressed genes are found without additional normalization. Despite this, additional normalization should be used with caution with BMA for two reasons: (1) there is a dramatic loss of parameter stability in lowess-normalized data, and (2) we believe that a fundamental assumption underlying lowess-normalization may be too harsh, as it fails to account for poor signal-to-noise at low signal intensities. Signals at low intensities tend to get distorted by the normalization process, and information about the level of background noise is lost.

Half-space integration of the bivariate normal density

Problem

Once the parameters of the two-color variance model are determined, the statistical significance of gene expression differences on two-color cDNA arrays may be converted into an integration of a bivariate normal density function. Assuming that there is no bias between measurements of the two treatment conditions, we are interested in whether the joint posterior density lies sufficiently distant from the diagonal line $\mu_1=\mu_2$. Geometrically, we can understand this as a diagonal cut through the density function in three-dimensional space. If the volume of the density function on one side of the cut is smaller than our significance threshold, then we may consider this a statistically significant change in gene expression.

We previously defined the *p*-value for statistical significance of changes in the *i*'th feature as

$$p_i = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\mu_2} \pi_i(\mu_1,\mu_2)\,d\mu_1 \cdot d\mu_2, & E[\mu_1] > E[\mu_2] \\ \int_{-\infty}^{\infty}\int_{\mu_2}^{\infty} \pi_i(\mu_1,\mu_2)\,d\mu_1 \cdot d\mu_2, & E[\mu_2] > E[\mu_1] \end{cases} \qquad (2)$$

where the joint posterior density for the gene expression of the *i*'th feature is $\pi_i(\mu_1,\mu_2)$. By our choice of a Bayesian conjugate-exponential hierarchical model, the joint posterior densities are bivariate normal.

The multivariate normal density function for the random vector $\vec{x} \sim N(\vec{\mu},\Sigma)$ is given by the expression,

$$f(\vec{x} \mid \vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})\right) \qquad (1)$$

where

$\vec{\mu} = (\mu_1 \quad \mu_2 \quad ...)^T$ is the vector of means,

$\Sigma = \begin{pmatrix} \sigma_1^{\,2} & \sigma_{12} \\ \sigma_{12} & \sigma_2^{\,2} \end{pmatrix}$ is the symmetric variance matrix, and

n is the dimensionality of the random vector.

Solution

The solution to this integral involves a series of careful transformations that preserve lines in space. By preserving continuity of lines, the half-spaces on each side of the lines also remain contiguous. We transform the arbitrary bivariate normal density (figure 3.10) to a bivariate unit normal, while simultaneously transforming our integration boundaries until we arrive at an expression that is more easily integrated.

Step 1: Transformation of the integral to the bivariate unit normal.

Let $\vec{z}(\vec{x}) = \sqrt{\Sigma^{-1}}(\vec{x} - \vec{\mu})$

The matrix $\sqrt{\Sigma^{-1}}$ that obeys $\sqrt{\Sigma^{-1}} \cdot \sqrt{\Sigma^{-1}} = \Sigma^{-1}$ can always be obtained from a symmetric, diagonalizable matrix Σ^{-1}. It can be obtained from the equation $\sqrt{\Sigma^{-1}} = V\sqrt{D}V^T$, where $\sqrt{D}$ is the diagonal matrix of square roots of eigenvalues, and V is the orthogonal matrix of eigenvectors. Furthermore, $\sqrt{\Sigma^{-1}}$ is itself symmetric.

The half-space integral is

$$F(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \iint \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) d\vec{x} \,. \qquad (2)$$

Substituting for Σ^{-1},

$$F(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \iint \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \sqrt{\Sigma^{-1}}^T \sqrt{\Sigma^{-1}} (\vec{x} - \vec{\mu}) \right) d\vec{x} \qquad (3)$$

Applying the transformation, the distribution function becomes bivariate unit normal.

$$F(\vec{z}) = \frac{1}{\sqrt{(2\pi)^n}} \iint \exp\left(-\frac{1}{2}\vec{z}^T\vec{z}\right) d\vec{z} \tag{4}$$

Step 2: Transformation of the integral boundaries.

The integral boundaries must also be transformed. The only boundary that we will be concerned with is the boundary $x_2=x_1$. The integration boundary $x_2=x_1$ may be expressed as the line passing through the origin with direction given by the unit vector

$$\vec{b} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T .$$

The transformation that we previously chose, $\vec{z}(\vec{x}) = \sqrt{\Sigma^{-1}}(\vec{x} - \vec{\mu})$, belongs to a class of transformations called affine transformations which preserve lines in space. This transformation may be expressed as a combination of a linear transformation and a simple translation.

$$\vec{z}(\vec{x}) = A\vec{x} + \vec{\delta} \tag{5a}$$

where

$$A = \sqrt{\Sigma^{-1}} \tag{5b}$$

$$\vec{\delta} = -\sqrt{\Sigma^{-1}} \cdot \vec{\mu} \tag{5c}$$

Since linear transformations preserve the origin, the integration boundary after transformation by A may again be expressed as the simple direction vector $\vec{c} = A\vec{b}$. After translation however, the origin is no longer preserved. The integration boundary is then given by the line passing through $\vec{\delta}$, with the direction vector $\vec{c}$.

Step 3: Rotation of the bivariate unit normal.

We can take advantage of the symmetry of the bivariate unit normal at the origin to perform one final transformation. We will rotate the coordinate system until the integration boundary lies perpendicular to the x-axis. Since rotation matrices are orthogonal, the preserve distance and will not affect the value of the integral. The clockwise-rotation of the axes (counter-clockwise rotation of two-dimensional space), may be written as

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{6}$$

The angle of rotation, θ, can be obtained from the direction vector $\vec{c}$ by

$$\theta = \tan^{-1}\left(\frac{c_1}{c_2}\right) \tag{7}$$

With this rotation, we can find the x-intercept of the rotated boundary by determining the minimum distance from a point on the unrotated boundary to the origin. The distance to the origin is given by the equation

$$d = \frac{|b|}{\sqrt{m^2 + 1}} \tag{8}$$

where

$m = \frac{c_2}{c_1}$ is the slope and

$b = \delta_2 - m \cdot \delta_1$ is the y-intercept of the unrotated boundary.

The x-intercept of the rotated boundary is located on the x-axis with the same distance to the origin. When the y-intercept of the unrotated boundary is greater than zero (Fig. 3.10), the rotated x-intercept is located on the *negative* x-axis. Similarly, when the y-

intercept of the unrotated boundary is less than zero, the rotated x-intercept is on the *positive* x-axis.

Step 4: Integration.

The final integral thus takes the form

$$p_i = \begin{cases} \frac{1}{(2\pi)} \int_{-\infty}^{\infty} \int_{-\infty}^{d} -\frac{1}{2}\left(z_1^2 + z_2^2\right) dz_1 \cdot dz_2, & E[\mu_1] > E[\mu_2] \\ \frac{1}{(2\pi)} \int_{-\infty}^{\infty} \int_{d}^{\infty} -\frac{1}{2}\left(z_1^2 + z_2^2\right) dz_1 \cdot dz_2, & E[\mu_2] > E[\mu_1] \end{cases} \tag{9}$$

where d is the x-intercept of the rotated integration boundary. This can be easily integrated to

$$p_i = \begin{cases} \frac{1}{2}\left(1 + erf\left(d/\sqrt{2}\right)\right), & E[\mu_1] > E[\mu_2] \\ \frac{1}{2}\left(1 - erf\left(d/\sqrt{2}\right)\right), & E[\mu_2] > E[\mu_1] \end{cases} \tag{10}$$

This function can then be quickly evaluated numerically for each feature on the array to test the null hypothesis that no change in gene expression is present.

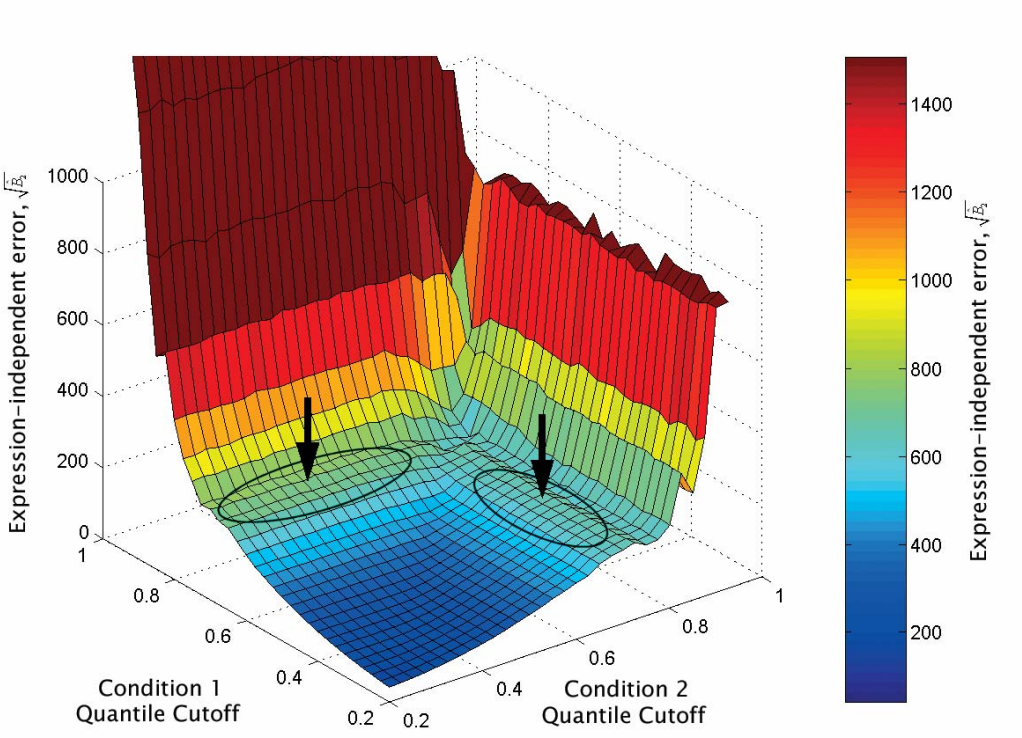
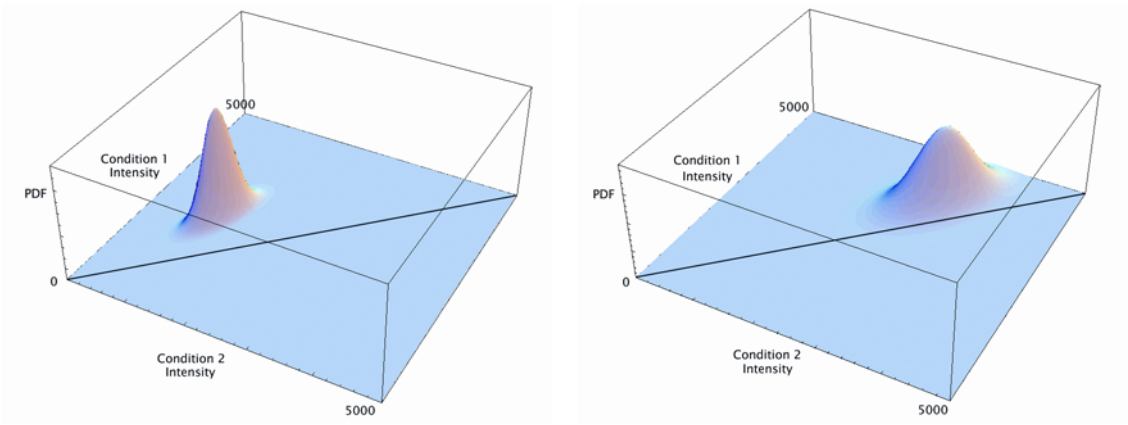


FIGURE 3.1

Estimation of expression-independent error in a sample data set. The estimation procedure computes parameters for a series of expression-level quantile cutoffs. Array features with average measurements below the cutoffs are discarded, and the most likely parameters for the remaining data are computed. In this figure, each point on the colored surface describes the estimated value of expression-independent error ($\sqrt{\hat{B}_2}$) for a given pair of cutoffs. There are two rectangular regions where expression-independent error appears to be stable. These regions are identified by arrows.



(a) (b)

FIGURE 3.2

Depiction of the significance integral. The statistical test performed by BMA involves integration of a bivariate normal density. The peak of the density appears on one side of the diagonal line. The *p*-value is defined as the ìtail volumeî on the side opposite of the peak. Statistical significance is achieved when the *p*-value is smaller than a specified threshold. The bivariate normal density depicted in (a) was sufficiently distant from the diagonal, and was considered a ìsignificantî change (α_{Bonf}=0.05). The density depicted in (b) was too close to the diagonal, and therefore does not represent a significant difference in gene expression.

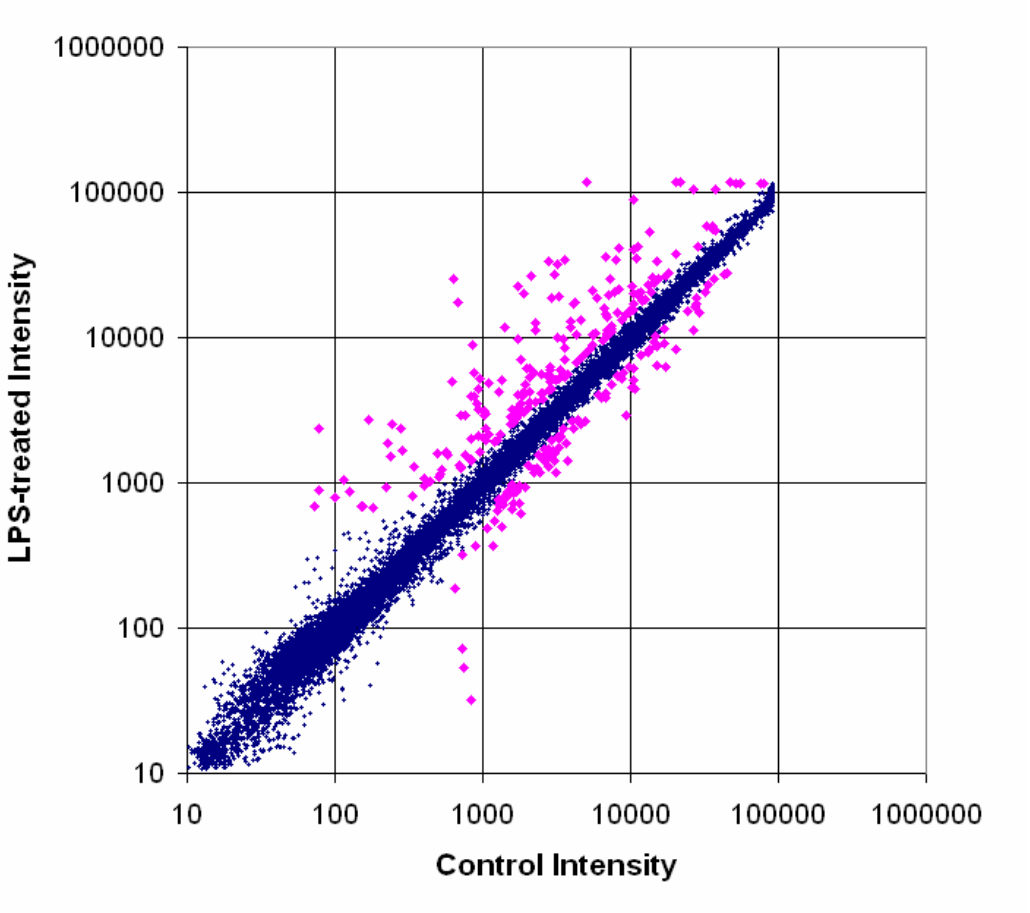
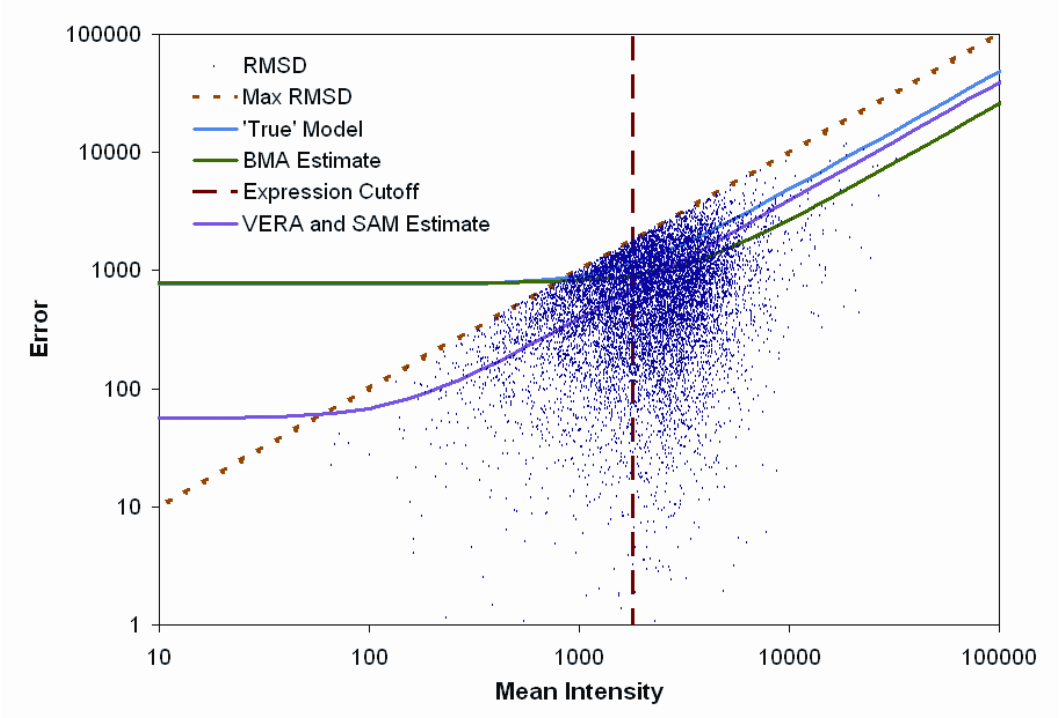


FIGURE 3.3

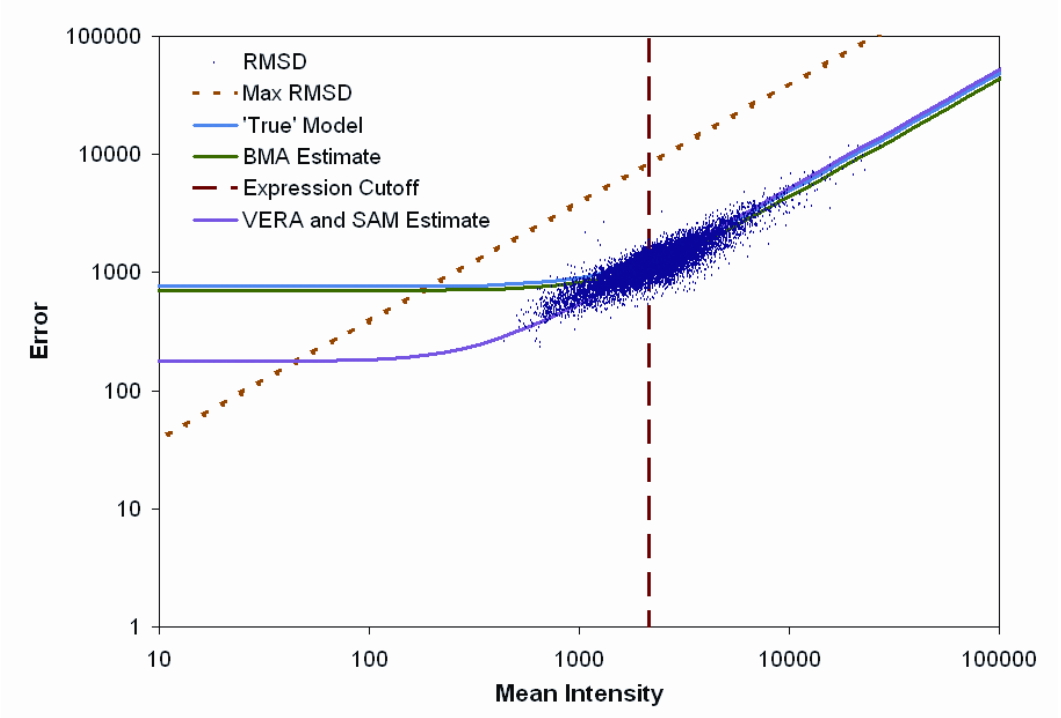
Scatterplot of significant changes. A scatter plot demonstrating the results of BMA on macrophages treated with LPS for 1 hour (α_{Bonf}=0.05) is shown above. Each point represents the average signal measured from an array feature ($n=2$). Highlighted dots appear farthest from the diagonal, indicating statistical significance.

FIGURE 3.4

The variance structure estimated by BMA tightly fits the RMSD (root mean square deviation) as well as the underlying variance model. The simulated data set shown here was obtained by sampling from a ëtrueí model derived from the Rome *et al*, 2003 two-color data set. The quality of fit to RMSD is shown for (a) n=2 and (b) n=16 in non-transformed coordinates. The vertical dashed line indicates the location of the quantile cutoff used to estimate the variance parameters. The diagonal line labeled ìMax RMSDî shows the maximum possible value for RMSD at each expression level. The variance model estimated by BMA tightly matches the ëtrueí model, particularly at low intensities. This accuracy was not matched by VERA and SAM, even when large numbers of replicates were used.


RMSD
Max RMSD
'True' Model
BMA Estimate
Expression Cutoff
VERA and SAM Estimate
Error
Mean Intensity
100000
10000
1000
100
10
1
10
100
1000
10000
100000

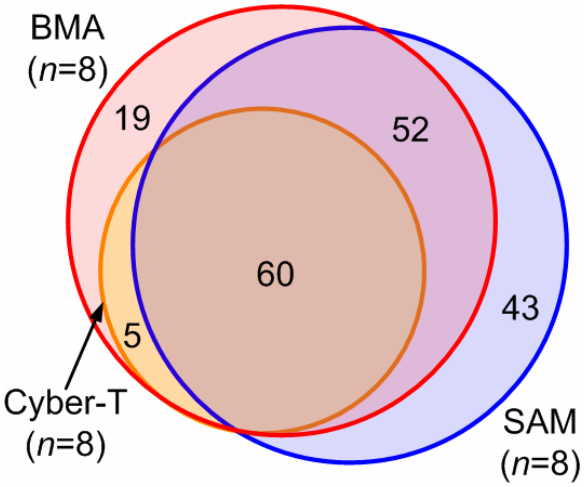
(a)


RMSD
Max RMSD
'True' Model
BMA Estimate
Expression Cutoff
VERA and SAM Estimate
Error
Mean Intensity
100000
10000
1000
100
10
1
10
100
1000
10000
100000

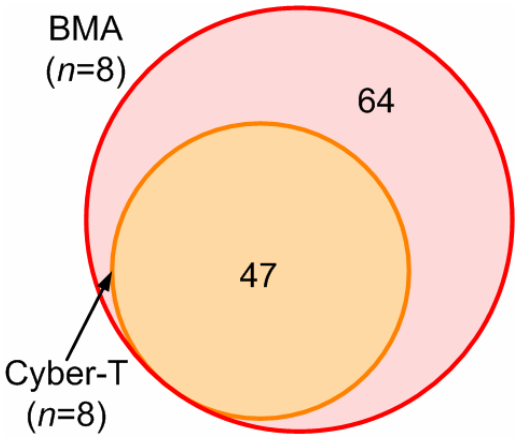
(b)

FIGURE 3.5

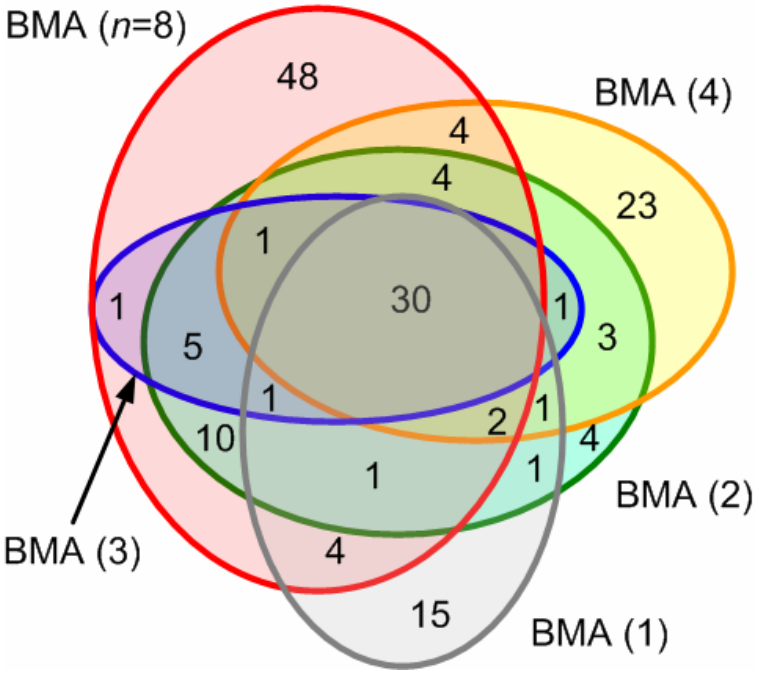
Comparison of methods used to analyze PDGF-treated T98G cells. (a) With four dye-swap pairs (n=8, FDR=0.005), BMA and SAM show strong correlation, sharing about 70% of detected features. Most features identified by Cyber-T were also detected the other two methods. The vast majority of genes detected by SAM and Cyber-T were also detected by BMA. (b) Similar results were obtained using Bonferroni-corrected thresholds for BMA and Cyber-T (n=8, αBonf=0.05). (c) Analysis of individual dye-swap pairs (n=2, αBonf=0.05) labeled 1, 2, 3, and 4 agreed substantially with each other and with the joint analysis of all four pairs (n=8, αBonf=0.05). Therefore, even at low replicate numbers, BMA can detect a substantial fraction of meaningful gene expression changes.


BMA
(n=8)
19
52
60
43
5
Cyber-T
(n=8)
SAM
(n=8)

(a)


BMA
(n=8)
64
47
Cyber-T
(n=8)

(b)


BMA (n=8)
48
BMA (4)
4
4
23
1
30
1
1
3
5
1
2
1
4
10
1
1
BMA (2)
BMA (3)
4
15
BMA (1)

(c)

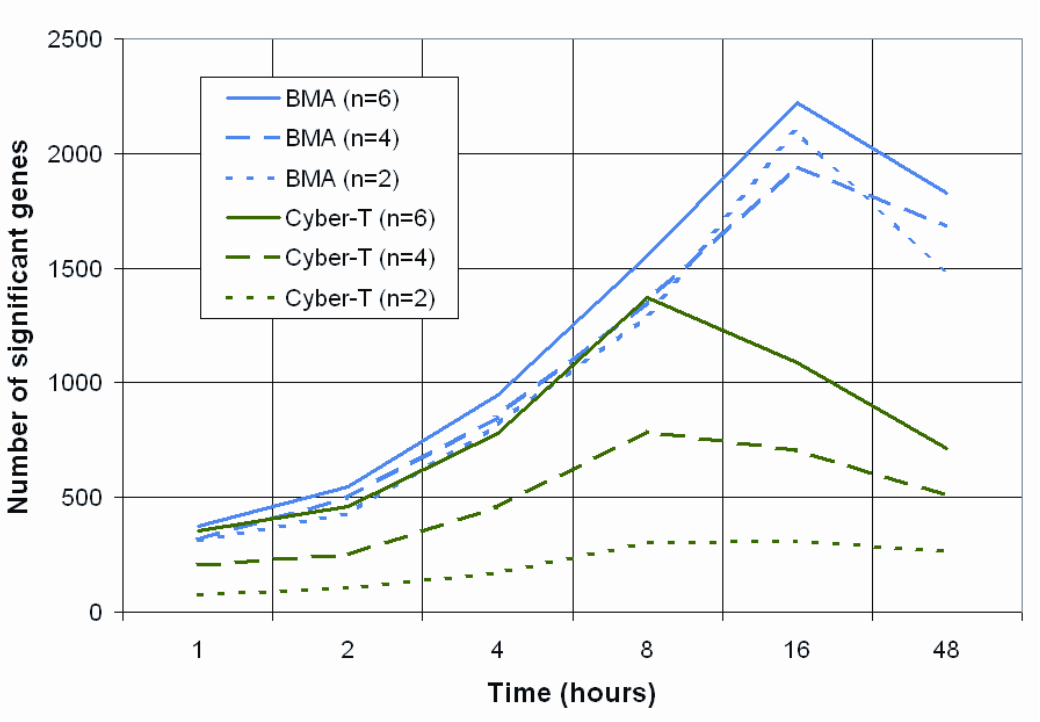
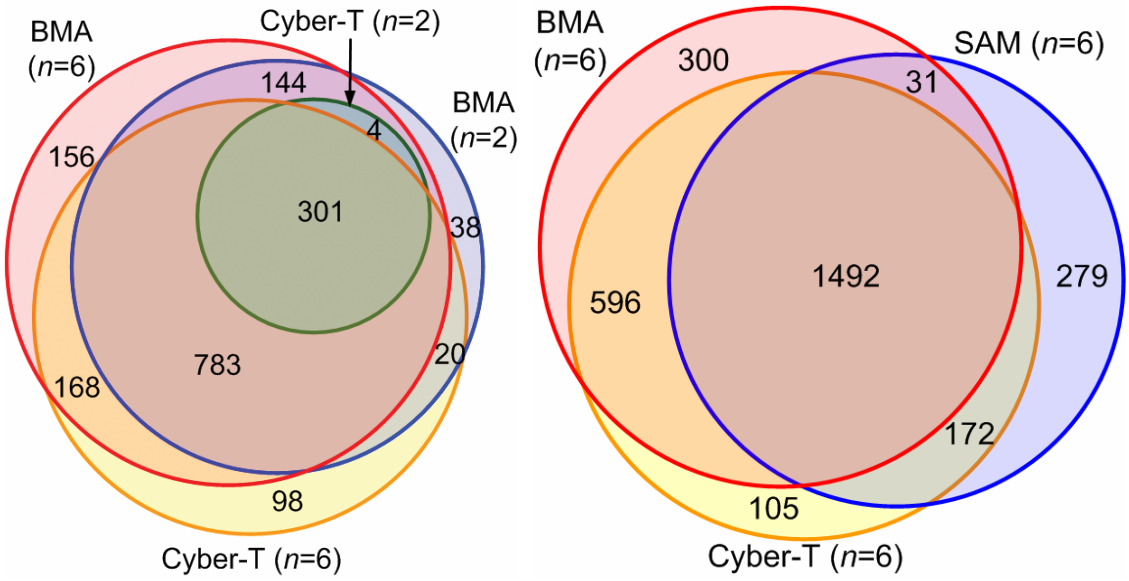


FIGURE 3.6

Comparison of methods and replicates in LPS-treated RAW 264.7 cells. BMA can use as few as a single pair of dye-swapped measurements to identify an equivalent number of statistically significant gene expression changes in the LPS-treated RAW 264.7 data set at all six time points. The quantity of significant gene changes (α_{Bonf}=0.05) stays relatively constant for BMA whether or not more replicates are used. Cyber-Tís results approach BMAís as the number of replicates increase.



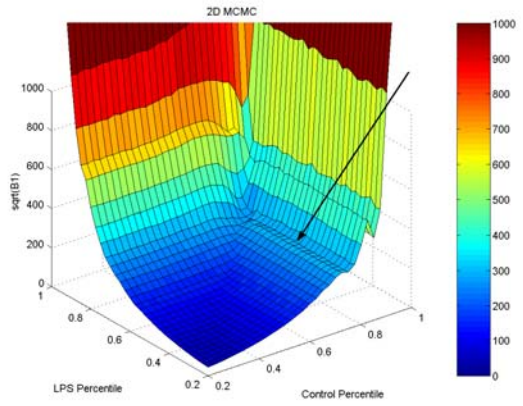
(a) (b)

FIGURE 3.7

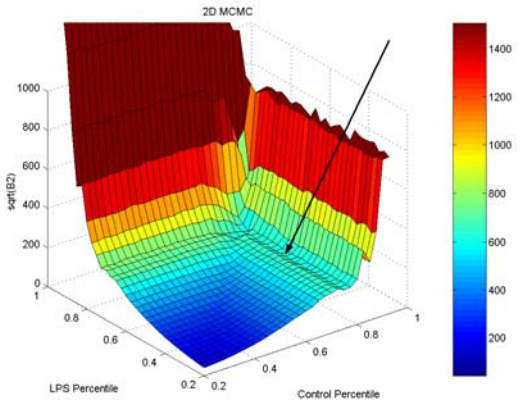
Comparison of methods used to interpret 8 hour LPS-treatment of RAW 264.7 cells. The identity of genes identified by BMA and Cyber-T are very similar at the 8 hour time point when all replicate data is used (n=6). BMA also obtains similar results when fewer replicates are used (n=2). A significance threshold of (a) α_{Bonf}=0.05 or (b) FDR=0.001 was used for all methods.

FIGURE 3.8

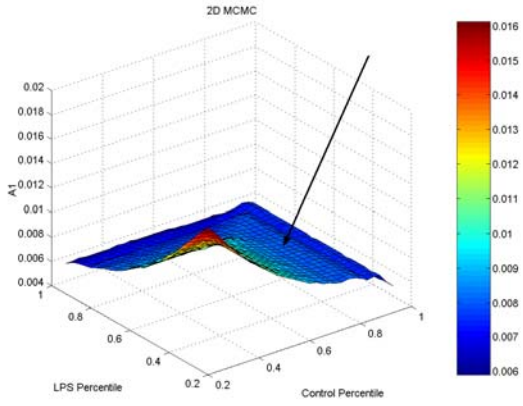
Identification of stable expression cutoffs in a lowess-normalized, LPS-treated RAW 264.7 data set (n=2). The selection of stable regions for the expression-independent variance parameters (a, b) results in stable parameter estimates in other variance parameters (c, d, e). The selected stable region is identified by the arrows. The correlation of expression-independent variances, ρB, approaches 1 (f).


2D MCMC
sqrt(B1)
LPS Percentile
Control Percentile

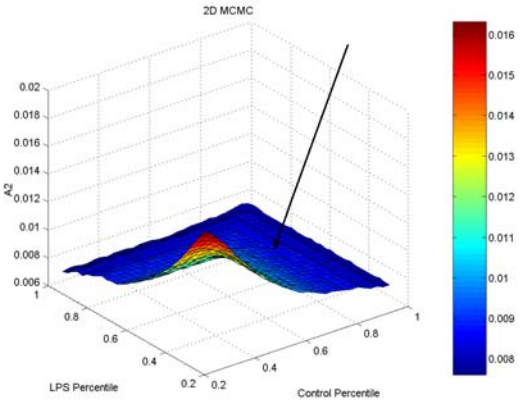
(a)


2D MCMC
sqrt(B2)
LPS Percentile
Control Percentile

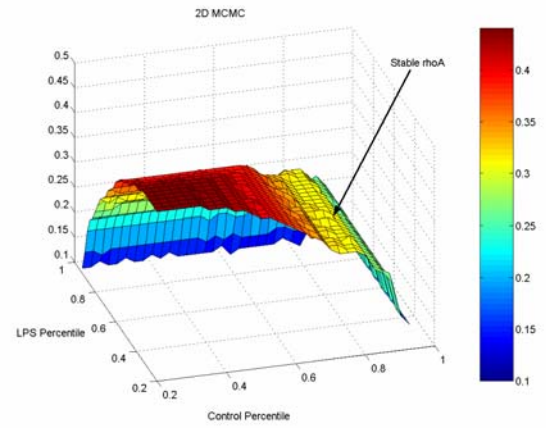
(b)


2D MCMC
A1
LPS Percentile
Control Percentile

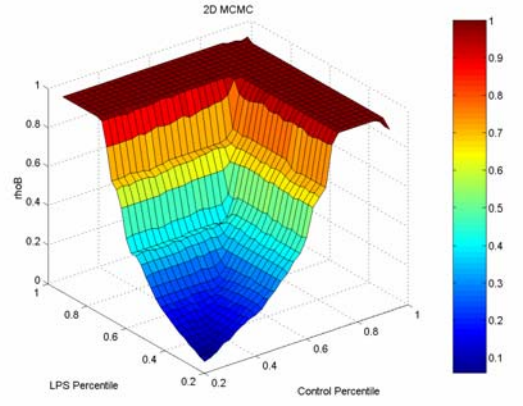
(c)


2D MCMC
A2
LPS Percentile
Control Percentile

(d)


2D MCMC
Stable rhoA
LPS Percentile
Control Percentile

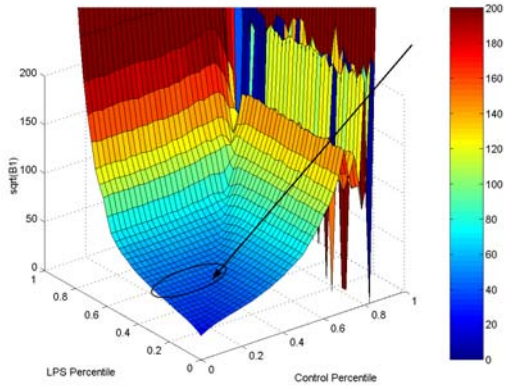
(e)


2D MCMC
rhoB
LPS Percentile
Control Percentile

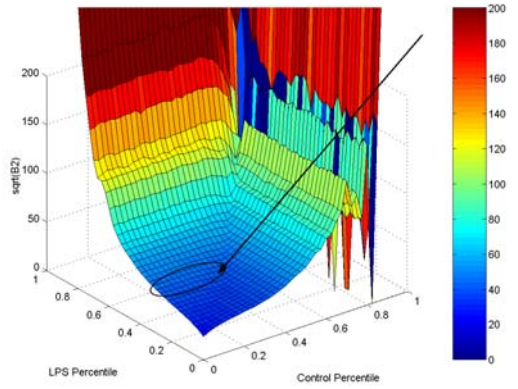
(f)

FIGURE 3.9

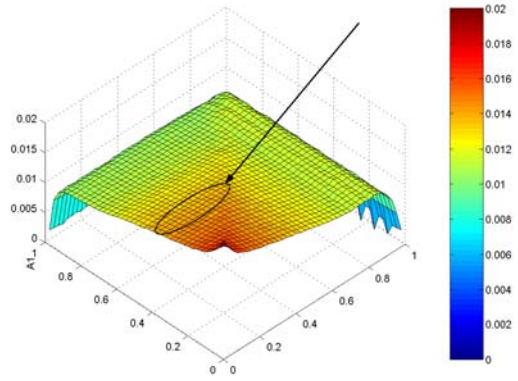
Identification of stable expression cutoffs in a median-normalized, LPS-treated RAW 264.7 data set (*n*=2). The selection of stable regions for the expression-independent variance parameters (a, b) results in stable parameter estimates in all other variance parameters (c, d, e, f). The selected stable region is identified by the arrows.


sqrt(B1)
LPS Percentile
Control Percentile

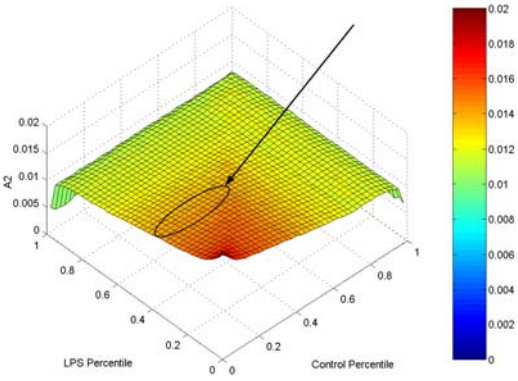
(a)


sqrt(B2)
LPS Percentile
Control Percentile

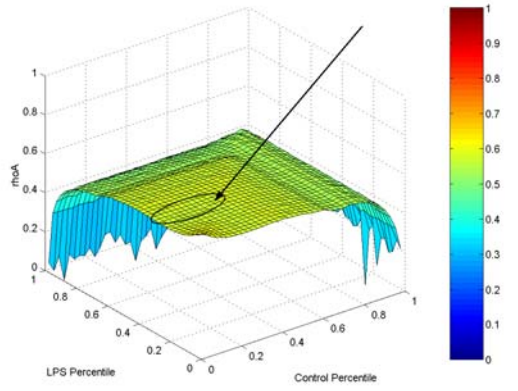
(b)


A1_

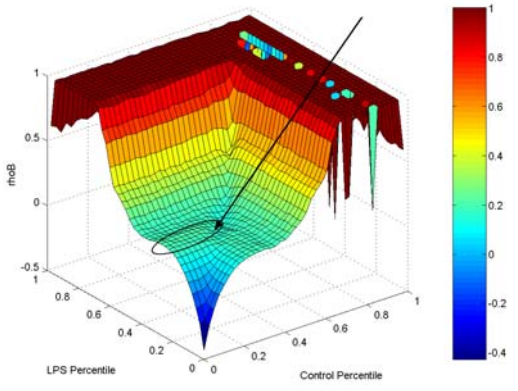
(c)


A2
LPS Percentile
Control Percentile

(d)


rhoA
LPS Percentile
Control Percentile

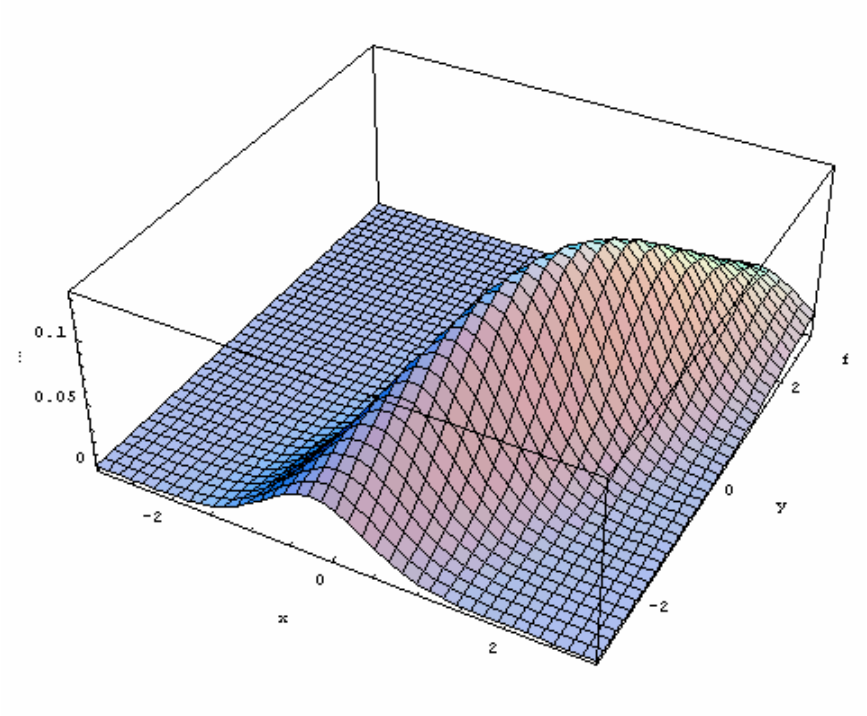
(e)


rhoB
LPS Percentile
Control Percentile

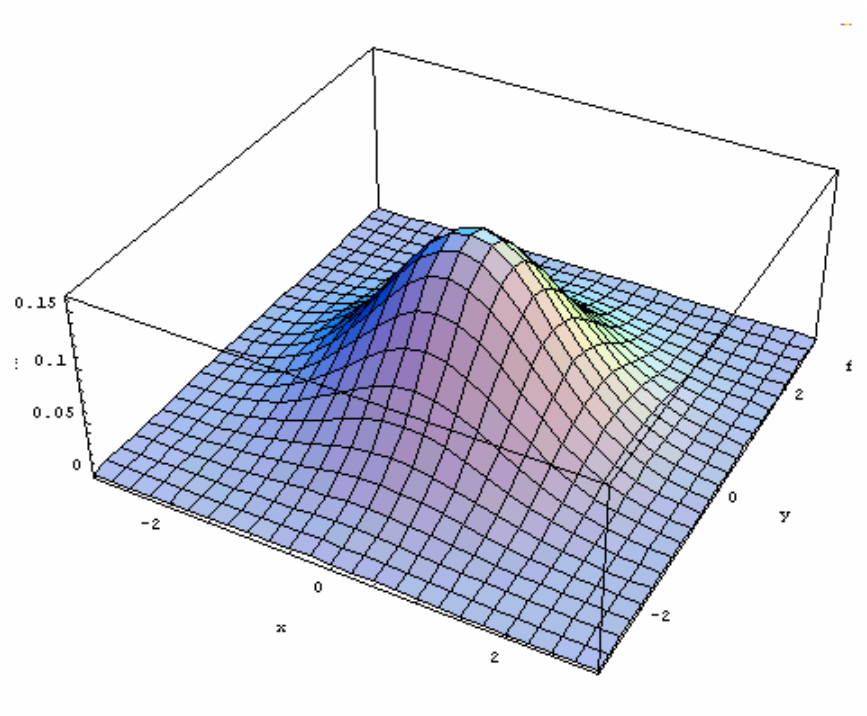
(f)

FIGURE 3.10

Depiction of the bivariate normal density. (a) A bivariate normal density with μ=(1,0), σ_1^2=1, σ_2^2=4, ρ=0.8 and (b) a unit bivariate normal density with μ=(0,0), σ_1^2=1, σ_2^2=1, ρ=0.



(a)



(b)

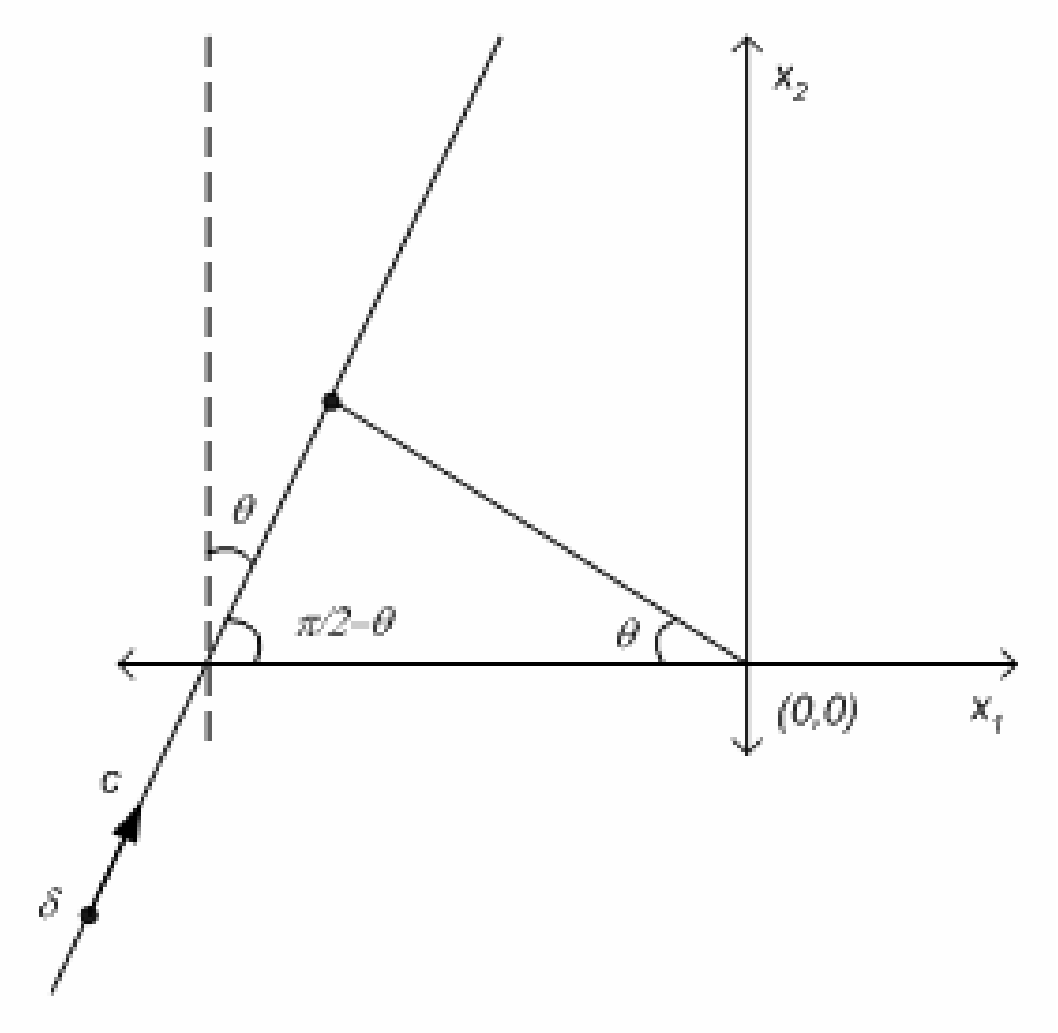


FIGURE 3.11

Rotation of the integration boundary. The integration boundary can be rotated by angle θ to produce a transformation which preserves volume, while simplifying the computation. Symbols used in the diagram are described in the text.

TABLE 3.1

Simulated data sets of 20000 features, 2000 spiked genes were analyzed with four different methods to determine the accuracy of predictions of differential-expression. The number of true-positives (TP), number of false-positives (FP), false-positive rate (FPR), false-negative rate (FNR), sensitivity, and specificity are reported for each method when either the Bonferroni-corrected threshold (αBonf) or false discovery rate (FDR) are controlled. BMA demonstrates substantial improvements in sensitivity across all significance thresholds investigated. BMA maintains specificity when either stringent significance thresholds are applied (FDR=0.001, αBonf=0.05) or replicate numbers are large (n=6, n=16).

***n*=2**

Threshold	Method	TP	FP	FPR	FNR	Sensitivity	Specificity
α_{Bonf}=0.05	BMA	**404**	0	0.000	0.081	**0.202**	1.000
	Cyber-T paired	78	0	0.000	0.096	0.039	1.000
FDR=0.001	VERA and SAM	274	9	0.032	0.088	0.137	1.000
	BMA	**657**	14	0.021	0.069	**0.329**	0.999
	Cyber-T paired	139	0	0.000	0.094	0.070	1.000
	SAM paired	0	0	N/A	0.100	0.000	1.000
FDR=0.05	BMA	**1171**	442	0.274	0.045	**0.586**	0.975
	BMA (*n*=6 model)	472	5	0.010	0.078	0.236	1.000
	Cyber-T paired	542	3	0.006	0.075	0.271	1.000
	SAM paired	130	1	0.008	0.094	0.065	1.000

***n*=6**

Threshold	Method	TP	FP	FPR	FNR	Sensitivity	Specificity
α_{Bonf}=0.05	BMA	**937**	0	0.000	0.056	**0.469**	1.000
	Cyber-T paired	269	0	0.000	0.088	0.135	1.000
FDR=0.001	VERA and SAM	1155	4	0.003	0.045	0.578	1.000
	BMA	**1211**	3	0.002	0.042	**0.606**	1.000
	Cyber-T paired	387	0	0.000	0.082	0.194	1.000
	SAM paired	0	0	N/A	0.100	0.000	1.000
FDR=0.05	BMA	**1514**	138	0.084	0.026	**0.757**	0.992
	Cyber-T paired	925	2	0.002	0.056	0.463	1.000
	SAM paired	266	8	0.029	0.088	0.133	1.000

***n*=16**

Threshold	Method	TP	FP	FPR	FNR	Sensitivity	Specificity
α_{Bonf}=0.05	BMA	**1479**	3	0.002	0.028	**0.740**	1.000
	Cyber-T paired	829	0	0.000	0.061	0.415	1.000
FDR=0.001	VERA and SAM	1554	34	0.021	0.024	0.777	0.998
	BMA	**1589**	16	0.010	0.022	**0.795**	0.999
	Cyber-T paired	1096	0	0.000	0.048	0.548	1.000
	SAM paired	930	0	0.000	0.056	0.465	1.000
FDR=0.05	BMA	**1718**	160	0.085	0.016	**0.859**	0.991
	Cyber-T paired	1441	6	0.004	0.030	0.721	1.000
	SAM paired	1601	61	0.037	0.022	0.801	0.997

TABLE 3.2

Statistically enriched GO terms among differentially-expressed features after 1 hour of LPS treatment (BMA, α_{Bonf}=0.05). The number of features annotated with the GO term and the *p*-value are also displayed. Only the ten GO terms with the lowest *p*-values are shown. All displayed GO terms are significantly enriched (α_{Bonf}=0.05). The complete lists are provided as tables 3.3 and 3.4.

***n*=2**

GO ID	GO Term Name	count	p-value
GO:0006952	defense response	46	0.0000000000
GO:0006954	inflammatory response	20	0.0000000000
GO:0006955	immune response	41	0.0000000000
GO:0009607	response to biotic stimulus	48	0.0000000000
GO:0045087	innate immune response	20	0.0000000000
GO:0009605	response to external stimulus	51	0.0000000001
GO:0009611	response to wounding	22	0.0000000001
GO:0009613	response to pest/pathogen/parasite	24	0.0000000041
GO:0050896	response to stimulus	53	0.0000000047
GO:0005125	cytokine activity	22	0.0000000060

***n*=6**

GO ID	GO Term Name	count	p-value
GO:0006952	defense response	48	0.0000000000
GO:0006955	immune response	44	0.0000000000
GO:0009607	response to biotic stimulus	51	0.0000000000
GO:0009611	response to wounding	25	0.0000000000
GO:0006954	inflammatory response	21	0.0000000001
GO:0045087	innate immune response	21	0.0000000001
GO:0006915	apoptosis	30	0.0000000007
GO:0012501	programmed cell death	30	0.0000000010
GO:0009613	response to pest/pathogen/parasite	27	0.0000000021
GO:0009605	response to external stimulus	54	0.0000000060

TABLE 3.3

Statistically enriched GO terms among differentially-expressed features after 1 hour of LPS treatment (BMA, $n=2$, $\alpha_{Bonf}=0.05$). The number of features annotated with the GO term and the p-value are also displayed. Only significantly enriched ($\alpha_{Bonf}=0.05$) GO terms are displayed.

GO ID	GO Term Name	count	p-value
GO:0006952	defense response	46	0.0000000000
GO:0006954	inflammatory response	20	0.0000000000
GO:0006955	immune response	41	0.0000000000
GO:0009607	response to biotic stimulus	48	0.0000000000
GO:0045087	innate immune response	20	0.0000000000
GO:0009605	response to external stimulus	51	0.0000000001
GO:0009611	response to wounding	22	0.0000000001
GO:0009613	response to pest/pathogen/parasite	24	0.0000000041
GO:0050896	response to stimulus	53	0.0000000047
GO:0005125	cytokine activity	22	0.0000000060
GO:0006915	apoptosis	25	0.0000000167
GO:0012501	programmed cell death	25	0.0000000222
GO:0017017	MAP kinase phosphatase activity	5	0.0000000561
GO:0008219	cell death	25	0.0000001005
GO:0016265	death	25	0.0000001504
GO:0050874	organismal physiological process	44	0.0000002362
GO:0007249	I-kappaB kinase/NF-kappaB cascade	7	0.0000003612
GO:0008009	chemokine activity	10	0.0000006489
GO:0042379	chemokine receptor binding	10	0.0000006489
GO:0007243	protein kinase cascade	13	0.0000006952
GO:0006950	response to stress	30	0.0000007011
GO:0019221	cytokine and chemokine mediated signaling pathway	7	0.0000008460
GO:0001664	G-protein-coupled receptor binding	10	0.0000011277
GO:0006935	chemotaxis	13	0.0000053891
GO:0042330	taxis	13	0.0000053891
GO:0005102	receptor binding	26	0.0000204434
GO:0042981	regulation of apoptosis	14	0.0000212681
GO:0043067	regulation of programmed cell death	14	0.0000239266
GO:0045727	positive regulation of protein biosynthesis	5	0.0000434267
GO:0007515	lymph gland development	6	0.0000449874
GO:0009891	positive regulation of biosynthesis	5	0.0000553049
GO:0008283	cell proliferation	30	0.0001693876

TABLE 3.4

Statistically enriched GO terms among differentially-expressed features after 1 hour of LPS treatment (BMA, $n=6$, $\alpha_{Bonf}=0.05$). The number of features annotated with the GO term and the p-value are also displayed. Only significantly enriched ($\alpha_{Bonf}=0.05$) GO terms are displayed.

GO ID	GO Term Name	count	p-value
GO:0006952	defense response	48	0.0000000000
GO:0006955	immune response	44	0.0000000000
GO:0009607	response to biotic stimulus	51	0.0000000000
GO:0009611	response to wounding	25	0.0000000000
GO:0006954	inflammatory response	21	0.0000000001
GO:0045087	innate immune response	21	0.0000000001
GO:0006915	apoptosis	30	0.0000000007
GO:0012501	programmed cell death	30	0.0000000010
GO:0009613	response to pest/pathogen/parasite	27	0.0000000021
GO:0009605	response to external stimulus	54	0.0000000060
GO:0008219	cell death	30	0.0000000061
GO:0016265	death	30	0.0000000099
GO:0000074	regulation of cell cycle	25	0.0000000844
GO:0017017	MAP kinase phosphatase activity	5	0.0000001451
GO:0045727	positive regulation of protein biosynthesis	7	0.0000003404
GO:0009891	positive regulation of biosynthesis	7	0.0000004891
GO:0050896	response to stimulus	55	0.0000005172
GO:0042981	regulation of apoptosis	18	0.0000005642
GO:0043067	regulation of programmed cell death	18	0.0000006591
GO:0005125	cytokine activity	21	0.0000007369
GO:0050789	regulation of biological process	76	0.0000009415
GO:0007243	protein kinase cascade	14	0.0000010168
GO:0007249	I-kappaB kinase/NF-kappaB cascade	7	0.0000012968
GO:0050874	organismal physiological process	48	0.0000014991
GO:0006950	response to stress	33	0.0000016729
GO:0001817	regulation of cytokine production	7	0.0000030046
GO:0042035	regulation of cytokine biosynthesis	7	0.0000030046
GO:0008283	cell proliferation	39	0.0000034395
GO:0001819	positive regulation of cytokine production	6	0.0000035027
GO:0042108	positive regulation of cytokine biosynthesis	6	0.0000035027
GO:0008270	zinc ion binding	26	0.0000049467
GO:0001816	cytokine production	7	0.0000078729
GO:0042089	cytokine biosynthesis	7	0.0000078729
GO:0042107	cytokine metabolism	7	0.0000078729
GO:0008243	plasminogen activator activity	4	0.0000101660
GO:0043066	negative regulation of apoptosis	9	0.0000106530
GO:0043069	negative regulation of programmed cell death	9	0.0000106530
GO:0050791	regulation of physiological process	56	0.0000168504
GO:0005634	nucleus	81	0.0000200197
GO:0019222	regulation of metabolism	54	0.0000242341
GO:0008009	chemokine activity	9	0.0000266418
GO:0042379	chemokine receptor binding	9	0.0000266418
GO:0007049	cell cycle	31	0.0000391496
GO:0006935	chemotaxis	13	0.0000403410
GO:0042330	taxis	13	0.0000403410
GO:0019221	cytokine and chemokine mediated signaling pathway	6	0.0000406230
GO:0001664	G-protein-coupled receptor binding	9	0.0000427076
GO:0006350	transcription	49	0.0002984442

TABLE 3.5

Statistically enriched KEGG pathways among differentially-expressed features after 1 hour of LPS treatment (BMA, α_{Bonf}=0.05). The number of features annotated with the pathway and the *p*-value are also displayed. Only pathways whose *p*-values lie below the significance threshold are shown (α_{Bonf}=0.05).

***n*=2**

KEGG ID	KEGG Pathway	count	p-value
KEGG:mmu04620	Toll-like receptor signaling pathway	19	0.0000000000
KEGG:mmu04210	Apoptosis	10	0.0002515006

***n*=6**

KEGG ID	KEGG Pathway	count	p-value
KEGG:mmu04620	Toll-like receptor signaling pathway	22	0.0000000000
KEGG:mmu04210	Apoptosis	14	0.0000026563
KEGG:mmu04010	MAPK signaling pathway	23	0.0000229708

May 19, 2005

Albert Hsiao has my permission to include the following paper, of which I was a co-author, in his doctoral dissertation.

Hsiao, Albert; Ideker, Trey; Olefsky, Jerrold M.; Subramaniam, Shankar. ì VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression dataî , Nucleic Acids Research, web server issue, 2005.

_____

Trey Ideker

_____

Jerrold M. Olefsky

_____

Shankar Subramaniam

This chapter, in full, is a reprint of the material as it appears in Nucleic Acids Research 2005, Hsiao, Albert; Ideker, T.; Olefsky, Jerrold M.; Subramaniam, Shankar., Oxford University Press, 2005. The dissertation author was the primary investigator and single author of this paper.

CHAPTER IV

VAMPIRE Microarray Analysis Suite

A. Abstract

Microarrays are invaluable high-throughput tools used to snapshot the gene expression profiles of cells and tissues. Among the most basic and fundamental questions asked of microarray data is whether individual genes are significantly activated or repressed by a particular stimulus. We have previously presented two Bayesian statistical methods for this level of analysis, collectively known as VAMPIRE. These methods each require a sophisticated modeling step followed by integration of a posterior probability density. We present here a publicly available, web-based platform that allows users to easily load data, associate related samples, and identify differentially-expressed features using the VAMPIRE statistical framework. In addition, this suite of tools seamlessly integrates a novel gene annotation tool, known as GOby, which identifies statistically overrepresented gene groups. Unlike other tools in this genre, GOby can localize enrichment while respecting the hierarchical structure of annotation systems like Gene Ontology (GO). By identifying statistically significant enrichment of GO terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and TRANSFAC transcription factor binding sites, users can gain substantial insight into the physiological significance of sets of differentially-expressed genes. The VAMPIRE microarray suite can be accessed at http://genome.ucsd.edu/microarray.

B. Introduction

Gene expression microarrays are commonly used to study the transcriptional responses of cells and tissues. Most studies involve comparisons of an experimental treatment with a corresponding control, often with only 2-3 replicates within each

treatment group. These experiments are commonly performed on either one-channel or two-channel microarray platforms, which simultaneously measure gene expression in one or two RNA samples, respectively. As biologists devise more sophisticated experimental designs, existing statistical tools for analyzing these data become increasingly unwieldy. For example, in large data sets of over 50 arrays, users may wish to begin analysis before all microarrays have been completely processed. As users define more statistical tests to perform on this data, accounting for all of the tests and changes to the underlying data becomes extremely challenging. Furthermore, because of the high-throughput nature of these experiments, and because of their non-uniform error structure, the resulting data can be difficult to interpret. To address these issues, we have devised a tightly integrated, web-based suite of microarray analysis tools based on a robust Bayesian approach known as VAMPIRE[27].

The microarray analysis platform we present here provides an integrated interface for data management, statistical analysis, and interpretation of gene expression data. To simplify the analysis of large data sets, this interface allows users to quickly load data and characterize experimental designs within the data set. Users can associate related samples and combine related groups. The variance structure of these groups can then be modeled to identify the coefficients of expression-dependent (A) and expression-independent variance (B). The analysis suite subsequently uses these models, stored in the userís account, and applies them to identify microarray features that are differentially-expressed between treatment groups. Once this analysis is complete, the corresponding gene lists must still be interpreted and related to biological function. We have therefore integrated a novel statistical tool known as GOby, which can identify overrepresentation of previously-defined functional categories. This is performed while respecting the hierarchical nature of annotation systems like Gene Ontology[8] (GO). Together, these

data management and analytical tools provide users with a powerful new approach to microarray gene expression analysis.

C. Implementation

The web application is implemented as a package of Java 1.5 servlets. In order to separate interface and the underlying content and logic, we applied a commonly-used three-tier scheme. At the foundation of the VAMPIRE analysis suite is a SQL-based relational database. The current implementation uses a combination of MySQL and Oracle databases. Java interfaces have been appropriately designed to isolate SQL-specific code, which allows us to quickly re-deploy the system on other database platforms. Each servlet gathers data through these SQL interfaces and constructs its own intermediate XML document. XML is subsequently transformed via XSLT into HTML. The end-user is ultimately presented with the HTML interface. By implementing VAMPIRE in this fashion, each component can be re-designed or re-implemented without disturbing the remainder of the site.

The VAMPIRE algorithm itself is computationally intensive. On an Intel Xeon 3.0 GHz processor, typical execution time for variance modeling ranges from 5 to 10 minutes depending on the number of features on the microarray, and whether unpaired or paired analysis is desired. Significance testing and GOby analysis usually complete within 2 to 3 minutes. Because of these computational costs, we have devised a Java RMI-based distributed processing solution (Fig. 4.1). Each analysis job constructed by the web-application is placed into a priority queue. An RMI host program provides access to each job as it arrives at the top of the queue to any number of RMI clients. The RMI hosts and clients may be executed on remote machines distant from the web-server. This design provides scalability to the VAMPIRE site, as additional machines may be added to handle the increased computational load. Since all of these transactions are stored in the VAMPIRE database, no data is lost when individual hosts or processing nodes fail. Jobs

may be restarted by other RMI nodes or provided through alternative RMI hosts as long as the database itself is accessible.

D. User Interface

The web-based interface of the microarray analysis suite can be divided into three major sections: 1. data management, 2. statistical analysis, and 3. interpretation. Throughout all three divisions is a consistent user-interface with multiple tools for exporting data (Fig. 4.2). To ease integration with third-party tools, results may be downloaded as tab-delimited text files or as XML documents.

Data Management

Management of microarray data sets and their corresponding analyses becomes increasingly complex as the number of treatment groups grows. Users must manage not only individual samples, but also all subsequent analyses. This is a particular problem for users with large data sets, who wish to begin interpreting gene expression data before all of samples have been completely processed. To accommodate these issues, we have created a data management system that can be quickly used to load, annotate, and associate microarray measurements. Summary measures of gene expression, such as those obtained from Affymetrix MAS 5.0 or Agilent processed signal intensities, can be imported into the server as tab-delimited text files, a file format that is easily accessible to most biologists.

Once data has been loaded into the web application, the user may associate related microarray samples. Characterizing these relationships is crucial, as they help to describe the analyses that will be later performed. For example, in a two-channel tutorial provided on the web site, users create sample groups for: 1. replicates of LPS-treated macrophages, 2. replicates of control-treated macrophages, and 3. paired samples obtained from the same chips. These groups can be further combined to create ìcategoriesî of related groups. Once these relationships are recorded, they can be used to compare gene

expression across different treatment conditions. Since further changes to each sample group are recorded by the analysis suite, VAMPIRE can subsequently inform users when analyses need to be re-executed.

Statistical analysis

Statistical analysis by VAMPIRE requires two distinct steps: 1. modeling of the error structure of sample groups and 2. significance testing with *a priori*-defined significance thresholds. This approach to microarray analysis is considerably different from the approaches taken by other analytical methods[5, 6, 20, 21]. Normalization methods are commonly applied to average out the error structure prior to performing statistical analysis. Statistical tests are then left to address the significance of expression differences found in the remaining data. In contrast, VAMPIRE studies the underlying error structure, without perturbing it, and uses this knowledge to distinguish signal from noise. The cutoff for statistical significance is then defined by the significance threshold and by the magnitude of the variance model coefficients. Because of the additional variance modeling step however, this kind of analysis can be quite challenging without a robust accounting system. In the web application that we present here, users can easily keep track of all variance models and the data sets for which they can be applied. We have initially incorporated two variants of VAMPIRE ñ classical unpaired analysis[27] and paired analysis (Chapter 3); both of which use variance models to detect significant changes in gene expression.

When users submit a request to model the variance structure of a group of samples, a new ìprocessing jobî is immediately submitted into a processing queue. Individual jobs require an average of 5-10 minutes to compute on a Intel Xeon 3.06GHz processor. In the meantime, users may continue to use the remainder of the site, without waiting for each job to complete. An estimate of the date and time of completion is prominently displayed. Similarly, users may request that specific statistical tests be performed. Since these tests

rely on variance model results, they will not be executed until their dependent models have been completed. As data can be continually added to the analysis platform, outdated models and tests are automatically flagged by the system to allow users to re-execute analyses with updated data. This particular feature facilitates ìon-the-flyî analysis. Users can monitor the results as they collect data, which may help them to decide which analyses require additional replication.

Interpretation

Differentially-regulated features obtained from any statistical test must be interpreted biologically. We have developed a novel tool, known as GOby, to initiate biological interpretation, independent of whether VAMPIRE itself was used to derive the feature lists. This database-driven application curates annotation data from several sources: NCBI, Gene Ontology (GO)[8], Kyoto Encyclopedia of Genes and Genomes[9] (KEGG), TRANSFAC[10], Biocarta, and Superarray. In addition, it can be readily updated with additional user-defined annotation lists.

GOby primarily uses its annotation database to identify overrepresented annotation groups. It does so by comparing a ìselectedî list to a ìbackgroundî. In our experience, using the comprehensive feature list for each microarray as the background gives quite meaningful results. In a manner similar to other recently published tools[12-14], GOby uses exact probabilities to compute enrichment likelihoods, and displays the enrichment likelihood as a *p*-value. GOby reports as its *p*-value, the probability of finding no more than *s* features annotated with a given term among *k* ìselectedî features:

$$P\left(annotated \le s\right) = \sum_{i=0}^{s} P\left(annotated = i\right)$$

$$P\left(annotated = i\right) = \binom{k}{i} \prod_{j=0}^{i-1} \frac{b-j}{N-j} \cdot \prod_{j=0}^{k-i-1} \frac{(N-b)-j}{(N-i)-j} = \frac{\binom{b}{i}\binom{N-b}{k-i}}{\binom{N}{k}}$$

where

b is the number of ìbackgroundî features annotated with the term

s is the number of ìselectedî features annotated with the term

N is the total number of ìbackgroundî features

k is the total number of ìselectedî features

Unlike similar tools however, GOby is also able to compute a ìconditional-enrichment likelihoodî, or *q*-value, for each term in a hierarchical annotation system. This *q*-value is based on the idea that truly meaningful enrichment in a hierarchical system like GO will occur at specific nodes in the annotation tree. The *q*-value computes the enrichment likelihood for a particular term *conditioned on the enrichment of its parent terms*. In other words, instead of using the entire set of array features as the ìbackgroundî, we use only the subset of features that are annotated with one of the parent terms. Unlike the *p*-value previously described, the *q*-value prevents annotation terms from reaching significance simply because they lie in an area of the tree near other terms that are enriched. It can therefore narrow the userís focus by reporting only the optimal level of functional detail while excluding both more general and more specific terms (unless these terms fall into a second area of functional enrichment independent of the first). Since both methods have their own advantages, both are displayed in the results.

GOby provides two options for viewing the results of its statistical analysis. Results may be downloaded in an XML format, or they may be viewed directly from the VAMPIRE web site (Fig. 4.3). Since each of the result pages is rendered for GOby via XSLT from the XML export file, all necessary information is easily accessible to third-party applications. Highly-skilled users may wish to render their own views of GOby reports. Most users, however, will rely on automatically generated web pages. For each annotation system, GOby renders at least two types of web pages. First, a result table

displays the annotation terms that are overrepresented given a specified significance cutoff. Significant *p*-values and *q*-values, according a user-selected significance threshold, are displayed in bold. Users may choose between Bonferroni-corrected thresholds and false-discovery rates (FDR) to correct for multiple testing (Chapter 3). Second, each entry in the result table can be clicked to view a page that displays differentially-expressed features that are annotated with the selected term. Because of careful integration with VAMPIRE statistical tests, fold-changes and *p*-values for each array feature can be viewed directly from these pages, eliminating the need to manually cross-reference additional lists. Each feature is linked by accession numbers and LocusLink IDs to corresponding web pages at NCBI. When available, each annotation term has also been linked to pages that display term-specific information. For example, clicking on the KEGG link from a KEGG report page will display the relevant KEGG pathway, while highlighting differentially-expressed genes. In addition, we have included several Javascript-based functions to show/hide columns and to sort tables by column data. For hierarchical systems such as GO, GOby also renders a third view, the tree view. Here, *p*-values and *q*-values are again visible to allow users to quickly address the importance of each node in annotation tree.

E. Conclusions

The VAMPIRE microarray analysis suite provides a fundamentally different approach to gene expression analysis. While other tools independently exist for management of microarray data, analysis, and interpretation, none are yet equipped with the statistical tools that we have described. The VAMPIRE suite gracefully handles incremental analysis of large data sets, applies some of the most rigorous statistical methods available for microarray analysis, and provides powerful tools for interpretation of the results. We have also presented a novel tool, GOby, for interpreting of sets of differentially-regulated genes. It can compute both ìglobal enrichmentî of annotation terms, as well as

ì conditional enrichmentî at specific nodes of a hierarchical annotation system, like Gene Ontology. This suite therefore represents a substantial advance in bringing the latest analytical algorithms into the hands of biologists.

E. Ackowledgements

We thank Chris Benner for several useful scripts, Yi-Xiong Zhou and Joshua Li for assisting in the Oracle-based implementation of the database backend, and Wendy Ching, Dorothy Sears, Nicholas Webster for their valuable comments. AH is a graduate student in the UCSD Medical Scientist Training Program and is supported by a Fellowship from the Whitaker Foundation and the UCSD MSTP Training Grant T35-GM07198. TI is supported by NIH grant 1R01GM070743-01 and a David and Lucille Packard Fellowship Award. SS and JMO acknowledge the NIH grants NIGMS K54 GM62114 and NIDDK RO1-DK33651 respectively, as well as a Life Science Informatics grant from the State of California, Pfizer La Jolla, and a grant from the Hillblom Foundation.

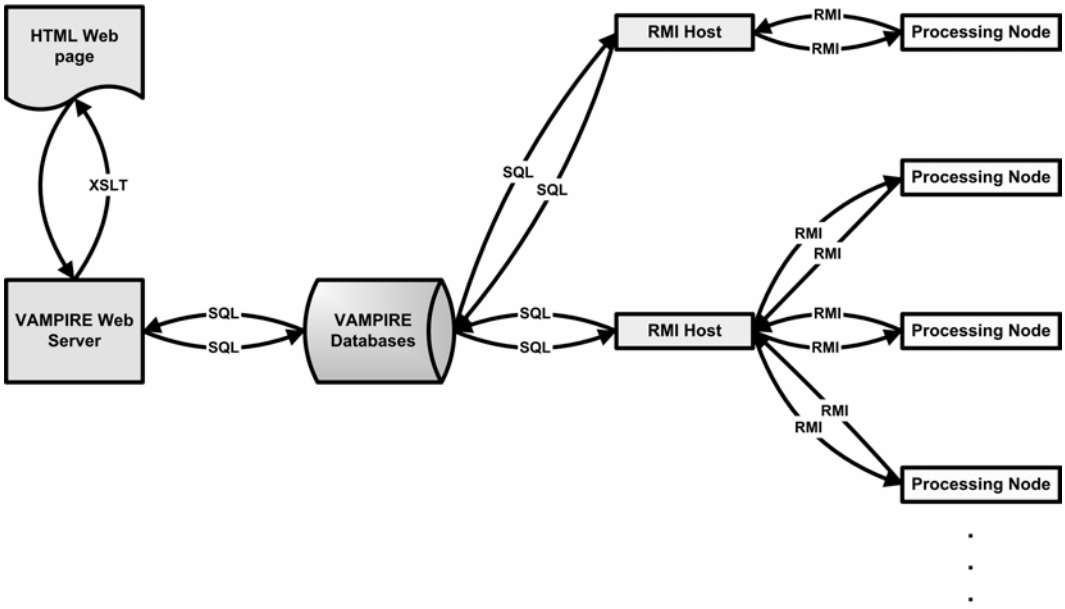


FIGURE 4.1

Design of the VAMPIRE processing pipeline. At the core of the application is a SQL-based relational database. The web-based user-interface generates intermediate XML documents from this database, and transforms these documents into the HTML interface via XSLT. A distributed processing framework based on Java/RMI disperses computational load on to remote processing servers.

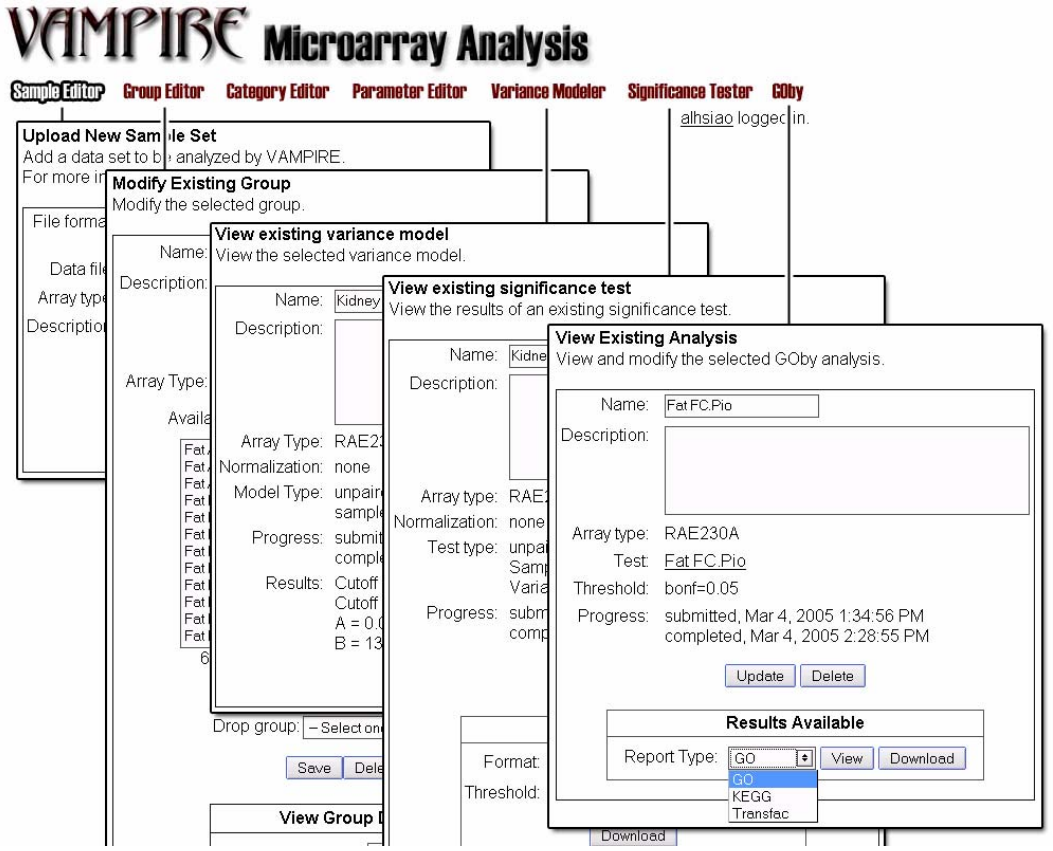
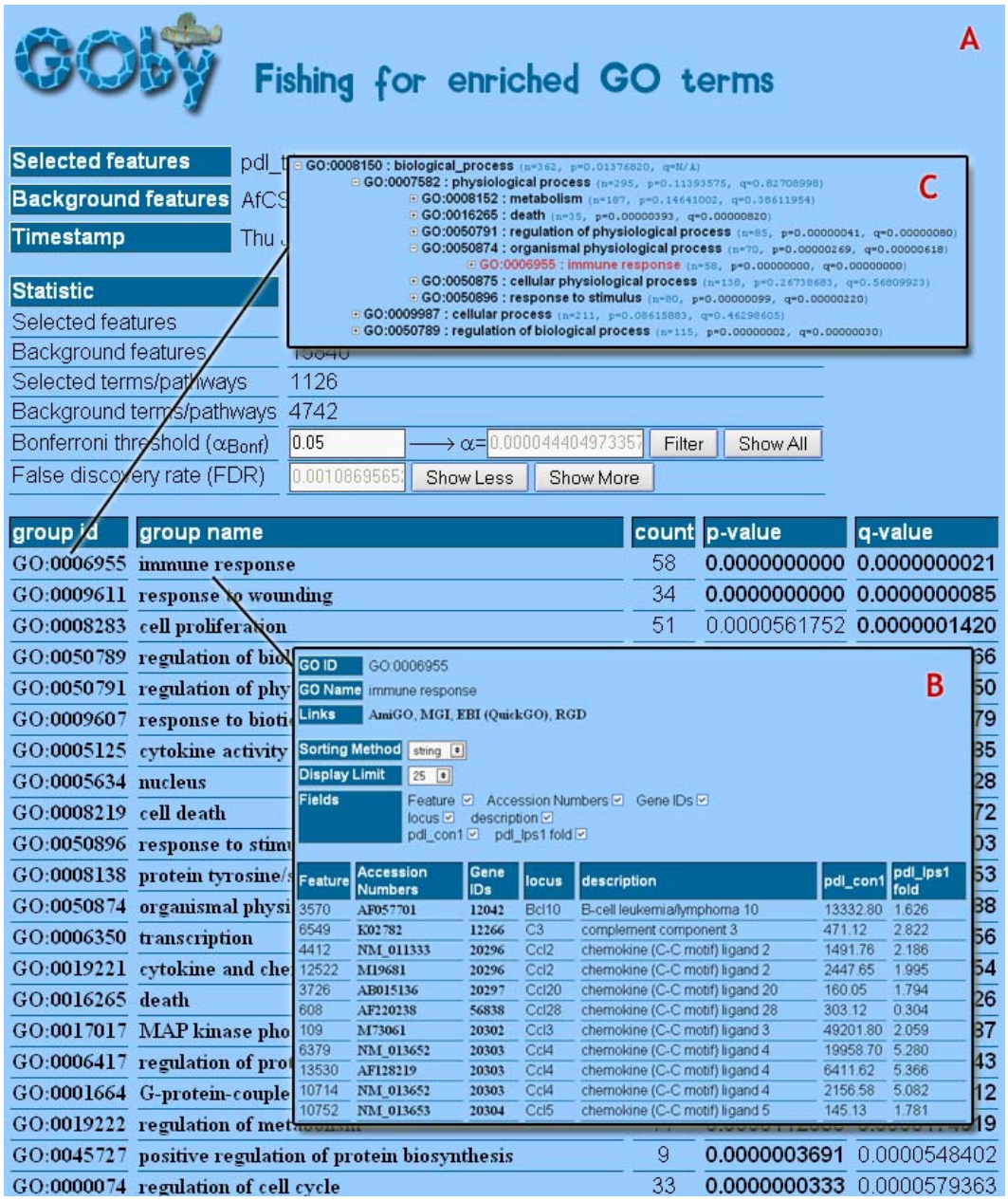


FIGURE 4.2

Screenshots of the VAMPIRE web interface. The intuitive design allows users to easily manage data sets, perform VAMPIRE statistical analysis, and interpret gene expression results with GOby.

FIGURE 4.3

GOby-rendered report pages. Three types of pages are automatically rendered by GOby for navigation of GOby results. The term table (A) displays annotation terms that are enriched among differentially-expressed features. The term pages (B) show differentially-expressed features that are annotated with each term. The tree term (C) displays the hierarchical structure of the annotation system with corresponding enrichment likelihoods.


Goby
Fishing for enriched GO terms
A
Selected features
Background features
Timestamp
C
GO:0008150 : biological_process
GO:0007582 : physiological process
GO:0008152 : metabolism
GO:0016265 : death
GO:0050791 : regulation of physiological process
GO:0050874 : organismal physiological process
GO:0006955 : immune response
GO:0050875 : cellular physiological process
GO:0050896 : response to stimulus
GO:0009987 : cellular process
GO:0050789 : regulation of biological process
Statistic
Selected features
Background features
Selected terms/pathways 1126
Background terms/pathways 4742
Bonferroni threshold (αBonf) 0.05
False discovery rate (FDR)
Filter
Show All
Show Less
Show More
group id
group name
count
p-value
q-value
GO:0006955 immune response 58 0.0000000000 0.0000000021
GO:0009611 response to wounding 34 0.0000000000 0.0000000085
GO:0008283 cell proliferation 51 0.0000561752 0.0000001420
GO:0045727 positive regulation of protein biosynthesis 9 0.0000003691 0.0000548402
GO:0000074 regulation of cell cycle 33 0.0000000333 0.0000579363
B
GO ID GO:0006955
GO Name immune response
Links AmiGO, MGI, EBI (QuickGO), RGD
Sorting Method string
Display Limit 25
Fields Feature Accession Numbers Gene IDs locus description pdl_con1 pdl_lps1 fold
Feature | Accession Numbers | Gene IDs | locus | description | pdl_con1 | pdl_lps1 fold
3570 | AF057701 | 12042 | Bcl10 | B-cell leukemia/lymphoma 10 | 13332.80 | 1.626
6549 | K02782 | 12266 | C3 | complement component 3 | 471.12 | 2.822
4412 | NM_011333 | 20296 | Ccl2 | chemokine (C-C motif) ligand 2 | 1491.76 | 2.186
12522 | M19681 | 20296 | Ccl2 | chemokine (C-C motif) ligand 2 | 2447.65 | 1.995
3726 | AB015136 | 20297 | Ccl20 | chemokine (C-C motif) ligand 20 | 160.05 | 1.794
608 | AF220238 | 56838 | Ccl28 | chemokine (C-C motif) ligand 28 | 303.12 | 0.304
109 | M73061 | 20302 | Ccl3 | chemokine (C-C motif) ligand 3 | 49201.80 | 2.059
6379 | NM_013652 | 20303 | Ccl4 | chemokine (C-C motif) ligand 4 | 19958.70 | 5.280
13530 | AF128219 | 20303 | Ccl4 | chemokine (C-C motif) ligand 4 | 6411.62 | 5.366
10714 | NM_013652 | 20303 | Ccl4 | chemokine (C-C motif) ligand 4 | 2156.58 | 5.082
10752 | NM_013653 | 20304 | Ccl5 | chemokine (C-C motif) ligand 5 | 145.13 | 1.781

CHAPTER V

Comparative Transcriptomics of Insulin-Regulated Expression

A. Abstract

Insulin is a key regulator of glucose homeostasis. While its acute effects on Glut4 translocation, glucose metabolism, and lipolysis have been extensively studied, its global transcriptional activity *in vivo* has not been adequately characterized. As insulin-responsiveness is a primary defect in type 2 diabetes, an understanding of insulinís normal transcriptional targets is necessary for the elucidation of diabetes-specific defects, which are independent of insulin activity. In order to address this *in vivo* response, we undertook a comparative transcriptomics approach. We investigate the global impact of hyperinsulinemic-euglycemic clamp on skeletal muscle and liver in three model organisms: human, mouse, and rat. In addition to novel, conserved expression changes in individual genes to this high-energy state, we characterize the physiological pathways that are most likely impacted by the accumulation of these transcripts. At the time scale of the hyperinsulinemic-euglycemic clamp, we observe in skeletal muscle, an inhibition of cell proliferation pathways with concomitant activation of a differentiation program. In liver, we observe simultaneous activation of inflammatory mediators and suppression of the gluconeogenic program. These results provide evidence for a role of hyperinsulinemia in regulating inflammatory activity, and suggest additional targets for study of energy metabolism.

B. Introduction

Diabetes mellitus is a heterogeneous group of metabolic disorders characterized by impaired regulation of energy metabolism and hyperglycemia. While type 1 diabetes is a disease of insufficient insulin production, type 2 diabetes is typically considered a disease of insulin resistance. To acutely control blood glucose, insulin induces metabolic changes

in several tissues. In skeletal muscle, it promotes glucose uptake and the synthesis of glycogen and protein. In adipose, insulin promotes glucose uptake to support fatty acid and triglyceride synthesis. In liver, it promotes glycogenesis while inhibiting gluconeogenesis. While many of these processes are acutely regulated by translocation of specific transporters and post-translational modification of metabolic enzymes, a large body of evidence suggests that transcriptional control may also be essential for insulin activity [37]. Through its action on a variety of response elements and over 150 genes, insulin regulates transcription of not only metabolic enzymes, but secreted factors, transmembrane receptors, signal transduction components, and transcription factors.

Insulin exerts its effects through several receptors and signal transduction pathways. Not only does it bind and activate the autophosphorylation of the insulin receptor, but insulin-like growth factor (IGF) receptors as well. The specific roles of insulin, IGFs, and their receptors in regulating cell growth and metabolism are fields of active investigation (Kim 2005, Entingh-Persall, 2004). Many of the actions of insulin have been attributed to the downstream phosphatidylinositol 3-kinase (PI3K) and mitogen-activated protein kinase (MAPK) pathways, activating acute effects through second messenger systems, and long-term effects through transcriptional regulation. In particular, the activation of PKB/Akt through PI3K and subsequent nuclear exclusion of Foxo-family transcription factors is believed to responsible for many direct transcriptional activities. A variety of other transcription factors including TTF-2, *c-fos*, *c-jun*, C/EBPα, and C/EBPβ, which also mediate direct and indirect actions of insulin, have also been studied[37].

The development of cDNA and oligonucleotide arrays has further inspired interest in transcriptional regulation of many biological processes. Insulin, in particular, is the subject of several such studies. Previous groups examined the effect of insulin in insulin-treated 3T3-L1 adipocytes[38], streptozotocin (STZ)-treated mice[39], muscle insulin receptor knockout (MIRKO) mice[40], and human subjects undergoing hyperinsulinemic-

euglycemic clamp[29]. Unfortunately, the results of these studies vary considerably in the genes they ultimately identified as insulin-regulated. It is not immediately clear whether these differences are due to weak sensitivity and specificity of previous high-throughput studies, or whether these are due to innate differences between the models being tested. To address this issue, and define a core set of genes coordinately regulated in several model systems, we undertook a comparative transcriptomics approach.

We present here a comparative approach toward the identification of transcriptional targets. In order to identify tissue-specific targets, we performed parallel hyperinsulinemic-euglycemic clamps in three model organisms: C57/Bl6 mice, lean Zucker fa/+ rats, and non-diabetic human patients. The hyperinsulinemic-euglycemic clamp is a well-recognized tool for assessing insulin responsiveness of target tissues, as well as for studying insulin-regulated gene expression. In these models, we examined the transcriptional profiles of skeletal muscle and liver in the presence or absence of hyperinsulinemia. In each of these tissues, we identify a core set of the conserved targets. In skeletal muscle, we demonstrate transcriptional regulation of mediators of cellular proliferation and differentiation. In liver, we show temporal modulation of the gluconeogenic transcriptional network and parallel transcriptional activation of inflammatory mediators. These results recapitulate some conventional concepts of the major roles insulin, while providing novel detail into the mechanisms of transcriptional control and insulin resistance.

C. Methods

Sample preparation.

Biopsies were collected from the vastus lateralus of fifteen non-diabetic human subjects preceding and immediately following hyperinsulinemic-euglycemic clamp at 60 $mU/m^2/min$ for a total of hyperinsulinemic exposure of five hours. Harvested tissue was generously provided Joseph Yu from patients of a forthcoming study (manuscript in

preparation). Total RNA was purified from each tissue with the Fibrous Tissue Mini Kit (Qiagen). Eluted total RNAs were quantified with a portion of the recovered total RNA adjusted to a final concentration of 1.25 ug/ul. All starting total RNA samples were quality assessed prior to beginning target preparation/processing steps by running out a small amount of each sample (typically 25-250 ng/well) onto a RNA Lab-On-A-Chip (Caliper Technologies Corp., Mountain View, CA) that was evaluated on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). Single-stranded, then double-stranded cDNA was synthesized from the poly(A)+ mRNA present in the isolated total RNA (typically 10 ug total RNA starting material each sample reaction) using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen Corp., Carlsbad, CA) and poly (T)-nucleotide primers that contained a sequence recognized by T7 RNA polymerase. A portion of the resulting ds cDNA was used as a template to generate biotin-tagged cRNA from an *in vitro* transcription reaction (IVT), using the Affymetrix GeneChip® IVT Labeling Kit. 15ug of the resulting biotin-tagged cRNA was fragmented to an average strand length of 100 bases (range 35-200 bases) following prescribed protocols (Affymetrix GeneChip® Expression Analysis Technical Manual). Subsequently, 10 ug of this fragmented target cRNA was hybridized at 45∞C with rotation for 16 hours (Affymetrix GeneChip® Hybridization Oven 640) to probe sets present on an Affymetrix HGU133 Plus 2.0 array. The GeneChip® arrays were washed and then stained (SAPE, streptavidin-phycoerythrin) on an Affymetrix Fluidics Station 450, followed by scanning on a GeneChip® Scanner 3000. The results were quantified and analyzed using GCOS 1.2 software (Affymetrix, Inc.) using default values (Scaling, Target Signal Intensity = 500; Normalization, All Probe Sets; Parameters, all set at default values).

Gastrocnemius muscle and liver of were harvested from male lean Zucker (fa/-) rats (Charles River, Wimington MA) obtained at 6 weeks of age and housed individually under controlled light (12:12 light:dark) and climate conditions. Six animals were

immediately sacked, while an additional six animals underwent the hyperinsulinemic-euglycemic clamp procedure following 12 hour fast as previously described[41]. Animals were exposed to hyperinsulinemic conditions for an average of 120 minutes, until glucose could be adequately controlled at 150mg/dL for at least 30 minutes. Harvested tissue was generously provided by Jason Wilkes and Jamie Resnik from animals of a forthcoming study (manuscript in preparation). Total RNA was purified from each tissue with the Fibrous Tissue Mini Kit (Qiagen). Equal amounts of RNA were pooled from each pair of rats. Labeling, amplification, and hybridization of pooled samples to Affymetrix RAE230A microarrays were performed as previously described[42].

Livers were harvested from C57/Bl6 mice catheterized for hyperinsulinemic-euglycemic clamp. Four days post-catheterization, four animals were sacked prior to insulin infusion. A total of eight additional animal subjects were infused with submaximal insulin (4 mU/kg/min) for a total of 90 or 240 minutes, with or without achieving steady-state euglycemia with concomitant dextrose infusion. Harvested tissue was generously provided by Philip Miles. Total RNA was purified as above, equal amounts of RNA were pooled from each pair of mice, and subsequently labeled as described above. The resulting cRNA was hybridized to a custom Affymetrix microarray chip designed for the Genomics Institute of the Novartis Research Foundation (GNF).

All procedures were performed in accordance with the *Guide for Care and Use of Laboratory Animals* of the National Institutes of Health and were approved by the University of California, San Diego, Animal Subjects Committee and Institutional Review Boards (IRB00000353, IRB00000354, IRB00000355, IRB00002758).

Microarray analysis.

Analysis of gene expression data was carried out with the VAMPIRE microarray analysis suite, accessible at http://genome.ucsd.edu/microarray [42-44]. Human skeletal muscle data was analyzed with a VAMPIRE paired test, while the VAMPIRE unpaired

test was performed for all remaining data. Gene annotation enrichment analysis was performed with GOby. A strict, Bonferroni-corrected significance threshold (α_{Bonf}=0.05) was applied for all statistical tests.

Conserved transcriptional targets

To provide an optimally-specific screen for insulin-regulated genes, we not only applied a stringent Bonferroni-corrected significance threshold during statistical analysis, but also selected for subsets of genes that were differentially-regulated in more than one model system. The presence of differential-expression in more than one system is strongly suggestive of conserved insulin regulation. The tissues available for comparison are displayed in table 5.1. For skeletal muscle, rats and human subjects were used to identify transcriptional targets. For liver, we obtained tissue from mice and rats. For comparisons of insulin-regulated gene expression in fat, we incorporated VAMPIRE-analysis of previously published IGF/PDGF murine myoblast data set[45] to supplement mouse and rat data collected in our lab.

D. Results

Effects of hyperinsulinemic-euglycemic clamp in skeletal muscle

Transcription factors are a major target of insulin-regulated expression

While a variety of known and yet uncharacterized transcription factors mediate the effects of insulin, we observed that insulin itself had a profound impact on the expression of many transcriptional regulators. In fact, in human skeletal muscle, the Gene Ontology (GO) term, *transcription*, was among the most significantly enriched annotation categories (p_{bonf}<0.05, q_{bonf}<0.05). A list of these transcriptional regulators and other enriched terms are shown as supplementary tables. Through this form of annotation analysis, it is possible to identify the physiological functions most affected by a biological perturbation, while accounting for the gene representation on the microarray platform.

Thus, enrichment is unlikely to occur simply because there are more transcription factors represented on the chip. These results suggest then, that in skeletal muscle, a major component of insulin activity is through its regulation of transcription factors. While these may not play a role acutely, over the course of hours, regulation of these transcripts can have a dominant effect.

Effect of hyperinsulinemia on cell proliferation and differentiation programs

Many transcriptional targets observed in human skeletal muscle were also observed to be differentially-regulated in Zucker fa/+ lean rats. Despite differences between rat and human physiology, technical differences between clamping procedures, and differences in microarray platforms, we found a distinct subset of genes which were differentially-expressed in both organisms. Among the genes coordinately regulated in both organisms were the transcription factors Cited2, Myod1, Nr1d1, and Nr4a3. These results are displayed in table 5.2. Perhaps the most prominent feature of this set of genes is the apparent effect of insulin on cell proliferation. Insulin dramatically upregulates Rrad *in vivo*, by over four-fold in both rats and human subjects. Rrad is a Ras-related GTPase whose expression is frequently lost in invasive breast cancer[46], and is implicated in the pathogenesis of malignant mesothelioma[47]. Its expression has been noted to be increased in human muscle during hyperinsulinemic-euglycemic clamp[48]. We observed similar results for other genes believed to behave as tumor suppressors. Ier3 (IEX-1) has been shown to enhance apoptosis [49, 50]. Its expression is upregulated over two-fold by insulin. The tumor suppressor gene, Tieg, is suppressed in both invasive and non-invasive forms of breast cancer [51]. Myod1 hypermethylation is associated with poorer outcomes in many forms of cancer, including cervical and colorectal[52, 53]. Gadd45a inhibits Cdc2 kinase to mediate cell cycle arrest[54]. In addition, the sole downregulated gene in this functional category, Cited2 (Mrg1), has transformation activity. Sun *et al* demonstrated this activity of Cited2 via overexpression in Rat1 cells[55]. This

combination of transcriptional responses is suggestive of an impaired state of skeletal muscle proliferation during the clamp.

In general, we observed upregulation of genes with tumor suppressor behavior and downregulation of genes with transformation activity. One potential exception to this pattern is the upregulation of TNFRSF12 (Fn14) in both humans and rats. Evidence for the influence of this gene on cell survival and proliferation is less clear. While it was originally described to influence apoptosis, it has also been observed to increase cell proliferation in certain cell types[56]. Nevertheless, the overall pattern of expression in this group of genes suggests a conserved role of insulin in suppressing cell proliferation in muscle tissue. In order to further assess the global transcriptional impact of insulin, and computationally address its roles, we examined more closely those genes regulated in rat skeletal muscle. Statistically enriched among these was the ìmuscle developmentî GO term (p_{bonf}<0.05, q_{bonf}<0.05). The apparent shift away from cell proliferation may therefore reflect an enhanced state of differentiation. Additional GO terms enriched in this data are shown in table 5.3, but all correlate to the upregulation of markers of skeletal muscle differentiation. To further explore the mechanism by which insulin mediates these effects, we subsequently examined a previously reported microarray data set describing the actions of a related ligand, insulin-like growth factor (IGF).

IGF has been known to promote skeletal muscle differentiation through its cell surface receptor. Insulin binds IGF receptors (IGFR) with modest affinity, and likely exerts some of its effects through these receptors. In addition, the insulin receptor can heterodimerize with IGFR to form hybrid receptors. A recent paper by Kuninger *et al* explored the transcriptional effects of 24 hour IGF-I treatment on an IGF-II-deficient murine myoblast cell line (C2AS12), and contrasted these with the proliferative effects of PDGF[45]. By VAMPIRE unpaired and subsequent gene annotation analysis on this data, we again discovered statistical enrichment of regulators and targets of skeletal muscle

differentiation. In fact, only GO terms related to muscle differentiation were statistically enriched (Table 5.4), and nearly all were upregulated. Furthermore, some of these murine targets were homologous to those we identified in rat skeletal muscle. The transcriptional profile of all homologous, differentially-regulated genes is shown in table 5.5. While genes involved in muscle differentiation were coordinately regulated by IGF in myoblasts and insulin in rat skeletal muscle, the same was not true for genes in other functional categories. Model-specific listings of this functional category, in both myoblasts and skeletal muscle, are provided as supplementary tables.

MyoD1 and Pde4b are both essential regulators of the muscle differentiation program [57-59]. While this transcriptional program is activated, regulators of skeletal muscle proliferation appear to be simultaneously inhibited. This is consistent with recently reported evidence that insulin can act as an inhibitor of growth-factor mediated cell proliferation[60]. Although it is not immediately clear what contribution of these actions are mediated by IGF or insulin receptors, or which cell types are ultimately responsive, this cross-species data suggest that a major *in vivo* role of insulin may be to regulate these transcriptional programs and facilitate energy storage in the form of skeletal muscle protein.

Transcriptional control of metabolism

Insulin acutely regulates metabolism through translocation of Glut4 and other transporters, and through post-translational modification of key metabolic enzymes. Although transcription does not typically operate on these time scales, we observed significant upregulation of several regulators of energy metabolism. When examining the shared profile of human and rat skeletal muscle we identified phosphodiesterase 4b (Pde4b), which metabolizes cAMP. In addition to its effects on skeletal muscle differentiation, by regulating cAMP levels, this transcriptional change may help desensitize skeletal muscle to the counter-regulatory influence of glucagon and

catecholamines. Downregulation of pyruvate dehydrogenase kinase, isotype 4 (Pdk4) was also seen in both models. Suppression of Pdk4 leads to release of its inhibitory influence on the pyruvate dehydrogenase complex, promotes glucose utilization, and increases flux into the citric acid cycle. In addition, while only reaching statistical significance in human muscle, hexokinase 2 was upregulated by over two-fold. Together with the expected influence of Pde4b and Pdk4, these observations are consistent with a state of enhanced glucose oxidation exhibited during the clamp.

Transcriptional control of signal transduction

Key mediators of insulin-mediated signal transduction were also found to be transcriptionally controlled (figure 5.1). The regulatory subunit of phosphatidylinositol 3-kinase, p85, was upregulated in response to hyperinsulinemia in insulin-sensitive human and rat models. Laville *et al* observed p85 transcriptional upregulation to a similar extent in human subjects [48]. As it has been reported that p85 deletion can lead to altered insulin-sensitivity[61-64], these observations suggest that a transcriptional feedback loop involving p85 may be a conserved regulator of insulin signal transduction. Since hyperinsulinemia itself can induce insulin-resistance, the upregulation of p85 may be an important mediator of this action. Other insulin signaling components, although not reaching statistical significance in both models, were also found to be insulin-regulated in skeletal muscle. Suppressor of cytokine signaling 3 (SOCS3), a gene implicated in TNF-α-mediated insulin-resistance[65], was upregulated nearly 8-fold in human muscle. The mechanism of insulin induction of SOCS3 expression has been studied in several model systems[66-68]. In addition, another gene currently known to be involved in insulin signaling, Grb2-associated binding protein 1 (GAB1), dropped 40% following clamp. While a liver-specific knockout of GAB1 was recently shown to enhance insulin sensitivity[69], it is not immediately clear whether a reduction in skeletal muscle GAB1 would share a similar effect.

Limitations of the comparative transcriptomics approach

SOCS3 and GAB1 are merely two examples of differentially-expressed genes that were not adequately detected in both human and mouse model systems, but likely mediate physiologically relevant effects of prolonged hyperinsulinemia. There are several factors that limit the sensitivity of a comparative transcriptomics approach. Microarrays themselves have limited sensitivity, depending on: (1) whether probes of sufficient quality are present on the chip, (2) whether gene expression is of sufficient intensity to be detected by these probes, (3) whether expression is sufficiently different between control and clamped subjects to be detected by the chosen statistical test. In addition, the quality of the statistical test and the current state of gene annotation knowledge will also contribute to the sensitivity of this high-throughput tool. Since it is difficult to decompose the contributions of each of these factors, negative results should not be interpreted as the absence of differential-regulation, but merely the absence of detection. A comparative approach between two organisms will be doubly affected by each of these issues. In addition, differences in the underlying genomes can further cloud comparisons. Despite only achieving statistical significance in the human or rat model then, it is still likely that these and many other single-model differential-responses represent important physiological actions of insulin. More complete listings these responses are provided as supplementary tables, but are ideally obtained by re-processing raw data with the statistical methods of choice.

Effects of hyperinsulinemic-euglycemic clamp in liver

Transcriptional regulation of gluconeogenesis

One of the most heavily-studied functions of insulin is its role in transcriptional regulation of gluconeogenesis. During the fed state, where insulin is elevated and glucagon is suppressed, two critical gluconeogenic enzymes are transcriptionally downregulated in the liver: phosphoenolpyruvate carboxykinase (Pck1) and glucose 6-

phosphatase (G6pc). These, and several other gluconeogenic enzymes, are regulated by a variety transcription factors. Through an inhibitory form of C/EBPβ, insulin inhibits Pck1 expression[70]. HNF-4α is also involved in insulinís negative transcriptional control of Pck1[71, 72]. Shen *et al* showed previously that gluconeogenic enzymes were elevated in liver in Foxa3$^{-/-}$ mice[73]. Stat3 mutants demonstrate a similar effect[74]. The peroxisome-proliferator activating receptor, gamma, coactivator 1 genes (PGC-1α, PGC-1β) have also been implicated in regulating these enzymes through interactions with HNF-4α and Foxo1[75-77]. While several factors involved in the transcriptional regulation of gluconeogenesis have been identified [78], the relationships between them and their mechanisms of regulation are not fully understood. It is also possible that additional unidentified factors and interactions still remain. Global transcriptional profiling of the liver therefore has the potential to identify new network components, and to bring insights into the regulatory roles of these transcription factors.

Interestingly, we not only observed the expected transcriptional changes in gluconeogenic enzymes Pck1, G6pc and the liver-specific glucose transporter Glut2[73] in response to hyperinsulinemic-euglycemic clamp, but also conserved regulation of several of the previously-mentioned transcription factors (table 5.6). C/EBPβ, Foxa3, HNF-4α, and Stat3 were each upregulated and reached statistical significance in both rat and mouse models of hyperinsulinemic-euglycemic clamp. These transcriptional regulators constituted a large fraction of the total list of conserved transcriptional responses, suggesting that the handful of additional genes that have not been characterized may also play equally important roles in the liver. In addition, we observed that while transcriptional activation of several factors achieved statistical significance after 240í clamp in mice, some changes were still within background noise at the earlier 90í time point. For PGC-1β, this relationship was reversed, suggesting that its role may occur earlier.

Insulin stimulates expression of inflammatory genes in liver

In each model system, GOby annotation analysis identified several inflammatory-related categories as the most significantly-enriched, including *response to biotic stimulus* (p_{bonf}<0.05). Several of the genes annotated by this GO term demonstrated coordinate regulation. These included several key players of the inflammatory response, including C-X-C motif chemokine ligand 1 (Cxcl1), type I interleukin 1 receptor (Il1r1), lipopolysaccharide binding protein (Lbp), C/EBPβ, and Stat3. Two of these, C/EBPβ and Stat3, have previously been mentioned in the context of glucose metabolism, indicating potential cross-talk between the inflammatory and gluconeogenic transcriptional networks. C/EBPβ is known to mediate LPS and cytokine-induction of acute phase proteins and inflammatory cytokines[79, 80], and Stat3 is a potent transcriptional activator of acute phase proteins[81] downstream of gp130/IL-6. In addition, in at least one model system, we observed that the TNF-α receptor (Tnfrsf1a), interleukin-1 beta (Il1b), components of the complement cascade, and many acute phase proteins themselves were upregulated as high as 80-fold. In fact, the vast majority of genes found in this functional category were upregulated during hyperinsulinemic-euglycemic clamp. Similar experiments, using in saline-clamped control mice, demonstrated similar upregulation of inflammatory mediators (unpublished data). It is likely then, that these inflammatory responses are independent of the clamp procedure itself. Furthermore, because the liver is the primary site of production for many of these inflammatory markers, these transcriptional changes reflect an enhancement of inflammatory activity during hyperinsulinemic-euglycemic clamp.

E. Discussion

We have demonstrated an approach for assessing the gross physiological impact of hyperinsulinemia in three mammalian systems based on transcriptional profiling. In skeletal muscle and liver, major transcriptional responses appear to be well-conserved

between model systems. While these organs are very different from each other, the same organs in different species are fairly similar. Where statistically significant, the transcriptional responses parallel this relationship. Very few genes showed opposing regulation between model systems. In addition to high levels of insulin, a variety of physiological changes occur during the hyperinsulinemic-euglycemic clamp, prior to our observation of transcriptional change. Thus, some transcriptional responses may be due to concomitant suppression of pancreatic glucagon secretion. The cumulative effect of these and other hormonal changes is an increase in skeletal muscle glucose uptake and suppression of hepatic glucose output. We demonstrate that outside of these gross metabolic functions, these tissues have several additional responses to this high-energy state.

In human patients, skeletal muscle appears to respond to prolonged hyperinsulinemia by modifying expression of transcriptional regulators. In response to clamp, conserved transcriptional changes in skeletal muscle appear to induce a program of differentiation while inhibiting proliferation. These events parallel the action of IGF in isolated myoblast cultures[45]. These results are also consistent with the observation that insulin co-treatment can inhibit PDGF-stimulated cell proliferation in NIH3T3 and C2C12 cells[60]. It is plausible then, that part of this differentiation program has been co-opted to support energy storage in the form of muscle protein. While aspects of skeletal muscle protein metabolism have been considered at the level of translational regulation, transcriptional control of this physiological process has not been well-explored. Furthermore, at the time scales of the glucose clamp, transcriptional effects may actually dominate. These data provide evidence for a novel mechanism of transcriptional control of skeletal muscle metabolism. In addition to these transcriptional programs, we also observed differential-expression of key regulators of insulin and glucagon signaling, including p85, SOCS3, GAB1, and Pde4b. While most of the genes identified in our screen can be accounted for

by these previously described functions, it is likely that the remaining uncharacterized targets play important roles as well.

The liver has considerably different transcriptional responses to insulin, reflective of its distinct role in mammalian physiology. It suppresses its gluconeogenic enzymes in the absence of glucagon, and modifies expression of key transcriptional regulators. From a core set of conserved, insulin-regulated genes, we were able to reconstruct a transcriptional network involved in the regulation of gluconeogenesis (figure 5.2). The expression and activity of these regulators takes a dynamic course. While some key genes may be detectably activated at a late time point, they may not necessarily be detectable at earlier time points. It is important to note that networks involving these factors are extremely complex, controlled by a variety of non-transcriptional mechanisms, which are not easily captured by a transcriptional network alone. For example, this transcriptional network does not consider insulin-regulated nuclear exclusion of Foxo transcription factors through PKB/Akt. Despite the coarseness of this approach however, it is clearly a valuable tool for defining novel mechanisms of insulin action. With prolonged stimulation, key transcriptional regulators may feedback and drive transition to new transcriptional steady states. It is perhaps because of this phenomenon, at longer time scales, that gene expression is able to provide insights into the mechanisms of its own regulation.

Downstream of these transcriptional regulators are not only gluconeogenic enzymes, however. Inflammatory genes, especially cytokines and acute phase proteins, are known targets of C/EBPβ. Through interleukin-6 and gp130, Stat3 is a potent transcriptional activator of acute phase proteins[81]. Both C/EBPβ and Stat3 are significantly upregulated in liver with prolonged insulin-stimulation *in vivo*. While energy metabolism and inflammation have historically been distant fields of investigation, recent evidence has brought these disparate worlds together. Several groups have begun to delineate the

crosstalk between the signal transduction networks of cytokine and insulin signaling pathways. Liver-specific Stat3 overexpression was recently shown to inhibit gluconeogenesis *in vivo*[74]. In a recent review, Hotamisligil observed that the development of the *Drosophila* fat body paralleled the developmental program of adipose, liver, hematopoetic/immune organs[82]. He postulated that the shared ancestry of these organs may be partly responsible for overlap the functional and molecular pathways in each of these tissues. On a similar note, we suggest that transcriptional networks governing metabolic and immune functions may not have adequately diverged to handle prolonged states of hyperinsulinemia. It is an interesting possibility then, that crosstalk between these transcriptional networks may contribute to the increased inflammatory state during pathologic states of hyperinsulinemia and insulin-resistance.

As biological knowledge databases further improve in their ability to systematically organize cell signaling and transcriptional networks, statistical and computational tools for integrating this knowledge will be increasingly valuable for mining high-throughput biological data. The interpretation of gene expression data will consequently need to evolve with additional insights into control at the transcriptional, translational, and post-translational level. At the present, these aspects of biological annotation are not adequately handled by existing databases. Despite the limitations of current systems however, the enrichment analysis approach we present here clearly has great utility, and will likely be invaluable for mining high-throughput data. Already, the human genome and related genomics projects are playing critical roles in shaping our study of biology and human disease. In the same way, because of the richness and accessibility of high-throughput transcriptomics data, it also has the potential to serve as a living body of biological knowledge, which should be constantly revisited and reinterpreted.

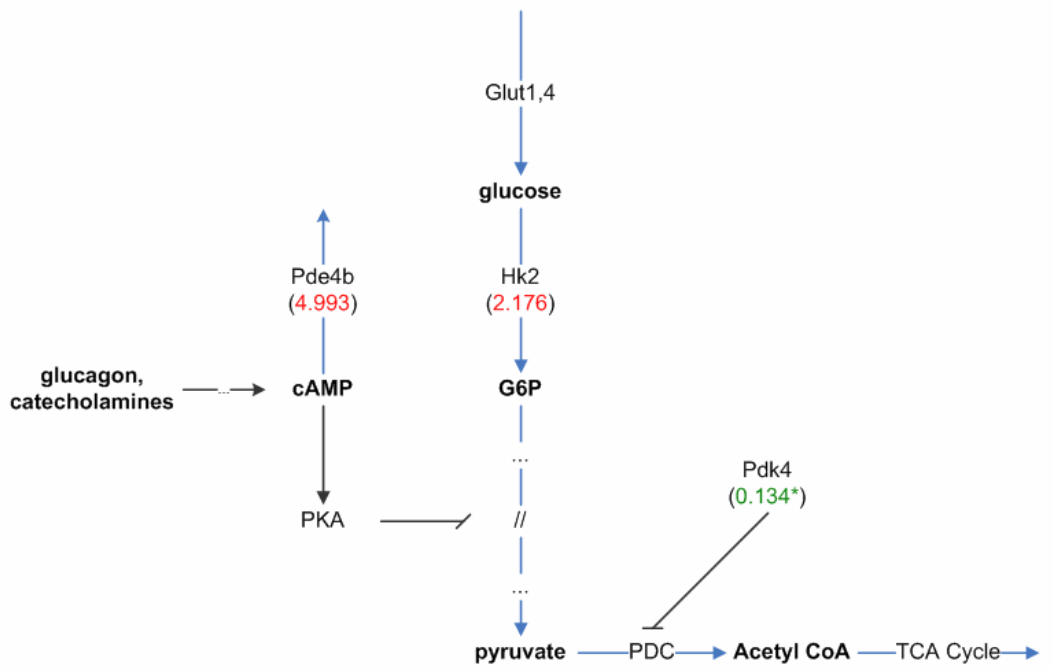


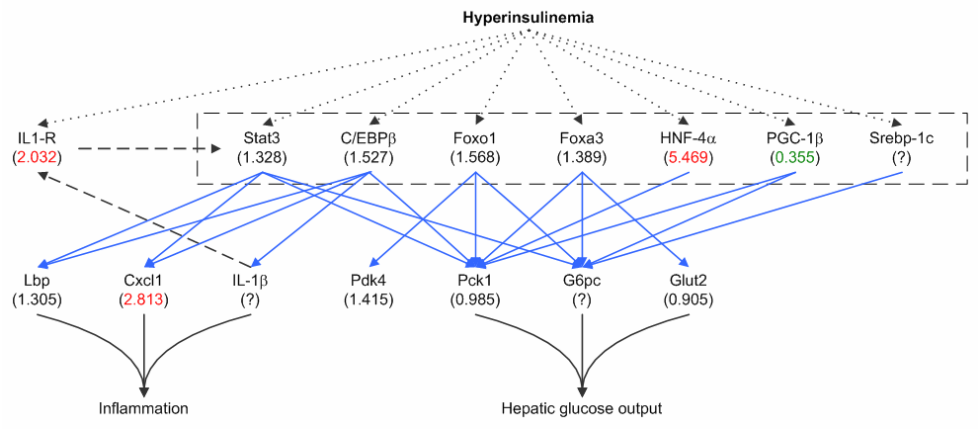
FIGURE 5.1

Key metabolic regulatory genes in skeletal muscle are transcriptionally controlled. The transcriptional responses mirror a state of increased glucose oxidation. Numbers shown represent statistically significant fold changes in human skeletal muscle. The asterisk indicates that the fold-change displayed is an average of multiple significant probe sets.

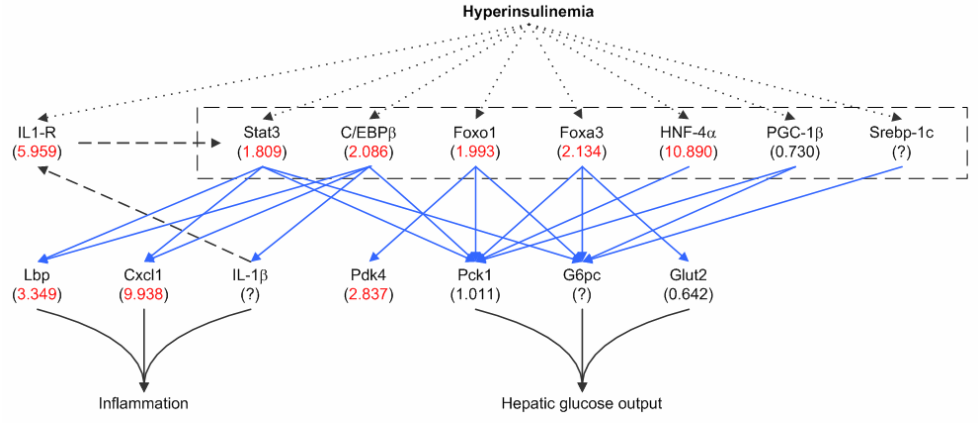
FIGURE 5.2

Reconstruction of the transcriptional network downstream of hyperinsulinemic-euglycemic clamp. Fold changes for (a) 90í clamp in mouse, (b) 240í clamp in mouse, or (c) rat clamp are shown in parentheses. Statistically significant changes in gene expression are colored in red if upregulated, or green if downregulated. Fold-changes that did not exceed background noise are in black text. Genes for which no measurements were obtained are marked by question marks. Blue, solid lines indicate known transcriptional regulatory connections. Dashed lines demonstrate some interconnections between transcriptional targets. No directionality is implied by either of these lines.

(a)


Hyperinsulinemia
IL1-R
(2.032)
Stat3
(1.328)
C/EBPβ
(1.527)
Foxo1
(1.568)
Foxa3
(1.389)
HNF-4α
(5.469)
PGC-1β
(0.355)
Srebp-1c
(?)
Lbp
(1.305)
Cxcl1
(2.813)
IL-1β
(?)
Pdk4
(1.415)
Pck1
(0.985)
G6pc
(?)
Glut2
(0.905)
Inflammation
Hepatic glucose output

(b)


Hyperinsulinemia
IL1-R
(5.959)
Stat3
(1.809)
C/EBPβ
(2.086)
Foxo1
(1.993)
Foxa3
(2.134)
HNF-4α
(10.890)
PGC-1β
(0.730)
Srebp-1c
(?)
Lbp
(3.349)
Cxcl1
(9.938)
IL-1β
(?)
Pdk4
(2.837)
Pck1
(1.011)
G6pc
(?)
Glut2
(0.642)
Inflammation
Hepatic glucose output

(c)

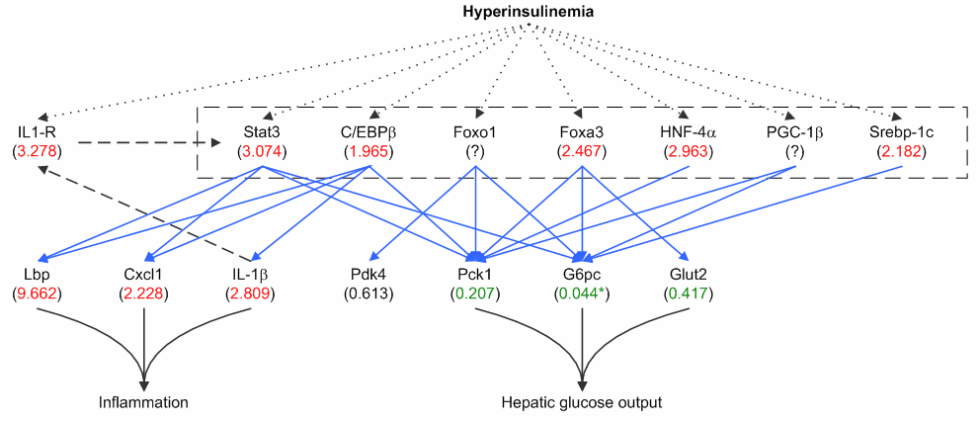

Hyperinsulinemia
IL1-R
(3.278)
Stat3
(3.074)
C/EBPβ
(1.965)
Foxo1
(?)
Foxa3
(2.467)
HNF-4α
(2.963)
PGC-1β
(?)
Srebp-1c
(2.182)
Lbp
(9.662)
Cxcl1
(2.228)
IL-1β
(2.809)
Pdk4
(0.613)
Pck1
(0.207)
G6pc
(0.044*)
Glut2
(0.417)
Inflammation
Hepatic glucose output

TABLE 5.1

Novel transcriptional profiling data collected for comparative analysis. Hyperinsulinemic-euglycemic clamp was performed for mouse, rat, and human for the time periods indicated.

	Mouse	**Rat**	**Human**
Muscle		x	x
Liver	x	x	

TABLE 5.2

Conserved targets of hyperinsulinemic-euglycemic clamp in skeletal muscle. Fold changes for transcriptional changes are displayed in the rightmost columns. Only statistically-significant fold-changes are displayed (α_{bonf}=0.05). Fold-changes marked by an asterisk represent the average change of more than one significant probe set on the array. Fold changes displayed were not necessarily obtained in the linear range of the microarray system.

locus	description	refs	human	rat
Cell Proliferation				
Cited2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	[55]	0.382 *	0.218 *
Tumor Suppressors				
Ier3	immediate early response 3	[49, 50]	2.256	2.527
Tieg	TGFB inducible early growth response	[51]	2.242	2.409
Gadd45a	growth arrest and DNA-damage-inducible 45 alpha	[54]	3.227	4.195
Rrad	Ras-related associated with diabetes	[46, 48]	4.519 *	103.7
Myod1	myogenic differentiation 1	[52, 53, 57]	2.218	2.454
Metabolic Signaling				
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha)	[48, 61-64]	2.719	2.547
Pde4b	phosphodiesterase 4B	[40]	4.993 *	2.264 *
Other				
Col1a2	procollagen, type I, alpha 2		0.556	0.569
Nr1d1	nuclear receptor subfamily 1, group D, member 1		0.349 *	0.562
Zfp36	zinc finger protein 36		3.909	0.503
Ankrd1	ankyrin repeat domain 1 (cardiac muscle)		2.087	7.227 *
TNFRSF12A	tumor necrosis factor receptor superfamily, member 12A		4.002	4.955
Nr4a3	nuclear receptor subfamily 4, group A, member 3		4.997 *	3.656

TABLE 5.3

Enriched Gene Ontology annotations in rat skeletal muscle following 1 hour clamp. Only terms achieving statistical significance are shown (p_{bonf}<0.05, q_{bonf}<0.05). Nearly all of the genes in these functional categories were found to be upregulated.

ID	description	genes	p-value	q-value
GO:0006936	muscle contraction	16	**0.00000000**	**0.00000000**
GO:0005622	intracellular	107	**0.00000002**	**0.00000004**
GO:0007517	muscle development	18	**0.00000000**	**0.00000005**
GO:0030484	muscle fiber	10	**0.00000002**	**0.00000565**

TABLE 5.4

Enriched Gene Ontology terms in 24hr IGF-treated myoblasts. Only terms achieving statistical significance are shown (p_{bonf}<0.05, q_{bonf}<0.05). Nearly all of the genes in these functional categories were found to be upregulated.

ID	description	genes	p-value	q-value
GO:0006936	muscle contraction	13	**0.00000000**	**0.00000000**
GO:0030484	muscle fiber	8	**0.00000000**	**0.00000000**
GO:0005861	troponin complex	5	**0.00000000**	**0.00000000**
GO:0007517	muscle development	13	**0.00000000**	**0.00000007**
GO:0005200	structural constituent of cytoskeleton	8	**0.00000003**	**0.00000241**
GO:0051239	regulation of organismal physiological process	8	**0.00000025**	**0.00000360**
GO:0005856	cytoskeleton	12	**0.00000439**	**0.00000556**
GO:0016529	sarcoplasmic reticulum	3	**0.00001002**	**0.00000863**

TABLE 5.5

Ligand-mediated regulation of genes in murine myoblasts and rat gastrocnemius. Only homologous genes differentially-regulated in both systems (α_{Bonf}<0.05) are shown. Genes marked by an asterisk display the average response of more than one probe set. Fold changes in bold colors represent statistically-significant changes in gene expression.

locus	description	pdgf	igf	rat
Muscle development				
Tnni1*	troponin I, skeletal, slow 1	0.952	**5.068**	**3.822**
Myh3	myosin, heavy polypeptide 3, skeletal muscle, embryonic	0.616	**11.878**	**27.794**
Csrp3	cysteine and glycine-rich protein 3	0.554	**10.811**	**4.105**
Other				
Fdps	farnesyl diphosphate synthetase	2.631	**3.433**	**0.504**
Pgam2	phosphoglycerate mutase 2	0.672	**8.861**	**0.524**

TABLE 5.6

Conserved targets of hyperinsulinemic-euglycemic clamp in liver. Fold changes for transcriptional changes are displayed in the rightmost columns. Fold changes in bold colors achieved statistical significance (α_{bonf}=0.05), while non-bolded numbers were within statistical error. A general trend of greater statistical significance can be seen at the 240í time point. In addition, fold-changes marked by an asterisk represent the average change of more than one significant probe set on the array.

locus	Description	ref	m 90	m 240	rat
Tumor suppressors					
Dusp6	dual specificity phosphatase 6	[83]	1.245	**1.862**	**2.978** *
Btg1	B-cell translocation gene 1, anti-proliferative	[84]	1.208	**1.972**	**1.904**
Btg2	B-cell translocation gene 2, anti-proliferative	[84]	**2.341**	**3.981**	**3.599**
Transcription factors					
Foxa3	forkhead box A3	[73]	1.389	**2.134**	**2.467**
Stat3	signal transducer and activator of transcription 3	[74, 85]	1.328	**1.809**	**3.074**
Cebpb	CCAAT/enhancer binding protein (C/EBP), beta	[37, 70]	**2.074**	**2.935**	**1.965**
Hnf4a	hepatic nuclear factor 4, alpha	[71, 72]	**5.469**	**10.890**	**2.963** *
Gluconeogenesis					
Car5a	carbonic anhydrase 5a, mitochondrial	[86]	0.711	**0.547**	**0.41**
Inflammation					
Cxcl1	chemokine (C-X-C motif) ligand 1	[87, 88]	**2.813**	**9.938**	**2.228**
Il1r1	interleukin 1 receptor, type I	[89]	**2.032**	**5.959**	**3.278**
Lbp	lipopolysaccharide-binding protein	[90, 91]	1.305	**3.349**	**6.900**
Other					
9430041C03Rik	UDP-glucuronosyltransferase 2B3 precursor, microsomal		0.490	**0.362**	**0.486**
Fdft1	farnesyl diphosphate farnesyl transferase 1		0.749	**0.457**	**0.404**
Arf6	ADP-ribosylation factor 6		1.254	**1.815**	**1.854**
Hspca	heat shock protein 1, alpha		1.498	**2.089**	**2.247**
Csda	cold shock domain protein A		1.454	**2.187**	**2.683**
Hspb8	heat shock 27kDa protein 8 / crystallin, alpha C		1.432	**3.194**	**1.974**
Dusp1	dual specificity phosphatase 1		**2.017**	**2.040**	**0.371** *
Egr1	early growth response 1		**2.485**	**3.400**	**3.355**
Atp1b1	ATPase, Na+/K+ transporting, beta 1 polypeptide		**2.327**	**4.145**	**2.605**
Hmox1	heme oxygenase (decycling) 1		**2.156**	**17.317**	**2.627**

TABLE 5.7

Supplementary table: Gene Ontology (GO) terms enriched by hyperinsulinemic-euglycemic clamp in human skeletal muscle. The column labeled *n* lists the number of probe sets annotated with the GO term that were found to be differentially-expressed. The *P*-values and *Q*-values shown represent the global and conditional probabilities of enrichment. Probabilities displayed in bold text were considered statistically significant after Bonferroni correction ($\alpha_{Bonf}<0.05$).

group id	group name	n	p-value	q-value
GO:0004879	ligand-dependent nuclear receptor activity	10	**1.39E-08**	**4.96E-08**
GO:0046870	cadmium ion binding	3	**1.35E-06**	**2.77E-06**
GO:0019222	regulation of metabolism	55	**2.88E-08**	**4.13E-06**
GO:0008147	structural constituent of bone	3	**1.59E-05**	**9.87E-06**
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	54	**6.3E-09**	**1.88E-05**
GO:0030528	transcription regulator activity	38	**1.76E-07**	**2.12E-05**
GO:0000185	activation of MAPKKK	3	**1.59E-05**	**4.61E-05**
GO:0050789	regulation of biological process	71	**1.06E-08**	**6.31E-05**
GO:0005507	copper ion binding	5	**3.24E-05**	8.52E-05
GO:0004740	[pyruvate dehydrogenase (lipoamide)] kinase activity	3	**1.59E-05**	8.89E-05
GO:0006350	transcription	54	**1.88E-08**	0.000157
GO:0048511	rhythmic process	5	**5.47E-05**	0.000172
GO:0050791	regulation of physiological process	61	**4.49E-07**	0.000392
GO:0005634	nucleus	72	**1.33E-06**	0.002543
GO:0048513	organ development	27	**2.38E-05**	0.0028
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	61	**1.71E-05**	0.003398
GO:0009653	morphogenesis	31	**4.57E-05**	0.004645
GO:0009649	entrainment of circadian clock	3	**7.46E-06**	0.004878
GO:0003714	transcription corepressor activity	8	**5.38E-05**	0.00495
GO:0043231	intracellular membrane-bound organelle	88	**1.7E-05**	0.007283
GO:0043227	membrane-bound organelle	88	**1.7E-05**	0.007283
GO:0003677	DNA binding	48	**3.53E-06**	0.007342
GO:0005488	binding	135	**8E-07**	0.008655
GO:0005584	collagen type I	3	**1.11E-05**	0.016935
GO:0003700	transcription factor activity	28	**4.02E-06**	0.044902
GO:0006351	transcription, DNA-dependent	54	**3.6E-09**	0.07172
GO:0006355	regulation of transcription, DNA-dependent	53	**3.3E-09**	0.148977
GO:0009887	organogenesis	27	**2.38E-05**	0.160464
GO:0050875	cellular physiological process	139	**2.61E-05**	0.243909
GO:0045449	regulation of transcription	53	**9.7E-09**	0.325879
GO:0003707	steroid hormone receptor activity	10	**5.8E-09**	0.391611
GO:0007582	physiological process	149	**4.06E-05**	0.786918
GO:0009987	cellular process	156	**1.59E-05**	0.796459
GO:0003674	molecular_function	179	**8.39E-06**	N/A
GO:0008150	biological_process	175	**3.89E-07**	N/A

TABLE 5.8

Supplementary table: Genes involved in *transcription* significantly downregulated by hyperinsulinemic-euglycemic clamp in human skeletal muscle (α_{Bonf}=0.05).

accn	locus	description	fold
NM_173173	NR4A2	nuclear receptor subfamily 4, group A, member 2	**0.165**
NM_173173	NR4A2	nuclear receptor subfamily 4, group A, member 2	**0.197**
NM_173173	NR4A2	nuclear receptor subfamily 4, group A, member 2	**0.214**
NM_021724	NR1D1	nuclear receptor subfamily 1, group D, member 1	**0.289**
NM_002616	PER1	period homolog 1 (Drosophila)	**0.29**
NM_016831	PER3	period homolog 3 (Drosophila)	**0.304**
NM_006079	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	**0.359**
NM_006079	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	**0.404**
NM_002616	PER1	period homolog 1 (Drosophila)	**0.407**
NM_021724	NR1D1	nuclear receptor subfamily 1, group D, member 1	**0.408**
NM_002469	MYF6	myogenic factor 6 (herculin)	**0.418**
NM_002616	PER1	period homolog 1 (Drosophila)	**0.425**
NM_014079	KLF15	Kruppel-like factor 15	**0.428**
NM_058176	HDAC9	histone deacetylase 9	**0.462**
NM_005126	NR1D2	nuclear receptor subfamily 1, group D, member 2	**0.494**
NM_005126	NR1D2	nuclear receptor subfamily 1, group D, member 2	**0.494**
NM_005077	TLE1	transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)	**0.516**
NM_005077	TLE1	transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)	**0.524**
NM_001352	DBP	D site of albumin promoter (albumin D-box) binding protein	**0.543**
NM_182524	ZNF595	zinc finger protein 595	**0.596**

TABLE 5.9

Supplementary table: Genes involved in *transcription* significantly upregulated by hyperinsulinemic-euglycemic clamp in human skeletal muscle (α_{Bonf}=0.05).

accn	locus	description	fold
NM_002467	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	9.774
NM_173199	NR4A3	nuclear receptor subfamily 4, group A, member 3	7.159
NM_005967	NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	6.23
NM_152878	MAFF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian)	6.133
NM_005384	NFIL3	nuclear factor, interleukin 3 regulated	4.224
NM_005524	HES1	hairy and enhancer of split 1, (Drosophila)	3.435
NM_005967	NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	3.293
NM_005239	ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	2.929
NM_173199	NR4A3	nuclear receptor subfamily 4, group A, member 3	2.834
NM_005178	BCL3	B-cell CLL/lymphoma 3	2.701
NM_012323	MAFF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian)	2.644
NM_013376	SERTAD1	SERTA domain containing 1	2.629
NM_130469	JDP2	jun dimerization protein 2	2.609
NM_005239	ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	2.599
NM_005524	HES1	hairy and enhancer of split 1, (Drosophila)	2.46
NM_002015	FOXO1A	forkhead box O1A (rhabdomyosarcoma)	2.316
NM_024336	IRX3	iroquois homeobox protein 3	2.251
NM_005655	KLF10	Kruppel-like factor 10	2.242
NM_002478	MYOD1	myogenic factor 3	2.218
NM_002229	JUNB	jun B proto-oncogene	2.194
NM_001300	KLF6	Kruppel-like factor 6	2.169
NM_001964	EGR1	early growth response 1	2.151
NM_000964	RARA	retinoic acid receptor, alpha	2.132
NM_003670	BHLHB2	basic helix-loop-helix domain containing, class B, 2	2.106
NM_005904	SMAD7	SMAD, mothers against DPP homolog 7 (Drosophila)	2.038
NM_005414	SKIL	SKI-like	2.011
NM_001546	ID4	inhibitor of DNA binding 4, dominant negative helix-loop-helix protein	2.006
NM_015995	KLF13	Kruppel-like factor 13	1.986
NM_001730	KLF5	Kruppel-like factor 5 (intestinal)	1.979
NM_002015	FOXO1A	forkhead box O1A (rhabdomyosarcoma)	1.977
NM_001964	EGR1	early growth response 1	1.94
NM_001964	EGR1	early growth response 1	1.907
NM_001300	KLF6	Kruppel-like factor 6	1.892
NM_003670	BHLHB2	basic helix-loop-helix domain containing, class B, 2	1.891

TABLE 5.10

Supplementary table: Genes involved in *muscle development* significantly regulated by hyperinsulinemic-euglycemic clamp in rat skeletal muscle (α_{Bonf}=0.05). The Gene Ontology term, *muscle development* was statistically-enriched among the set of genes differentially-expressed during hyperinsulinemic-euglycemic clamp (p_{bonf}<0.05, q_{bonf}<0.05)

accn	locus	description	fold
NM_017131	Casq2	calsequestrin 2	2.632
NM_017131	Casq2	calsequestrin 2	3.28
NM_057144	Csrp3	cysteine-rich protein 3	4.105
NM_145669	Fhl1	four and a half LIM domains 1	2.172
NM_019242	Ifrd1	interferon-related developmental regulator 1	4.105
NM_012604	Myh3	myosin, heavy polypeptide 3, skeletal muscle, embryonic	27.794
NM_017239 NM_017240	Myh6 Myh7	myosin heavy chain, polypeptide 6 myosin, heavy polypeptide 7, cardiac muscle, beta	7.128
NM_017240	Myh7	myosin, heavy polypeptide 7, cardiac muscle, beta	2.83
NM_017240	Myh7	myosin, heavy polypeptide 7, cardiac muscle, beta	5.419
NM_012606	Myl3	myosin, light polypeptide 3	2.254
NM_176079	Myod1	myogenic differentiation 1	2.454
NM_022499	Pvalb	parvalbumin	0.568
NM_057132	Rhoa	plysia ras-related homolog A2	0.443
NM_057132	Rhoa	plysia ras-related homolog A2	0.539
NM_017184	Tnni1	troponin I, skeletal, slow 1	3.596
NM_017184	Tnni1	troponin I, skeletal, slow 1	4.048
NM_019131	Tpm1	tropomyosin 1, alpha	1.974
NM_057208	Tpm3	tropomyosin 3, gamma	2.663

TABLE 5.11

Supplementary table: Genes involved in *muscle development* significantly regulated by 24-hour IGF treatment in C2AS12 murine myoblasts. The Gene Ontology term, *muscle development* was statistically-enriched among the set of genes differentially-expressed during IGF treatment (p_{bonf}<0.05, q_{bonf}<0.05)

accn	locus	description	fold
M74753	Myh3	myosin, heavy polypeptide 3, skeletal muscle, embryonic	11.878
NM_013808	Csrp3	cysteine and glycine-rich protein 3	10.811
NM_009394	Tnnc2	troponin C2, fast	8.113
NM_016749	Mybph	myosin binding protein H	8.078
NM_010867	Myom1	myomesin 1	6.799
NM_013593	Mb	myoglobin	6.66
NM_009608	Actc1	actin, alpha, cardiac	6.25
NM_011619	Tnnt2	troponin T2, cardiac	6.165
NM_009393	Tnnc1	troponin C, cardiac/slow skeletal	5.8
NM_011620	Tnnt3	troponin T3, skeletal, fast	5.395
NM_021467	Tnni1	troponin I, skeletal, slow 1	5.068
NM_009606	Acta1	actin, alpha 1, skeletal muscle	4.839
NM_010858	Myl4	myosin, light polypeptide 4	4.553

CHAPTER VI

Conclusion

A. Overview

In this text, we introduced a novel framework for the interpretation of high-throughput functional genomics data. This Bayesian statistical construction, VAMPIRE, takes advantage of the error structure of array data to identify an estimate of noise that is more robust at low-replicate numbers than standard deviation, and uses this to identify statistically significant differences across groups of samples. We introduce two statistical tests based on this framework for (1) unpaired array data and (2) paired array data. We also implemented these tools in a web-based application to make these them more widely accessible. To facilitate interpretation and discovery in this high-throughput data, we developed a novel tool for integrating biological annotation systems. This application, GOby, is applicable to both flat and hierarchical systems of annotation such as KEGG and GO. It too has been integrated into a web-based application. While VAMPIRE analysis provides insight into the individual transcriptional targets of a biological perturbation, GOby aggregates these genes into functional categories to isolate statistically enriched physiological functions. We subsequently demonstrate the applicability of these computational and statistical tools to study an important fundamental question in mammalian physiology and human disease.

In particular, we investigate the primary physiological roles of insulin through comparative transcriptomics. Through a model of insulin action during the fed state, the hyperinsulinemic-euglycemic clamp, we characterized the dominant transcriptional responses in skeletal muscle and liver. Skeletal muscle is typically considered the primary site of insulin-stimulated glucose uptake, and liver the primary source of glucose production. Insulinís regulation of transcription factors dominated the conserved responses of mouse, human, and rat tissue, suggesting that transcriptional networks may

play a major role in meditating glucose homeostasis. Only a few signal transduction components were consistently altered by prolonged hyperinsulinemic-euglycemic clamp, including p85 and Pde4b. These results provide evidence for the theory that at the time scale of the clamp, physiological control is primarily mediated not through signal transduction and acute changes in compartmentalization, but through changes in transcriptional state.

In skeletal muscle, we observe that transcriptional regulators of cell proliferation and muscle differentiation are overwhelmingly prominent. While somewhat unexpected, this idea is consistent with established ideas concerning insulin action. Stimulation of a differentiation program in skeletal muscle may serve to promote energy storage in the form of muscle protein. Previously, this aspect of insulin action was mostly studied through regulation of the translational apparatus. In addition, insulin has recently been shown to have inhibitory effects on growth-factor stimulated proliferation. Because of the dichotomy between programs of cellular proliferation and differentiation, and the role of insulin-like growth factors in promoting skeletal muscle differentiation, it is ultimately not surprising that insulin may act in this manner. We present here then, a promising alternative mechanism for insulin-induced metabolic changes in skeletal muscle. Thus, while transcriptional regulation of the gluconeogenic program has only been widely studied in the liver, metabolic functions in skeletal muscle may also be mediated through transcriptional programs.

Transcriptional responses in the liver are remarkably different from those of skeletal muscle. In this tissue, we observe control of gluconeogenesis with spillover into the inflammatory transcriptional network. While the list of conserved targets is brief, a surprisingly large proportion of these have already been characterized as key regulators of gluconeogenesis. Several of these are also important regulatory components of a seemingly unrelated program, inflammation. We explicitly describe C/EBPβ and Stat3 as

highly-conserved insulin-regulated targets which act simultaneously to control transcription in both processes. We speculate that coordinate regulation of these disparate pathways may be due to a common developmental origin of liver and immune cells. Although these tissues and networks have diverged to perform distinct physiological roles, they may have not sufficiently diverged to support states of hyperactivation, such as the state of prolonged hyperinsulinemia. In pathological states of hyperinsulinemia, such as those found in early stages of insulin-resistance and type 2 diabetes, increased inflammation may be partially due to this phenomenon.

B. Future Directions

In the course of these studies, we have not only established novel methods for interpretation of high-throughput functional genomics data, but identified several new directions for future work. It is evident that the linear error model and the modeling procedures utilized in VAMPIRE provide a reliable indicator of noise. In some situations, particularly in low-replication high-throughput data, it can perform better than standard deviation. We have demonstrated this quantitatively in a simulated data set, and qualitatively in several biological data sets. Statistical significance in our Bayesian framework correlates well qualitatively with conventional fold-change approaches, while limiting ìinvestigator biasî commonly present in those techniques. Ironically though, the greatest strength of this parametric approach is also its greatest weakness. The unpaired and paired statistical tests presented here cannot be generally applied to all statistical questions that may arise. Additional mathematical derivations are necessary. For example, consider the reference-designed microarray experiment. In this experiment, RNA from control and treated experiments are hybridized to different two-channel microarrays, while a common reference is hybridized to all. Without additional normalization, neither of the two tests we described can precisely handle this statistical question. In addition to these two-class comparison problems, there are several categories

of additional statistical tests that should be derived. Multiple-class comparisons, for which ANOVA is commonly used, may further benefit from adaptations of the VAMPIRE framework. Regression tests and clustering algorithms, which do not account for the heteroskedasticity inherent in microarray data, can also be equivalently modified to provide more reliable results.

At the level of gene annotation analysis, the most practical line of further work may be the development of highly-refined annotation databases and statistical algorithms to make use of them. These databases should precisely describe the spectrum of splice variants, translational control, post-translational modification, and cellular localization processes which modify the activity of each gene. They should also extend beyond current concepts of biological processes and functions to consider factors affecting compartmentalization, sources of functional modification, and directionality (activation/repression). A considerable amount of study has been devoted to these areas of biology, and can be systematically organized. Without such organization of this knowledge, interpretation of high-throughput functional genomics data will continue to require ìexpert eyesî, with broad training in many areas of biological research. While this will often be the case at the forefront of research, a substantial amount of automation is possible, and will serve to accelerate the pace of new discoveries. The tools we present here are only glimpses of this possibility.

Other forms of high-throughput biological data are rapidly being developed. Promoter arrays have the potential to more directly address the numerous interconnections between each transcription factor and their target genes. They have additional potential for identifying important co-regulatory factors in a position-specific manner across the entire genome. The statistical considerations involved in interpreting this kind of data highly similar to those of cDNA and oligonucleotide arrays. As these technologies are developed

and integrated with expression data, we may begin to address higher order properties of this highly connected biological network.

In this text, we have also alluded to continuing work in studying the relationships between various models of insulin-resistance, and the behavior of adipose tissue, liver and skeletal muscle in the insulin-resistant state. Each tissue has characteristic metabolic abnormalities, whose mechanisms are not well understood. In insulin-resistant skeletal muscle, insulin-stimulated glucose uptake is greatly impaired. In adipose and liver, insulin does not efficiently suppress lipolysis and hepatic glucose production, respectively. Within each tissue, it has not been determined what contribution primary defects in post-receptor signaling, accumulation of metabolites, or dysregulation of transcriptional networks have in this pathological state. Between tissues, it is unclear whether additional, yet unidentified hormones or metabolites somehow antagonize normal metabolic regulation. Additional transcriptional studies of tissues in each of these models will likely address some of these questions.

With the bioinformatics tools currently available, we have been able to identify some key interactions between the transcriptional regulatory networks governing gluconeogenic and inflammatory pathways. Along with the transcriptional evidence we provide, these overlaps suggest that the hyperinsulinemic state can itself induce production of acute phase proteins indicative of an inflammatory response. Several recent studies have begun to tie infiltration of inflammatory cells and cytokine-induced dysregulation of signaling to insulin-resistance [67, 92, 93]. Future works should address the effect of hyperinsulinemia in both normal and insulin-resistant states. In addition, one major tissue we have not addressed here is adipose tissue. Parallel experiments may help determine the role of insulin transcriptional regulation in adipose tissue. As adipose has been proposed to be a primary regulator of energy metabolism, and secretes a spectrum of

cytokines and adipokines to influence other tissues, the results of these experiments may be enlightening.

We have, in the course of this text, also alluded to the use of insulin-sensitizing thiazolidinediones (TZDs) to further understand the pathologic state of insulin-resistance. While these drugs are commonly used to improve insulin-resistance in diabetic patients, a substantial proportion of subjects do not respond to treatment. Other patients exhibit atypical edema and other untoward side-effects. Because of these limitations, it is of considerable scientific and economic interest to further understand the underlying mechanisms of TZD action. While PPAR-γ is expressed most strongly in adipose, skeletal muscle and liver also demonstrate improvements in insulin-sensitivity following treatment. Transcriptional profiling for differences between patients before and after TZD-treatment, whether absolute expression or differences in expression following hyperinsulinemic-euglycemic clamp, may provide some insight here. In addition, we can investigate the underlying differences between responsive and non-responsive patients. This may assist not only in identifying characteristics of receptive populations, but also in understanding the concomitant conditions that prevent non-responsive populations from improving. A variety of mechanisms have been proposed as primary defects that cause insulin resistance, such as infiltration of immune mediators and lipid accumulation in peripheral tissues. While each of these likely play some role, TZDs and related drugs may give us an opportunity to tease apart and characterize the contributions of each of these mechanisms.

REFERENCES

1. Braunwald, et al., *Harrison's Principles of Internal Medicine*. 15 ed. 2001, New York: McGraw-Hill.
2. Elbein, S.C., *Perspective: the search for genes for type 2 diabetes in the post-genome era.* Endocrinology, 2002. **143**(6): p. 2012-8.
3. Sanders, L.J., *From Thebes to Toronto and the 21st Century: An Incredible Journey.* Diabetes Spectr, 2002. **15**(1): p. 56-60.
4. Fodor, S.P., et al., *Multiplexed biochemical assays with biological chips.* Nature, 1993. **364**(6437): p. 555-6.
5. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
6. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
7. Sasik, R., E. Calvo, and J. Corbeil, *Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model.* Bioinformatics, 2002. **18**(12): p. 1633-1640.
8. Consortium, T.G.O., *Creating the Gene Ontology Resource: Design and Implementation.* Genome Res., 2001. **11**(8): p. 1425-1433.
9. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucl. Acids Res., 2004. **32**(90001): p. D277-280.
10. Wingender, E., et al., *The TRANSFAC system on gene expression regulation.* Nucl. Acids Res., 2001. **29**(1): p. 281-283.
11. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.
12. Beissbarth, T. and T.P. Speed, *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics, 2004. **20**(9): p. 1464-5.
13. Zhang, B., et al., *GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.* BMC Bioinformatics, 2004. **5**(1): p. 16.
14. Zhong, S., et al., *Comparative Analysis of Gene Ontology Space under the Multiple Hypothesis Testing Framework.* Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB2004), 2004.
15. Olefsky, J.M. and A.R. Saltiel, *PPAR gamma and the treatment of insulin resistance.* Trends Endocrinol Metab, 2000. **11**(9): p. 362-8.
16. Suzuki, A., et al., *Alteration in expression profiles of a series of diabetes-related genes in db/db mice following treatment with thiazolidinediones.* Jpn J Pharmacol, 2000. **84**(2): p. 113-23.
17. Albrektsen, T., et al., *Novel genes regulated by the insulin sensitizer rosiglitazone during adipocyte differentiation.* Diabetes, 2002. **51**(4): p. 1042-51.

18. Schadt, E.E., et al., *Analyzing high-density oligonucleotide gene expression array data.* J Cell Biochem, 2000. **80**(2): p. 192-202.
19. Miles, M.F., *Microarrays: lost in a storm of data?* Nat Rev Neurosci, 2001. **2**(6): p. 441-3.
20. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
21. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.* Bioinformatics, 2001. **17**(6): p. 509-19.
22. Rocke, D.M. and B. Durbin, *A model for measurement error for gene expression arrays.* J Comput Biol, 2001. **8**(6): p. 557-69.
23. Morris, C., *Natural Exponential Families with Quadratic Variance Functions: Statistical Theory.* Ann Statis, 1983. **11**(2): p. 515-529.
24. Ideker, T., et al., *Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.* J Comput Biol, 2000. **7**(6): p. 805-17.
25. Smyth, G.K., *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1).
26. Durbin, B.P. and D.M. Rocke, *Variance-stabilizing transformations for two-color microarrays.* Bioinformatics, 2004. **20**(5): p. 660-667.
27. Hsiao, A., et al., *Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes.* Bioinformatics, 2004.
28. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.
29. Rome, S., et al., *Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp.* J Biol Chem, 2003. **278**(20): p. 18063-8.
30. Tullai, J.W., et al., *Identification of transcription factor binding sites upstream of human genes regulated by the phosphatidylinositol 3-kinase and MEK/ERK signaling pathways.* J Biol Chem, 2004. **279**(19): p. 20167-77.
31. Bokemeyer, D., M. Lindemann, and H.J. Kramer, *Regulation of Mitogen-Activated Protein Kinase Phosphatase-1 in Vascular Smooth Muscle Cells.* Hypertension, 1998. **32**(4): p. 661-667.
32. Karlsson, M., et al., *Both Nuclear-Cytoplasmic Shuttling of the Dual Specificity Phosphatase MKP-3 and Its Ability to Anchor MAP Kinase in the Cytoplasm Are Mediated by a Conserved Nuclear Export Signal.* J. Biol. Chem., 2004. **279**(40): p. 41882-41891.
33. Ulevitch, R.J. and P.S. Tobias, *Recognition of Gram-negative bacteria and endotoxin by the innate immune system.* Current Opinion in Immunology, 1999. **11**(1): p. 19-22.
34. Oshikawa, K. and Y. Sugiyama, *Gene expression of Toll-like receptors and associated molecules induced by inflammatory stimuli in the primary alveolar macrophage.* Biochem Biophys Res Commun, 2003. **305**(3): p. 649-55.

35. Yokochi, T., et al., *In vivo effects of bacterial lipopolysaccharide on proliferation of macrophage colony-forming cells in bone marrow and peripheral lymphoid tissues.* Infect Immun, 1985. **47**(2): p. 496-501.
36. Schaible, U.E., et al., *Apoptosis facilitates antigen presentation to T lymphocytes through MHC-I and CD1 in tuberculosis.* Nat Med, 2003. **9**(8): p. 1039-46.
37. O'Brien, R.M. and D.K. Granner, *Regulation of gene expression by insulin.* Physiol Rev, 1996. **76**(4): p. 1109-61.
38. Sartipy, P. and D.J. Loskutoff, *Expression Profiling Identifies Genes That Continue to Respond to Insulin in Adipocytes Made Insulin-resistant by Treatment with Tumor Necrosis Factor-{alpha}.* J. Biol. Chem., 2003. **278**(52): p. 52298-52306.
39. Yechoor, V.K., et al., *Coordinated patterns of gene expression for substrate and energy metabolism in skeletal muscle of diabetic mice.* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10587-92.
40. Yechoor, V.K., et al., *Distinct pathways of insulin-regulated versus diabetes-regulated gene expression: an in vivo analysis in MIRKO mice.* Proc Natl Acad Sci U S A, 2004. **101**(47): p. 16525-30.
41. Hevener, A.L., et al., *Thiazolidinedione Treatment Prevents Free Fatty Acid-Induced Insulin Resistance in Male Wistar Rats.* Diabetes, 2001. **50**(10): p. 2316-2322.
42. Hsiao, A., et al., *Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes.* Bioinformatics, 2004. **20**(17): p. 3108-27.
43. Hsiao, A., et al., *VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression data.* Nucleic Acids Research, accepted.
44. Hsiao, A. and S. Subramanium, *Bivariate microarray analysis: Statistical interpretation of two-channel functional genomics data.* (in review).
45. Kuninger, D., et al., *Gene discovery by microarray: identification of novel genes induced during growth factor-mediated muscle cell survival and differentiation.* Genomics, 2004. **84**(5): p. 876-889.
46. Tseng, Y.-H., et al., *Regulation of Growth and Tumorigenicity of Breast Cancer Cells by the Low Molecular Weight GTPase Rad and Nm23.* Cancer Res, 2001. **61**(5): p. 2071-2079.
47. Suzuki, M., et al., *Aberrant methylation profile of human malignant mesotheliomas and its relationship to SV40 infection.* Oncogene, 2005. **24**(7): p. 1302-8.
48. Laville, M., et al., *Acute regulation by insulin of phosphatidylinositol-3-kinase, Rad, Glut 4, and lipoprotein lipase mRNA levels in human muscle.* J Clin Invest, 1996. **98**(1): p. 43-9.
49. Arlt, A., et al., *The early response gene IEX-1 attenuates NF-kappaB activation in 293 cells, a possible counter-regulatory process leading to enhanced cell death.* Oncogene, 2003. **22**(21): p. 3343-51.
50. Osawa, Y., et al., *Expression of the NF-{kappa}B Target Gene X-Ray-Inducible Immediate Early Response Factor-1 Short Enhances TNF-{alpha}-Induced*

Hepatocyte Apoptosis by Inhibiting Akt Activation. J Immunol, 2003. **170**(8): p. 4053-4060.
51. Reinholz, M.M., et al., *Differential gene expression of TGF-beta family members and osteopontin in breast tumor tissue: analysis by real-time quantitative PCR.* Breast Cancer Res Treat, 2002. **74**(3): p. 255-69.
52. Hiranuma, C., et al., *Hypermethylation of the MYOD1 gene is a novel prognostic factor in patients with colorectal cancer.* Int J Mol Med, 2004. **13**(3): p. 413-7.
53. MULLER, H.M., et al., *Prognostic DNA Methylation Marker in Serum of Cancer Patients.* Ann NY Acad Sci, 2004. **1022**(1): p. 44-49.
54. Jin, S., et al., *The GADD45 Inhibition of Cdc2 Kinase Correlates with GADD45-mediated Growth Suppression.* J. Biol. Chem., 2000. **275**(22): p. 16602-16608.
55. Sun, H.B., et al., *MRG1, the product of a melanocyte-specific gene related gene, is a cytokine-inducible transcription factor with transformation activity.* Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13555-60.
56. Brown, S.A., et al., *The Fn14 cytoplasmic tail binds tumour-necrosis-factor-receptor-associated factors 1, 2, 3 and 5 and mediates nuclear factor-kappaB activation.* Biochem J, 2003. **371**(Pt 2): p. 395-403.
57. Florini, J.R., D.Z. Ewton, and K.A. Magri, *Hormones, Growth Factors, and Myogenic Differentiation.* Annual Review of Physiology, 1991. **53**(1): p. 201-216.
58. De Arcangelis, V., et al., *IGF-I-induced Differentiation of L6 Myogenic Cells Requires the Activity of cAMP-Phosphodiesterase.* Mol. Biol. Cell, 2003. **14**(4): p. 1392-1404.
59. Naro, F., et al., *Involvement of Type 4 cAMP-Phosphodiesterase in the Myogenic Differentiation of L6 Cells.* Mol. Biol. Cell, 1999. **10**(12): p. 4355-4367.
60. Cirri, P., et al., *Insulin inhibits platelet-derived growth factor-induced cell proliferation.* Mol Biol Cell, 2005. **16**(1): p. 73-83.
61. Terauchi, Y., et al., *Increased insulin sensitivity and hypoglycaemia in mice lacking the p85 alpha subunit of phosphoinositide 3-kinase.* Nat Genet, 1999. **21**(2): p. 230-5.
62. Mauvais-Jarvis, F., et al., *Reduced expression of the murine p85{alpha} subunit of phosphoinositide 3-kinase improves insulin signaling and ameliorates diabetes.* J. Clin. Invest., 2002. **109**(1): p. 141-149.
63. Ueki, K., et al., *Positive and Negative Roles of p85{alpha} and p85{beta} Regulatory Subunits of Phosphoinositide 3-Kinase in Insulin Signaling.* J. Biol. Chem., 2003. **278**(48): p. 48453-48466.
64. Brachmann, S.M., et al., *Phosphoinositide 3-Kinase Catalytic Subunit Deletion and Regulatory Subunit Deletion Have Opposite Effects on Insulin Sensitivity in Mice.* Mol. Cell. Biol., 2005. **25**(5): p. 1596-1607.
65. Emanuelli, B., et al., *SOCS-3 Inhibits Insulin Signaling and Is Up-regulated in Response to Tumor Necrosis Factor-alpha in the Adipose Tissue of Obese Mice.* J. Biol. Chem., 2001. **276**(51): p. 47944-47949.
66. Emanuelli, B., et al., *SOCS-3 Is an Insulin-induced Negative Regulator of Insulin Signaling.* J. Biol. Chem., 2000. **275**(21): p. 15985-15991.

67. Sadowski, C.L., et al., *Insulin Induction of SOCS-2 and SOCS-3 mRNA Expression in C2C12 Skeletal Muscle Cells Is Mediated by Stat5*.* J. Biol. Chem., 2001. **276**(23): p. 20703-20710.
68. Peraldi, P., et al., *Insulin Induces Suppressor of Cytokine Signaling-3 Tyrosine Phosphorylation through Janus-activated Kinase.* J. Biol. Chem., 2001. **276**(27): p. 24614-24620.
69. Bard-Chapeau, E.A., et al., *Deletion of Gab1 in the liver leads to enhanced glucose tolerance and improved hepatic insulin action.* Nat Med, 2005.
70. Duong, D.T., et al., *Insulin Inhibits Hepatocellular Glucose Production by Utilizing Liver-enriched Transcriptional Inhibitory Protein to Disrupt the Association of CREB-binding Protein and RNA Polymerase II with the Phosphoenolpyruvate Carboxykinase Gene Promoter.* J. Biol. Chem., 2002. **277**(35): p. 32234-32242.
71. Hall, R., F. Sladek, and D. Granner, *The Orphan Receptors COUP-TF and HNF-4 Serve as Accessory Factors Required for Induction of Phosphoenolpyruvate Carboxykinase Gene Transcription by Glucocorticoids.* PNAS, 1995. **92**(2): p. 412-416.
72. Odom, D.T., et al., *Control of pancreas and liver gene expression by HNF transcription factors.* Science, 2004. **303**(5662): p. 1378-81.
73. Shen, W., et al., *Foxa3 (Hepatocyte Nuclear Factor 3gamma) Is Required for the Regulation of Hepatic GLUT2 Expression and the Maintenance of Glucose Homeostasis during a Prolonged Fast.* J. Biol. Chem., 2001. **276**(46): p. 42812-42817.
74. Inoue, H., et al., *Role of STAT-3 in regulation of hepatic gluconeogenic genes and carbohydrate metabolism in vivo.* Nat Med, 2004. **10**(2): p. 168-74.
75. Yoon, J.C., et al., *Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1.* Nature, 2001. **413**(6852): p. 131-8.
76. Puigserver, P., et al., *Insulin-regulated hepatic gluconeogenesis through FOXO1-PGC-1alpha interaction.* Nature, 2003. **423**(6939): p. 550-5.
77. Lin, J., et al., *PGC-1{beta} in the Regulation of Hepatic Glucose and Energy Metabolism.* J. Biol. Chem., 2003. **278**(33): p. 30843-30848.
78. Barthel, A. and D. Schmoll, *Novel concepts in insulin regulation of hepatic gluconeogenesis.* Am J Physiol Endocrinol Metab, 2003. **285**(4): p. E685-692.
79. Bretz, J.D., et al., *C/EBP-related protein 2 confers lipopolysaccharide-inducible expression of interleukin 6 and monocyte chemoattractant protein 1 to a lymphoblastic cell line.* Proc Natl Acad Sci U S A, 1994. **91**(15): p. 7306-10.
80. Akira, S., et al., *A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family.* Embo J, 1990. **9**(6): p. 1897-906.
81. May, P., et al., *Signal transducer and activator of transcription STAT3 plays a major role in gp130-mediated acute phase protein gene activation.* Acta Biochim Pol, 2003. **50**(3): p. 595-601.
82. Hotamisligil, G.S., *Inflammatory pathways and insulin action.* Int J Obes Relat Metab Disord, 2003. **27 Suppl 3**: p. S53-5.
83. Xu, S., et al., *Abrogation of DUSP6 by hypermethylation in human pancreatic cancer.* Journal of Human Genetics, 1900.

84. Tirone, F., *The gene PC3(TIS21/BTG2), prototype member of the PC3/BTG/TOB family: regulator in control of cell growth, differentiation, and DNA repair?* J Cell Physiol, 2001. **187**(2): p. 155-65.
85. Schumann, R., et al., *The lipopolysaccharide-binding protein is a secretory class 1 acute- phase protein whose gene is transcriptionally activated by APRF/STAT/3 and other cytokine-inducible nuclear proteins.* Mol. Cell. Biol., 1996. **16**(7): p. 3490-3503.
86. Heck, R., et al., *Catalytic properties of mouse carbonic anhydrase V*. J. Biol. Chem., 1994. **269**(40): p. 24742-24746.
87. Klein, C., et al., *The IL-6-gp130-STAT3 pathway in hepatocytes triggers liver protection in T cell-mediated liver injury.* J Clin Invest, 2005.
88. Sekine, O., et al., *Insulin Activates CCAAT/Enhancer Binding Proteins and Proinflammatory Gene Expression through the Phosphatidylinositol 3-Kinase Pathway in Vascular Smooth Muscle Cells.* J. Biol. Chem., 2002. **277**(39): p. 36631-36639.
89. Yoshida, Y., et al., *Interleukin 1 Activates STAT3/Nuclear Factor-{kappa}B Cross-talk via a Unique TRAF6- and p65-dependent Mechanism.* J. Biol. Chem., 2004. **279**(3): p. 1768-1776.
90. Mathison, J., et al., *Plasma lipopolysaccharide (LPS)-binding protein. A key component in macrophage recognition of gram-negative LPS.* J Immunol, 1992. **149**(1): p. 200-206.
91. Kirschning, C.J., et al., *The transcriptional activation pattern of lipopolysaccharide binding protein (LBP) involving transcription factors AP-1 and C/EBP beta.* Immunobiology, 1997. **198**(1-3): p. 124-35.
92. Sartipy, P. and D.J. Loskutoff, *Expression profiling identifies genes that continue to respond to insulin in adipocytes made insulin-resistant by treatment with tumor necrosis factor-alpha.* J Biol Chem, 2003. **278**(52): p. 52298-306.
93. Sartipy, P. and D.J. Loskutoff, *Monocyte chemoattractant protein 1 in obesity and insulin resistance.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 7265-70.