Genomic Analysis of Hepatic Insulin Resistance

by

R. Michael Raab

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[February, 2006]
October 2005

Author .... 27-Oct-05 ...
Department of Chemical Engineering
October, 2005

Certified by .... 10/27/05 ....
Gregory Stephanopoulos
Bayer Professor of Chemical Engineering
Thesis Supervisor

Accepted by ....
Daniel Blankschtein
Chairman, Department Committee on Graduate Students

Genomic Analysis of Hepatic Insulin Resistance

by

R. Michael Raab

Submitted to the Department of Chemical Engineering
on October, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Type II Diabetes mellitus is a genetically complex disease characterized by insulin resistance in peripheral tissues, which results in simultaneous hyperglycemia and hyperinsulinemia. Because of the prevalence of type II diabetes, many researchers are investigating the genetics of glucose homeostasis, however, traditional mapping techniques have not been successful in determining all of the genes that regulate glycemia. To complement these efforts, we used DNA microarrays to find differentially expressed genes and combinatorial siRNA screening to investigate the effects of hepatic gene transcription during periods of high and low glucose production. This strategy provides a new approach to studying the molecular mechanisms of disease pathogenesis.

Our investigations focused on discovering new genes that influence hepatic metabolism and glucose production. Hepatocytes help maintain whole body glycemia by providing glucose and other substrates during non-feeding periods. DNA microarrays containing 17,000 unique gene probes were used to study hepatic gene transcription during normal, insulin resistant, and fasting states in C57/BL/6J mice. We analyzed this data set using a combination of statistical and multivariate techniques to determine 41 different genes that are differentially expressed and highly discriminatory of the treatment groups.

Hepatocytes perform many physiological roles, thus to investigate which genes from the microarray analysis affected hepatic metabolism, we developed combinatorial RNA-interference (RNAi) based gene silencing techniques. Using combinatorial siRNA screening, we silenced genes that were over–expressed within the microarray data set to study loss of function effects on hepatic metabolism, which was quantified by measuring intracellular metabolite concentrations in relevant metabolic pathways. Based upon the metabolite dependent clustering of experimental and control samples using Fisher Discriminant Analysis, four of the silenced genes had a significant effect on key metabolites involved in hepatic glucose output. Of these four genes, three were shown to influence hepatic glucose output in our primary cell model.

Thesis Supervisor: Gregory Stephanopoulos
Title: Bayer Professor of Chemical Engineering

Acknowledgments

The last five years have been a tremendous learning experience for me intellectually, professionally, and personally. Over that time I have thoroughly enjoyed living in Cambridge, MA and attending the Massachusetts Institute of Technology, where I have been steadily working to assemble the document you now hold. The Boston area is a truly vibrant place to pursue research, and the community provides no shortage of seminars, interesting ideas, alternative viewpoints, and opportunities. That being said, what really made the events of the past five years particularly rewarding, are the incredible people I have had the privilege to know. A few are locals, however, many arrived in Boston from across the United States, and some even hail from distant continents. Despite their different backgrounds, motivations, and perspectives, they all share a common respect for each other, a lot of ambition, and a refreshing thirst for learning. The combination of these characteristics has made my time in graduate school a special period of my life and one for which I am very grateful.

For these reasons I have to thank a number of people that both enriched my experiences and enabled me to enjoy them in the first place. Clearly none of my achievements would ever have been possible without the unwavering support of my family. Although I probably do not acknowledge them enough, their help, guidance, and friendship has always been appreciated. Indeed, having such a wonderful family foundation allows me to take risks that I may have forgone in their absence, which only could have limited my success.

At MIT, I would like to thank the combined Stephanopoulos research group for their support, and my committee: Robert Langer, Christos Mantzoros, Joanne Kelleher, Doug Lauffenburger, and my advisor, Gregory Stephanopoulos. It may seem no small feat for faculty to advise students and help them develop, however, this group's input has been very valuable to me. Dr. Langer, despite his enormously busy schedule, *always* reads my emails and up–dates, which he demonstrates by replying with inquisitive questions and supportive comments. Dr. Mantzoros was particularly important to helping set the scope of the thesis and providing focus to relevant problems. In addition, Dr. Mantzoros introduced me to many people at the hospital and made me aware of all the resources that a teaching hospital provides, which I think will serve me well in the future. Finally, I really have to thank Dr. Stephanopoulos for helping me make it to MIT, providing guidance, and giving me a chance to perform. His relaxed style was a good fit for someone in my position and I consider it a real honor to have been able to participate in his group's research.

There are also a few people that deserve mentioning for their prior influence on me. Although we may not be in regular contact, this group has impacted my development for various reasons and this is one of the rare opportunities I have to formally acknowledge them. Dr. Kathy McKinney, who was my first supervisor at Merck, gave me a chance out of college that has really led to my current position. Kathy probably did not know what she was in for when she hired me, but despite our up's and down's, she has always been incredibly supportive and a person I am lucky to know. Several other mentors at the Merck Research Laboratories (MRL) include Dr. Ann Lee, Dr. Michael Washabaugh, and importantly Dr. Wayne Herber. Wayne was the director of our research group in MRL and the best manager I have ever known or seen. I continue to revisit and try to apply the lessons he implicitly taught us, as my own leadership roles develop. Wayne assembled a great group, kept us motivated, and despite a grueling schedule and set of expectations, he always knew the right thing to say. I hope I can do the same.

Prior to Merck there were three other people I should mention for their mentoring and support. First is Dr. Doug Cameron, who was the professor at the University of Wisconsin that provided my first research opportunity and for whom I have a lot of respect and admiration. I have been lucky to keep up with Doug over the years and watch his career grow rapidly, for which he is truly deserving. Second is Dr. Edwin Lightfoot at the University of Wisconsin. I had the real privilege of spending one semester working closely with Ed, a person who has a level of creativity I had never encountered before, nor have I encountered since. Ed's great personality, commitment to excellence, and incredible creativity are really contagious. Third, I need to thank a high–school English teacher, Mr. Aunans, for literally and figuratively waking me up in class one day when I was sleeping. I am not sure why his tirade had such an effect, but I do recognize it as the day I woke up and started looking beyond myself. Like science, life is very serendipitous, so I am lucky he would not allow me to continue sleeping as opportunity passed.

I have purposely saved my friends' acknowledgements for last. Of all groups, this is the one I spend the most time with on a daily basis and therefore continuously enriches my life. There are so many, in so many different places, that I doubt I can give them the attention they deserve. Nonetheless, I will try my best. From Merck there is Dr. Stephen Decker and Dr. Cian Ryle, Dr. Hugh George, Dr. Kara (van Struck) Calhoun, Bethany Rogers, Eva Gefroh, Jen (Bautista) Brown, Charlie Parker, Jim Bailey, Mark Mikola, Dr. Fran Meacle (and Clarissa), Alison Snyder, Scott Meyers, Dr. Andy Bett, Dr. Dave Gerhold, and at least a couple dozen others that made MRL a great place to be during those years. I look forward to keeping in touch with all of you in the future and wish you only the best of luck in your endeavors. From Minnesota there are my old friends, now scattered throughout the United States. Next to my family, these friends provide their own foundation, and I love how balanced their lives have become. In particular I want to thank Craig Barbee for always keeping in touch, and John Sippola for organizing our fantasy football league (which routinely lines my pockets with their spare cash). It's a wonderful group.

Finally I need to acknowledge all of my friends from the Boston area. Eric Barbee, who was here and now has left, was always great to catch up with and exchange stories about Minnesota, and his Harvard experiences. From Agrivida, Karl Ruping has become not only a fantastic work colleague, but also a friend and a great mentor in the business of starting small companies. Karl's patience, resourcefulness, and help in getting Agrivida off the ground has not only been critical to our success, but also has provided one of the most incredible experiences I have ever enjoyed. In Agrivida's laboratory I need to thank Dr. Humberto de la Vega and Marisha Youngblood for continuing to make our research successful and creating a great company culture. Jeremy Johnson, K.J., Fred Koenig, Dr. Jeff Tester, Scott Pearce, Dr. Elizabeth

Hood, and Dr. Fred Ausubel also contribute to Agrivida's success and I hope that will continue in the future. Alice Jacobs, although not directly involved with Agrivida and whom I only recently met, has provided inspiration and education in the start–up world through her own successes. I am not sure I have ever met anyone with her drive and dedication before, which has renewed my own motivation, commitment, and ambition. I am very grateful to know someone as talented as Alice and hope our friendship continues to grow.

The majority of my friends in Boston were students from our class at MIT. It is a very talented group and I think one that will be extremely successful in the future. Like most people that share common hardships, I doubt we will ever forget all of those difficult nights of problem sets, studying for exams, and finishing experiments. Luckily, the Muddy was usually available and within this group someone could always be counted on for a beer. Jeremy Johnson, Joe Moritz, and Prem Pavoor have all become great friends and I hope they continue to sponsor Premapalooza in the future. Stephen Fox and Kathryn Miller from the "other" office were never late to the TG's, and always provided interesting conversation. For those nights when conversation was taking a backseat to malted beverages, I was lucky to share the company of Dr. Greg Randall, one Dr. Michael Berg ("The Berg") and the future president of the United States, Dr. Brian Baynes. I look forward to the day when Brian is in office and I can write about our late nights eating low grade Chinese food at the Hong Kong in Harvard Sq. Kyle Jensen, who has become my best friend at MIT, has taught me so much that it cannot be summed up here. It will be interesting to see where Kyle's talents lead him in the future. Many of these folks have already moved on to promising careers in consulting, small businesses, and research. Those that have not yet, will soon. Regardless of where their paths lead them, I was honored to have shared some common experiences with this group while in Cambridge, and I sincerely look forward to observing your future endeavors. Thank you for your friendship!!

Cambridge, MA R. Michael Raab
October 27, 2005

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diabetes is a growing problem throughout the world and the subject of intense research. The control of blood glucose levels within a person is mediated through a number of complex systems, whose molecular basis is not completely understood. Further insight can be gained by identifying genes that affect blood glucose levels and determining their biochemical and physiological effects. This chapter provides an overview of the disease and its significance to human health.

Chapter two introduces the general concept of quantitative and polygenic traits, and discusses methods for discovering disease genes. Blood glucose control can be considered a quantitative, polygenic trait, and therefore the background provided in this chapter is necessary to compare and contrast our strategy with previous methodologies. Methodologies have evolved over the years, and the work presented herein employed a novel approach whose advantages, disadvantages, and development are presented.

Chapter three describes methods for gene characterization, once a gene has been implicated as having a potential role in the phenotype being studied. Different experimental model systems are discussed, as well traditional and new methods for manipulating gene expression including the introduction of RNA interference (RNAi) as an efficient method for gene silencing. Development and application of RNAi to cells in culture is presented.

Chapter four describes the application of the methods developed in chapter two to hepatic gene discovery. This work identified genes that may mediate differences between control mice, insulin resistant, diet–induced obese (DIO) mice, and DIO mice who were fasted for 48 hours, returning their weights back to baseline. Forty–one genes were identified that were differentially expressed and discriminatory of the experimental treatments, implying an important role in the underlying physiology.

Chapter five applies combinatorial RNAi screening to genes identified in chapter four. The effect of gene silencing on cellular metabolite pools was used to determine which genes had a significant impact on hepatic metabolism. Of the 15 genes over–expressed in the mouse experiments, four had an impact on metabolism, and three of these reduced hepatic glucose output in our primary cell culture model.

1.1 Overview of Diabetes mellitus

Diabetes mellitus is a condition broadly characterized by elevated blood glucose levels (hyperglycemia). The condition has been observed in humans for over two thousand years and its name, "diabetes mellitus", was coined in 1674 by Thomas Willis [52]. *Diabetes* stems from the Greek word meaning "to syphon" in reference to the chronic wasting that occurs in untreated patients, while *mellitus* comes from the Latin word for honey in reference to the sweet taste of the urine that was often used in ancient times for diagnosis [197].

Although overt diabetes presents clear symptoms, which include sweet smelling urine, acetic breath, frequent urination, and weight loss, little was known about the ailment until relatively modern times. The first conceptual break through in the pathology of diabetes came in 1889, when Dr. von Mering and Dr. Minkowski of Germany discovered a link to the pancreas [159]. These physicians found that if they removed a dog's pancreas (pancreatectomy), the animal's urine would accumulate high levels of sugar and it would develop a condition similar to diabetes. Thirty years later, in 1921, Banting and Best discovered insulin as the key pancreatic factor controlling glucose levels. After perfecting the isolation technique for bovine insulin, they were able to treat the first patient, Leonard Thompson, in 1922. At the time of treatment Thompson was extremely ill, however, he made a miraculous recovery and lived for an additional 13 years on bovine insulin.

Since these early observations research into diabetes mellitus has been widespread. It is now accepted that diabetes mellitus actually constitutes a number of different conditions, all portraying the hallmark hyperglycemia. These conditions are broadly classified among two types (Type I and Type II), gestational diabetes, and a series of rare syndromes [185]. Most diabetics fall under the modern classification system, which diagnoses them as either Type I or Type II depending upon the underlying cause. Type I diabetics have an autoimmune disorder that causes destruction of β–cells in the pancreas, resulting in total absence of insulin. Type I diabetes is often first observed in the early teen years and presents with the common symptoms of frequent urination, weight loss, and may be associated with elevated serum and urinary ketone levels (diabetic ketoacidosis). In contrast, Type II diabetics do not have an autoimmune disorder, but instead are insulin resistant. Type II diabetes is the most common form of the disease, often being detected during middle age, but its incidence is rising dramatically in younger people [69, 68]. Type II diabetics are frequently over weight and are usually diagnosed by elevated glucose levels during clinical visits. Type II diabetics account for 90 – 95% of all known cases.

The rising incidence of Type II diabetes is a concern for both developed and developing countries. Global estimates project an increase from 150 million diabetics in 2000 to 220 million in 2010 [277]. This growth in disease incidence is largely associated with a sedentary lifestyle and increasing levels of obesity, and has important implications for the associated costs of health care. Diabetes is a leading risk factor for retinopathy and blindness, nephropathy and kidney disease, neuropathy, and is associated with a number of cardiovascular disease risk factors. In all, it is estimated that diabetes accounts for over $130 billion in health care costs in the United States and is the fifth leading cause of death [111]. Even with its rich history in research, the spread of the disease has provided further incentives for understanding its molecular basis, which will hopefully contribute to new therapies.

Although many hormones influence glucose levels, in healthy people glucose homeostasis is maintained largely by insulin signaling. Following a meal glucose is absorbed into the blood where it is sensed by the pancreas. Pancreatic β–cells absorb glucose, thereby increasing their ATP/ ADP ratio. This increase causes ATP–sensitive potassium channels to close, depolarizing the cells, which subsequently causes voltage–regulated calcium channels to open. This leads to an increase in Ca^{2+} levels in the cell and excretion of insulin via exocytosis from insulin containing vesicles [15]. Once secreted, insulin travels through the blood to bind receptors on target tissue cells. In glycemic terms, the primary insulin sensitive tissues are muscle, adipose, and liver. Muscle cells take up glucose in response to insulin and store it as glycogen, but also use it to synthesize proteins. Adipocytes respond to insulin by taking up glucose and storing it as fat. Unlike muscle and adipose cells, liver cells do not modulate glucose up–take in response to insulin; instead insulin binding induces hepatocytes to suppress glucose production and store glucose as glycogen. Glucose homeostasis in the postprandial state is shown in Figure 1-1.

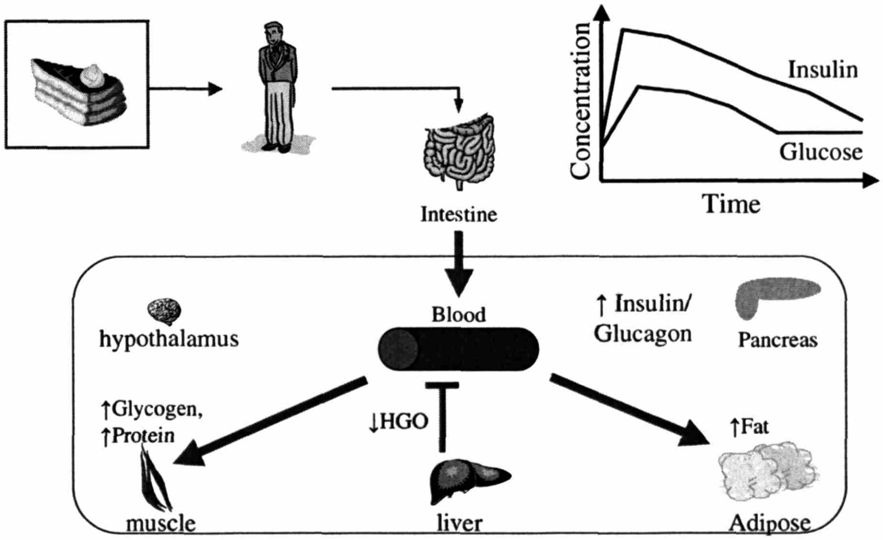


Figure 1-1: Glucose homeostasis in the postprandial state.

In Figure 1-1 rising glucose levels increase insulin secretion by the pancreas, which drives glucose up–take by the muscle and adipose, and importantly suppresses glucose output by the liver. Over time glucose and insulin levels fall back to normal. This usually takes two to three hours depending upon the person, size of the meal, and meal composition.

Type II diabetics are insulin resistant and have simultaneous hyperglycemia and hyperinsulinemia, as shown in Figure 1-2. In this condition, glucose is absorbed into the blood stream and sensed properly by the pancreas, but target tissues no longer respond adequately to the released insulin. Thus muscle and adipose tissues do not remove glucose as they would normally, and hepatic glucose output continues even in the presence of elevated glucose and insulin levels. To compensate, the pancreas increases its level of insulin secretion, thereby resulting in the elevated insulin concentration observed in the blood of patients. Over time secretion of insulin by the pancreas may diminish, leading to overt diabetes. The relative contributions of each tissue (muscle glucose up–take, adipose glucose up–take, hepatic glucose output, pancreatic insulin secretion) to elevated glucose levels vary depending upon the specific patient, contributing to the overall heterogeneity of the disorder.

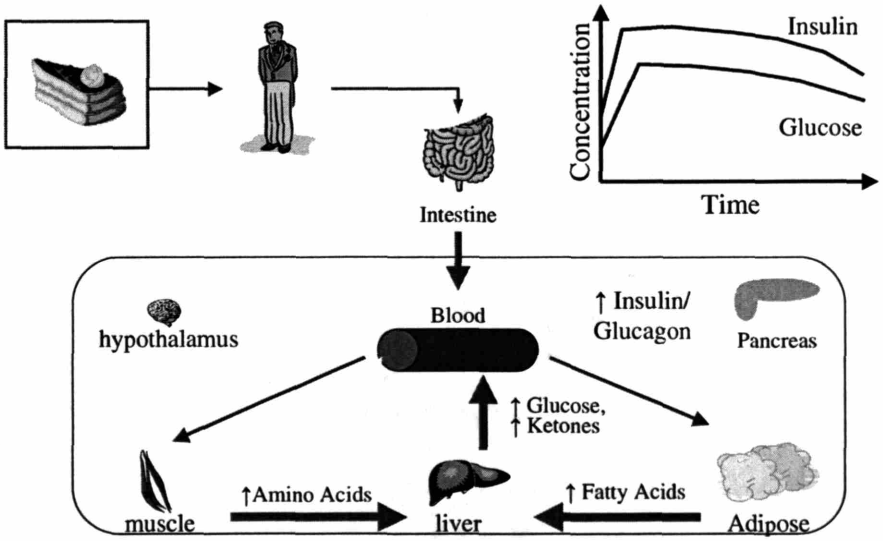


Figure 1-2: Impaired glucose homeostasis that takes place in Type II diabetics.

The exact molecular cause of Type II diabetes is unknown and many cross–sectional and prospective epidemiological studies have been performed to determine disease risk factors. These studies have shown several important characteristics exist among various populations including ethnicity, obesity, age, sex, and genetics [95]. In addition, several biochemical and physiological markers, such as body mass index (BMI), waist-to-hip ratio, fasting insulin concentration, and impaired glucose tolerance have also been identified as independent risk factors [94].

In the U.S., African–Americans and Hispanics have at least a twofold increase in risk relative to Caucasians, whereas Native Americans have a fivefold increase [99, 231]. Although the risk within these ethnic populations is inversely related to socioeconomic status, the differences between ethnic groups still arise when correcting for socioeconomic class and other variables, suggesting that some genetic factors within different ethnic populations contribute to their risk of Type II diabetes [93].

The strongest evidence for genetic risk factors comes from twin studies and rare syndromes that demonstrate the potential for genetic influence. One study showed that from 53 twin pairs, in which one twin had Type II diabetes, 48 of the co-twins developed the disease in later assessments, representing a 91% concordance rate [10]. Notably in the same study, the five twins that did not have diabetes at the time, did have mild glucose intolerance and abnormal insulin responses during oral glucose tolerance tests, suggesting they are also at risk of developing the disease with time. The high concordance between twins, coupled with the differences between ethnic populations, are strong evidence that genetic factors are important to understanding the molecular basis for the disease. This situation is vividly demonstrated by a number of known syndromes of insulin resistance resulting from single gene mutations [74]. Although rare, these syndromes prove that specific genetic mutations, in a variety of genes, can give rise to severe insulin resistance. Genes for which mutations have been identified include those encoding insulin, the insulin receptor, and glucokinase [52].

Besides genetic risk factors, environmental effects also predispose certain populations to Type II diabetes. Obesity is a primary risk factor that is becoming more common because of an environment that provides readily available food and an increasingly sedentary lifestyle. Indeed, 85% - 90% of all patients clinically diagnosed with Type II diabetes are overweight or obese [133]. Although numerous studies have demonstrated the link between obesity and diabetes, and have even quantified the increased risk due to various obesity related measurements (such as BMI and waist-to-hip ratio), the mechanisms that link obesity to diabetes remain elusive. In this regard, because both disorders are complex, multigenic traits, discovering some of the genes that are potentially involved would be a large contribution to the field.

Despite our clear deficiency in understanding the genetic basis for Type II diabetes, clinical diagnosis [1] has improved tremendously and treatments exist for controlling diabetes. These treatments have evolved empirically over the last century and currently present most patients with effective alternatives such that they can lead relatively normal lives and often avoid many of the debilitating diabetic complications.

Diabetes management has a number of levels depending upon the severity and complications of the specific patient. Because most patients are overweight, the first line of defense is to control glycemia through life style changes. These usually require the patient to lose weight by changing their diet and increasing their level of exercise. For patients that have insulin resistance, but not overt diabetes, weight loss in the range of 5% to 10% usually reduces insulin resistance and improves insulin tolerance [133], which may be enough to control glycemia. Because sustained weight reduction is extremely difficult for many patients and because insulin resistance may not be diagnosed until endogenous insulin secretion has already been diminished, additional treatments are available. Thus if life style changes are ineffective, the next level of treatment is oral hypoglycemic agents. There are currently several classes of oral agents available, detailed in Table 1.1, as well as an increasing number of new compounds [107]. If oral hypoglycemic reagents still do not adequately control blood glucose levels, or if pancreatic secretion of insulin is not sufficient, then Type II diabetics must start insulin therapy.

Because of the increasing incidence of Type II diabetes and the patient require-

[1] The clinically useful tests for diagnosing diabetes in patients are measurement of serum glucose levels, glucose tolerance test, urinary and serum ketone levels, hemoglobin A_{1C}, and urinary microalbumin excretion [133]. Of these, the most commonly used for diagnosis is the measurement of serum glucose levels, which in normal patients is less than 115 mg/ dL (6.4 mM) following an overnight fast. A fasting serum glucose level of greater than 126 mg/ dL (7.0 mM) and an accompanying negative result in another clinical test, confirm the diagnosis of diabetes. Urinary serum and ketone levels help identify Type I diabetics who may be developing ketoacidosis from excessive fat metabolism in the absence of insulin secretion. Hemoglobin A_{1C} and urinary microalbumin excretion help determine the extent of diabetes and effectiveness of disease management. Finger-stick blood sugar is commonly used by patients to monitor their day-to-day glycemic control.

Drug Class	Mode of Action
Sulfonylureas	Increase the secretion of insulin by β-cells.
Biguanides	Increases glucose utilization.
Glucosidase Inhibitor	Decreases glucose absorption in the small intestine.
Thiazolidinediones	Reduce insulin resistance.

Table 1.1: Oral Hypoglycemic Agents.

ments for managing glycemia, continued research and improved therapies are required. As a primary defect in any single insulin sensitive tissue may lead to detrimental insulin resistance, studying tissue specific molecular pathogenesis is a promising approach for providing targeted and efficacious treatments. By understanding the genes involved in the disease, it may be possible to more accurately define the underlying molecular mechanisms that regulate glycemia, and thereby develop more potent therapies for this complex disease.

Chapter 2

Methods of Gene Identification

This chapter provides an overview of methods and technologies used to identify disease related genes. For monogenic disorders this is a relatively straightforward task, amenable to mendelian techniques such as pedigree analysis, gene mapping, and genotyping. For complex diseases associated with quantitative traits, whose phenotypes are continuous in nature and are most often influenced by multiple genes, other methods are required, which have been far less elucidating in determining the complete set of genes involved. For this reason, new experimental techniques, such as expression profiling using DNA microarrays, are being incorporated as a way of facilitating the identification of genes that may be involved with complex traits, such as glucose homeostasis. Because DNA microarrays are capable of generating massive amounts of data, efficient analysis of the data is necessary and several methods are presented for different experimental designs.

2.1 Genetics of Quantitative Traits

The goal of genetics is the analysis of the genotype of an organism [89]. This analysis usually depends at least initially on the organism's phenotype; that is, some characteristic of the organism is observed and its variation among different individuals is studied. Depending upon the frequency of a characteristic's occurrence and how the characteristic varies in relation to other characteristics, models can be built that describe and eventually identify, which genes primarily determine the phenotype. For example, among Gregor Mendel's initial observations in peas were experiments determining pea color [177]. Mendel observed that if he took one line of plants with all green peas and crossed them with one another, all of the progeny[1] peas were also green. Likewise if he took a line of plants with yellow peas and crossed them, the progeny all had yellow peas. However, if he took a plant with green peas and crossed it with a plant having yellow peas, all progeny were yellow. Even more bizarre was his observation that if he crossed the F_1 generation, which were all yellow, with themselves, he obtained a mixture of yellow and green peas, thus recapturing the phenotypes of the parental lines. Mendel's analysis of these observations, built primarily upon enumerating the distribution of pea colors in different crosses, allowed him to propose a theory for inheritance that he could explain empirically based upon his data [101]. In this specific case Mendel had demonstrated not only heritability of a discrete trait, but also dominance as explained by the following simple model:

> If the gene version, or allele, for yellow pea color is represented by "Y", and green pea color is represented by "g", then the parental yellow peas have a genotype YY, while the green peas are gg. A cross between these two is represented as
>
> | Parental Cross | YY x YY | YY x gg | gg x gg |
> |---|---|---|---|
> | F_1 | YY | $Yg + gY$ | gg |
> | F_1 Cross | YY x YY | (gY, Yg) x (Yg, Yg) | gg x gg |
> | F_2 | YY | $YY + Yg + gY + gg$ | gg |
>
> By counting the frequency with which the yellow and green phenotype appeared, Mendel could test his model against observation. For pea color the yellow gene allele, Y, is dominant, so it would be anticipated that the F_2 progeny would have three plants with yellow peas for every one plant that had green peas. Indeed, this was observed experimentally.

[1] Progeny of a parental cross are called the "first filial generation", or F_1. Subsequent progeny are represented as F_2, F_3, and so on.

Since Mendel performed his simple experiments with plants, researchers have expanded upon Mendel's concept of heredity and segregation of multiple traits to identify genes involved with many phenotypes in a variety of organisms. As time has evolved, new methods have emerged from technological innovations. What began with the simple enumeration of phenotypes, has progressed to the analysis of cosegregating phenotypes [11], mapping the relative positions of genes [167, 263], enumerating chromosomal stains and differential banding [66, 162, 212], enumerating the distribution of restriction length polymorphisms [144], and finally on to DNA sequencing [71, 114]. These techniques have been used to explain many different observations ranging from eye color in fruit flies to the occurrence of cystic fibrosis in humans.

Unfortunately, the vast majority of phenotypes are not qualitative, or discrete, like Mendel's yellow and green peas. Instead most characteristics of interest have a continuous, or quantitative, range of variability. For example, a person's height, weight, skin color, or even fasting serum glucose concentration, may vary in an interval that is intermediate to parental attributes. In these circumstances, even crosses between extreme individuals of a population does not yield a Mendelian segregation result, making the observations difficult to explain. This complication arises because a given genotype may produce a range of phenotypes depending upon the environment, and most phenotypes are polygenic, depending upon multiple genes, each of which contributes a small portion to the characteristic. Thus any given phenotype may be represented as:

$$\text{Phenotype} = f(\text{environment}, \; g(\alpha \; \text{gene}_{11}, \; \beta \; \text{gene}_{12}, \; \gamma \; \text{gene}_{21}, \; \delta \; \text{gene}_{22}, \; \ldots))$$

where $gene_{ij}$ denotes allele j of gene i contributing to the observed phenotype, and the Greek letters represent the relative contribution of the gene to the phenotype. In type II diabetes serum blood glucose levels are a continuous phenotype dependent upon the environment (age, diet, climate, physical activity) and a host of different genes.

Geneticists have explained the continuous variation in phenotype by the concept of quantitative inheritance, showing that it is sometimes possible to detect putative *quantitative trait loci* (QTLs) [56]. A QTL is any region in the genome that contributes to a quantitatively measured phenotype. It is anticipated, that within the QTL region a segregating allele (or genetic variant) occurs that gives rise to the observed statistical association with the phenotype variation.

In the 1990's it became possible to systematically map QTLs and over 2,000 different QTLs have been identified in a range of rodent phenotypes including obesity [26, 276] and diabetes [195]. Despite new genetic technologies that improve the feasibility of association studies [109], linkage studies [144], admixture studies and others that can identify QTLs, less than 1% of these QTLs have been characterized at the molecular level [75]; that is, an important region of the genome has been identified, but the actual gene(s) or genetic element(s) contributing to the QTL remain unknown.

The value of QTL analysis to discovering disease genes is in reducing the region of the genome under investigation. Once this has been done, other techniques such as DNA sequencing, positional cloning, and transgenic knockouts can be used to search for genes within the identified locus. By the end of 2001, this approach had resulted in the discovery of 29 disease genes, eight of which were involved in diabetes or obesity [137]. Genes discovered through QTL analysis are often highly penetrant[2], with a large effect size[3]. This is a major drawback to finding all relevant genes to a particular phenotype through QTL analysis alone. In addition, the experiments are time consuming and require a large number of samples: 1,000 animals will only map a QTL contributing 5% of the phenotype variation onto a 10 centimorgan (cM) interval with 50% power [47].

There are a number of other problems with relying upon QTL analysis for determining genes involved in quantitative traits or complex diseases. False positives can still arise, even at the level of the gene coding sequence. For one thing there

[2] Penetrance is the number of individuals within a population that have a specific genotype and the corresponding phenotype. Thus dominant genes, such as those responsible for Mendel's yellow pea color, are almost completely penetrant; that is, all individuals that have an allele for yellow color, look yellow.

[3] Effect size is the amount, or percentage, of phenotypic variation that is attributable to a QTL. A QTL with a large effect size therefore contributes substantially to the observed phenotype.

is degeneracy in the genetic code, thus a number of differentially segregating gene sequences may result in the same protein and proteins with minor variations may not be problematic. Genetic regulatory elements may also be responsible for significant QTLs, however, because of their typically small size (and therefore usually low information content) they may be extremely difficult to isolate, particularly if the gene they regulate lies a considerable distance from the element or outside the QTL. Because of these constraints other techniques are required to help dissect which genes are ultimately involved in various quantitative disease phenotypes and piece together the disease mechanism at the molecular level. It is unlikely that any single technique by itself will be capable of conclusively determining all of the relevant genes, however, combining techniques can build a body of corroborating evidence from which a consensus may arise.

2.2 Genomic Technologies

The human genome project (HGP) was biology's first major foray into the era of big science. Like the physics projects that had come before it, it garnered an enormous level of funding amid unprecedented fanfare for the field. The goals of the human genome project were to [260]:

- Sequence all three billion base pairs of human DNA.
- Determine all of the approximately 20,000–25,000 genes in the human genome.
- Store the sequence and gene information in databases.
- Improve and develop new tools for data analysis.
- Transfer emerging technologies to the private sector.
- Address and discuss ethical, legal, and social issues arising from the project.

It was anticipated that even while the HGP advanced, the sequence data would provide a greater density of genetic markers for researchers mapping traits, particularly

QTLs. Thus one priority was to make as much of the sequence publicly available as soon as possible [186]. Once the entire sequence was complete, it would enable new comparative studies with other organisms and provide the genes that lie within QTL regions.

When the project was first conceived the idea of "genome–scale" experiments was in its infancy, however, midway through the sequencing effort new technologies were emerging that rapidly enabled many researchers to perform very large scale experiments [82]. Currently large, high-through-put experiments are relatively common (in comparison to their occurrence before the HGP) and can investigate DNA sequences, sequence variation, RNA abundance, protein abundance, and metabolite concentrations.

2.2.1 DNA Sequencing

As the human genome project began in 1990, it was well known that it could not be completed using the existing technologies [151, 184, 262]. At that time the two most commonly used sequencing techniques were a chemical degradation method developed by Maxam and Gilbert [156], or an enzymatic method developed by Sanger [213]. Both methods used radioactively labeled DNA fragments that were separated by gel electrophoresis and detected using autoradiography. These techniques were very laborious and time consuming [259], which often precluded their use in QTL gene discovery. Therefore, in addition to funding physical mapping experiments that would help with sequence assembly, the HGP also funded new technology development, particularly in DNA sequencing [184, 262].

Technology funding for the HGP served as a catalyst for the development of high-through-put technologies and methods of analysis [184]. When the project began the largest DNA sequence determined was the 250,000 bp cytomegalovirus sequence that cost approximately $10 per basepair (bp) and required several years to complete [261]. By the end of the project, more than 1,400 megabases (Mbp) per year could be sequenced at a cost of less than $0.09 per finished base pair [39]. These results were extremely impressive and even outperformed the technology goals for the project.

More importantly, the increase in sequencing capacity and decrease in cost, has now made DNA sequencing fairly routine and amenable for most research groups. This provides a technical resource for identifying mutations, genotyping, and determining gene alleles within QTL regions that was largely unavailable before the HGP.

2.2.2 DNA Microarrays

Among the technologies emerging from the HGP were DNA microarrays, which are used to separate DNA from a complex mixture. In its simplest form, the microarray is composed of a substrate (such as a glass slide or nylon membrane) covered with a specific arrangement of known DNA strands or fragments. The idea of covalently attaching DNA to a substrate and using it to probe multiple DNA fragments from a mixture was not new; many researchers had used the technique as an improved method of Southern Analysis (dot blot and reverse dot blot), Northern Analysis, and *in situ* hybridization for gene discovery [229]. In the HGP, researchers were trying to improve the technology for use in sequencing methods known as *sequencing by hybridization with oligonucleotide matrix* (SHOM) [27, 134, 149, 192].

The real innovation came by combining the use of DNA microarrays with reverse transcription using labeled nucleotides. This created a system in which each mRNA could be linearly converted into a labeled cDNA, separated by hybridization to its complementary probe on a DNA microarray, and then quantified by measuring the label abundance as shown in Figure 2-1. The first attempts at transcription monitoring were rather modest: Patrick Brown's 1995 Science paper [215] measured the transcript levels of only 45 genes simultaneously on one array. Today, arrays containing more than 20,000 gene probes are not uncommon [150, 258].

DNA microarrays provide a new, and potentially more efficient, route to finding gene targets involved in quantitative traits and biological processes associated with complex diseases. The core concept is simple: genes that are differentially expressed between unaffected (or control) samples and affected (or experimental) samples, potentially play a role in the observed differences in phenotypes. The information derived does not determine mutations or the sequence of segregating alleles, instead it

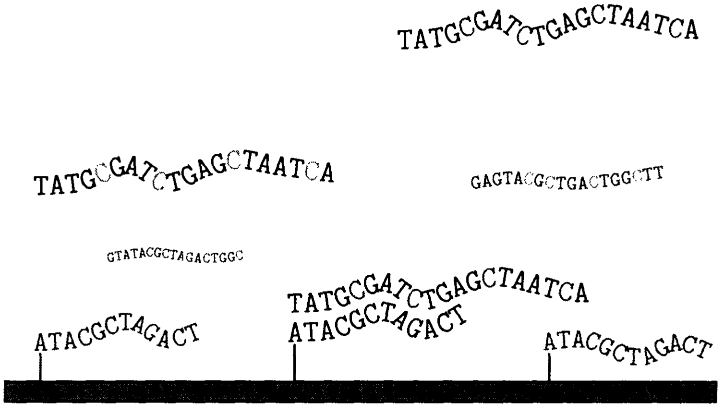


Figure 2-1: DNA microarrays work by exploiting the specificity of DNA base pairing. The initial rules for hybridization were discovered by Erwin Chargaff and dictate that each guanine noncovalently pairs with a cytosine and each adenine is paired with a thymine [32]. The affinity and stability of the hybridized, double stranded DNA is therefore directly related to sequence complementarity. In this figure the labeled "target" molecules, representing the mRNA transcripts, compete for binding to their immobilized *complementary* "probe" molecules on the array. Once equilibrium is achieved, the arrays are washed and scanned to measure the transcript abundance.

quantitatively determines which genes are active or inactive in the environment from which the samples are taken. Thus instead of looking for specific gene mutations that in some way affect the observed phenotype, the cell's "reaction" (as defined by its gene transcription) under one state is compared with the cell's "reaction" under a different state to determine which genes mediate the observed differences in phenotype.

The advantages of using DNA microarrays for gene discovery, particularly with respect to complex diseases, are that they provide information *on actual genes*, do not require as many samples as QTL analyses, are highly parallel, and allow direct, hypothesis based testing on a genomic scale. The fact that microarrays can directly implicate specific genes is a considerable advantage given all of the work that QTL analysis requires. Indeed, combining QTL analysis with DNA microarray results is a complementary approach that has already resulted in the identification of two disease–related genes [137], one of which is involved in insulin–mediated glucose uptake in rats [2]. Another advantage is that DNA microarray analysis typically does not require 100's of samples. So long as the variance in the array measurements can be quantified, direct statistical comparisons of transcript levels can be made with a moderate number of replicates. Additionally, environmental changes can be used to further parse differentially expressed genes and help determine their relevance. Thus if it is known that a certain diet results in insulin resistance in one population, but not another population, more complex experiments involving diet composition can be designed to more accurately find the relevant genes involved.

The caveats of using DNA microarrays are that changes in gene transcription alone may not be responsible for phenotypic changes, and analysis can be challenging when confronted by 20,000 different transcript measurements. It is often, wrongly, inferred that changes in transcript levels correlate to changes in protein levels, or even worse, changes in protein activity; that is, sometimes it is assumed that translation occurs with little or no regulation and that transcription is the dominating effect. Certainly this is not true in a large number of cases [35, 92]. While increases or decreases in transcription *may* alter protein levels, there is no single correlation or function that tells how the concentration of mRNA is linked to the concentration of protein. Since it is often held that most phenotypes are the results of protein activity, measuring transcript levels alone does not necessarily correlate with a given phenotype. Thus the change in the level of a specific mRNA, although highly correlated with the phenotype, does not necessarily mean it causes the changes in phenotype.

There are currently two DNA microarray technologies that are commonly used. One is a high density oligonucleotide system commercially available through Affymetrix[4] (Santa Clara, CA), the other is typically referred to as a "cDNA system." While there are substantial differences between the two types of technologies [141], both quantify the distribution of transcripts from a pool of RNA. While the technology surrounding DNA microarrays continues to evolve, we developed and validated an assay for use in our laboratory using the "cDNA system."

Our transcription monitoring system incorporates a fairly diverse set of experimental and computational methods, from chemistry to molecular biology to image

[4] The National Human Genome Research Institute provided funding for DNA Microarray research that helped establish Affymetrix [39].

and signal processing, all of which can affect the data. Thus, we first standardized our protocols to understand what experimental artifacts may be introduced and how the data may be affected by each step in the procedure. Our starting point for assay development was based largely on previous published work done in similar laboratories.

DNA Microarray Development

To investigate genes that may mediate biochemical processes involved in Type II diabetes, we developed a DNA microarray assay to measure transcript levels of over 16,000 mouse genes. The assay itself is relatively simple and uses the following procedure:

- Microarrays are printed with known DNA *probe* sequences on a substrate and then blocked prior to hybridization.
- RNA is isolated from tissue or cell samples of interest.
- The RNA is labeled during reverse transcription with fluorescent nucleotides.
- The labeled cDNA *target* is hybridized to the microarray.
- The microarray is washed and then scanned with a laser to quantify the amount of label hybridized to each probe.

Each of these steps requires some optimization and study in order to produce reliable data.

To develop our microarray printing protocols we conducted a large set of control experiments detailed below. These experiments usually used either a red, water soluble dye with high autofluorescence, or COT-1 DNA with a stain.

In studying each of the sample preparation steps we have relied primarily on control experiments where the same sample is divided in two and each half labeled with either Cy3– or Cy5–dCTP dyes. The labeled samples are then hybridized to the same array and should yield expression ratios of unity for every probe. In this way we can evaluate the variance in the ratios obtained, across a wide range of signal intensities and experimental conditions.

Printing and Preparing DNA Microarrays

Printing of DNA microarrays containing several thousand DNA probes, or *features*, is still an art in that the printing process must be developed and optimized empirically. Because of the many variables (robotics, array substrate, array surface chemistry, printing buffer, humidity, pin type, pin wear, pin blotting, pin capacity, probe type (cDNAs or oligonucleotides), probe concentration, library size, run time), each of which can change substantially based upon the specific system and laboratory, there is no simple general way of optimizing array printing. Fortunately there are a number of references on each of these subjects [103, 215, 220, 229, 257], and several good on–line resources to guide development:

- The BioMicro Center at the Massachusetts Institute of Technology
  - http://biomicro.mit.edu/forms/index.htm
- Microarrays.org, a public source of protocols and software hosted by the University of California, San Francisco
  - http://www.microarrays.org/index.html
- The Institute for Genomic Research
  - http://www.tigr.org/tdb/microarray/
- Patrick Brown's laboratory at Stanford University
  - http://cmgm.stanford.edu/pbrown/mguide/index.html
- Whitehead Institute Center for Microarray Technology
  - http://www.wi.mit.edu/CMT/Microarrayhome.html

To print DNA microarrays we used a Versarray Chipwriter Pro arrayer from Virtek (now owned by Bio-Rad), shown in Figure 2-2. This robotic arrayer can print 126 arrays per run and has automated pin cleaning, a robotic arm designed for handling 96 or 384-well plates, an environmentally controlled printing chamber, and is user programmable. In conducting a series of control experiments to optimize the printing conditions we found that pin cleaning was primarily dependent upon pin drying (when any residual material was effectively removed by the vacuum), that the washing station water cycle required monitoring, and that the mechanical arm for handling the plates of the probe library was not reliable and also had to be monitored.

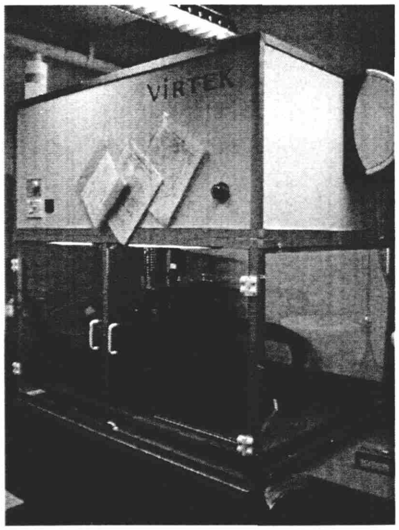


Figure 2-2: Virtek Chipwriter Pro Arrayer used to print DNA Microarrays.

In developing our microarray assay, we tested a variety of printing conditions using three different array substrates: Corning Gamma-Amino Propyl Silane (GAPS) slides, Telechem Superamine slides, and Cel Associates poly-L-lysine coated slides. The chemistry on each of these substrate surfaces is similar in that the functional group for binding DNA is a primary amine, which can provide two kinds of interactions [233]. The amine moiety can bind DNA through ionic interactions between positively charged amino groups and the DNA's negatively charged phosphodiester backbone. The other interaction is through the UV catalyzed formation of covalent bonds between thymidine residues and the alkyl chains to which the amines are attached. Of these, the Telechem Superamine slides had significantly lower signals compared to the GAPS and poly-L-lysine slides. The GAPS slides were ultimately selected for array printing because of poor quality control by Cel Associates in producing the poly-L-lysine slides, which commonly had visible surface imperfections.

The probe library that was printed onto the arrays contained 17,021 DNA probes, supplied in 44 384–well round bottom Genetix plates. To conduct a single print run of 126 microarrays using 16 printing pins would take approximately 72 hours. Thus 16 new Telechem pins were purchased, cutting the arraying time down to 36 hours per run. When the new pins arrived, they had to be "worn in" to remove any metal burrs or shavings and ensure they were capable of proper operation in the Telechem printhead.

Several print runs were conducted using either herring sperm DNA or COT-1 DNA in 3X sodium chloride–sodium citrate (SSC) or dimethyl sulfoxide (DMSO) buffers at different humidity settings. The humidity level can vastly affect spot morphology and size, depending upon the buffer used. We found that DMSO worked well as a buffer at humidity levels between 10% and 30%, while humidity levels greater than 30% lead to large spots and very poor morphology. In contrast, 3X SSC worked well at elevated humidities between 40% and 60%. At humidity levels below 40%, features printed in 3X SSC became very small and did not attach well to the surface. Because the ambient humidity level in the laboratory was routinely above 30%, we chose 3X SSC as our printing buffer. This allowed us to control the humidity at approximately 50% using the arrayer's environmental control, and enabled the printing of 100 micron features, spaced 175 microns apart (center-to-center) with very good spot morphology.

For each probe in the library, 600 pmol of material was provided. The recommended printing concentration was 40 - 50 μM, however, adequate printing and array signals from concentrations as low as 10 μM had been communicated by other users [252]. Using a subset of purified cDNA and oligonucleotide probes and RNA target, the probe printing concentration was tested at 0.1, 10, and 40 μM, as shown in Figure 2-3 on the following page. When the corresponding ratios were compared at different probe concentrations, they were highly correlated indicating that although signal intensities varied, the signal ratios, which represent the primary data obtained from a microarray, were largely unchanged as shown in Figure 2-4 on the next page. In Figure 2-4, for each data series the X–axis represents the fluorescence ratio of the probe printed at the lower concentration and the Y–axis represents the ratio for the probe printed at the higher concentration. For example, blue diamonds show the comparison of signal ratios obtained by hybridizing samples on arrays printed with 0.1 μM (X–axis) and 10 μM (Y–axis) probe concentrations.

In order to balance the number of arrays that could be obtained from the library with the signal intensity, we printed our arrays at 20 μM. This concentration allows printing of 1,200 arrays from the library.

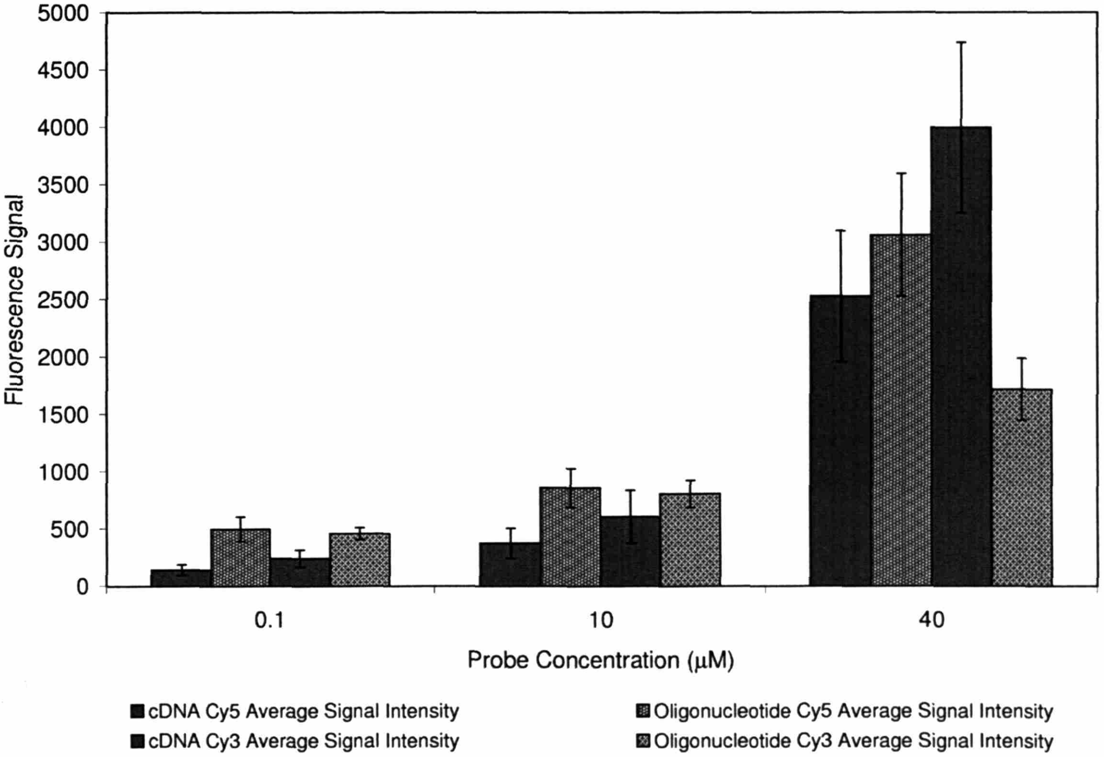


Figure 2-3: Average signal intensity of cDNA and oligonucleotide probes as a function of printing concentration.

The mouse gene library we printed on our arrays contained 17,021 probes based on the Unigene Database, build Mm 102 (www.ncbi.nlm.nih.gov/Unigene/). This library (Operon Mouse Oligo Set Version 2) was composed of 70-mers representing 16,463 mouse genes from the UniGene Database [266]. The UniGene database automatically

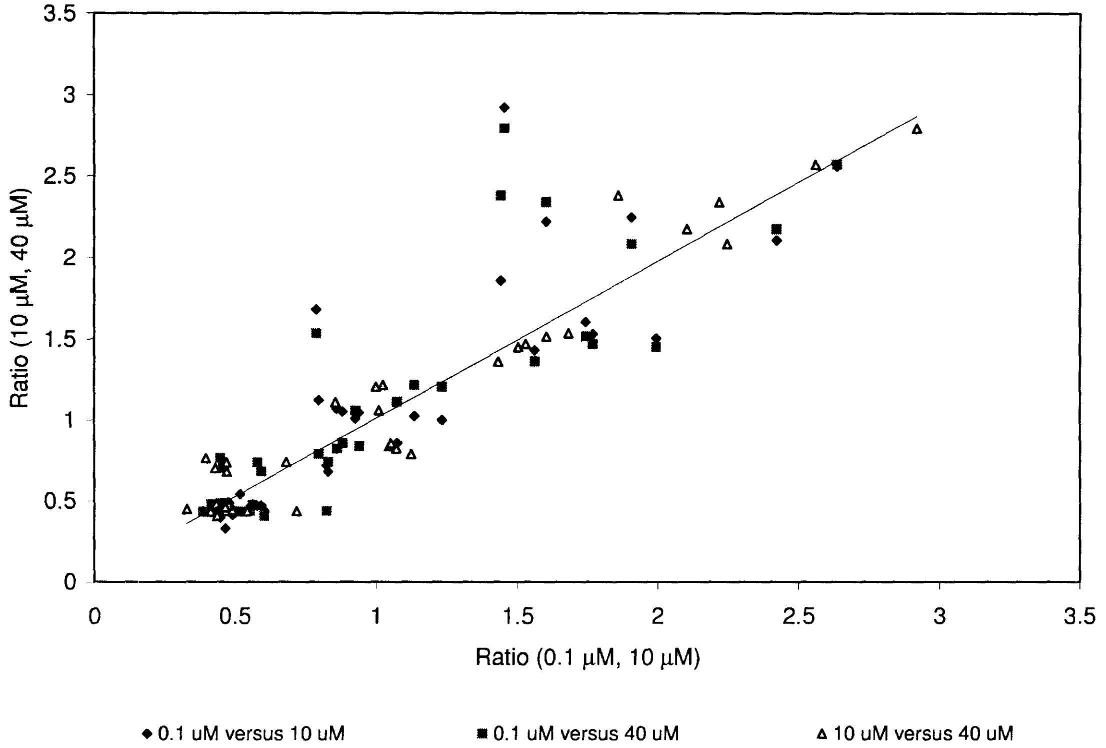


Figure 2-4: Comparison of probe ratios across different probe printing concentrations.

clusters all mouse sequences in GenBank into a non-redundant set of genes. Each cluster in the UniGene Database represents one unique gene, representative of all cluster sequences, based on the longest region of high-quality sequence data.

Operon tries to minimize cross–hybridization by optimizing their oligonucleotide sequences using BLAST (Basic Local Alignment Search Tool). This is not an optimal strategy for probe selection as small homologous regions have been shown to contribute significantly to cross hybridization [125], and BLAST does not evaluate these regions well because it is based on pair-wise similarity, as opposed to block similarity [70, 146]. In addition, the melting temperature, T_m, of each 70-mer probe is normalized to 78 °C (± 5 °C) based upon its nucleotide (GC) content. This temperature reflects the probe affinity for its complementary target, with higher temperatures indicative of greater affinity. The T_m normalization ensures a consistent and stringent hybridization across all selected probes. When a 70-mer could not be designed within the specified T_m range, a shorter or longer sequence is substituted to maintain the

T_m criteria.

Three print runs were conducted in series using these parameters:

- Corning GAPS slides
- 3X SSC buffer
- 20 μM probe concentration
- 50% — 60% humidity
- 32 printing pins

Each print run contained 126 slides, printed all 45 plates (17,021 features) of the library, and required approximately 50 minutes per plate, hence a single run took approximately 35 hours to complete. The first run printed each feature one time on each slide without redipping the pins. SYBR stains (Molecular Probes) following the print run showed that array 126 did not possess all features, and visual inspection implied that arrays 110 and after were deficient in some spots. For this reason, pins were redipped in the second and third print runs. Redipping occurred for every gene following printing of the 60th slide. After the 60th slide was printed, the pins returned to the wells, reloaded, blotted eight times, and then restarted printing with slide 61. All of the resulting slides contained all features.

After each print run, the probes were crosslinked to the microarray using a Stratagene Crosslinker (Stratagene, La Jolla, CA). Microarrays were subsequently blocked using boric acid and methyl-pyrrolidinone. The blocking inactivates the surface where features are not printed and prevents substantial binding of the target cDNA in those areas [53, 238].

RNA Isolation and Labeling

Our RNA harvesting methods primarily used the commercially available RNeasy RNA Extraction kit available from Qiagen (Valencia, CA). Briefly, the procedure entails lysing the cells in a guanidine isothiocyanate buffer (that is highly denaturing and inactivates RNases), homogenizing the sample, adding ethanol and adjusting the pH, then binding the sample to a silica-gel based membrane. Contaminants are washed away, and the RNA is eluted in water. The procedure can be used to directly lyse adherent cells while still attached to the flask, or following trypsinization and cell harvest. The yields of the RNA isolation are quantitative, depending upon the cell line and conditions used, with a coefficient of variation between 10 - 30%. Figure 2-5 shows the yield of RNA as a function of cell number.

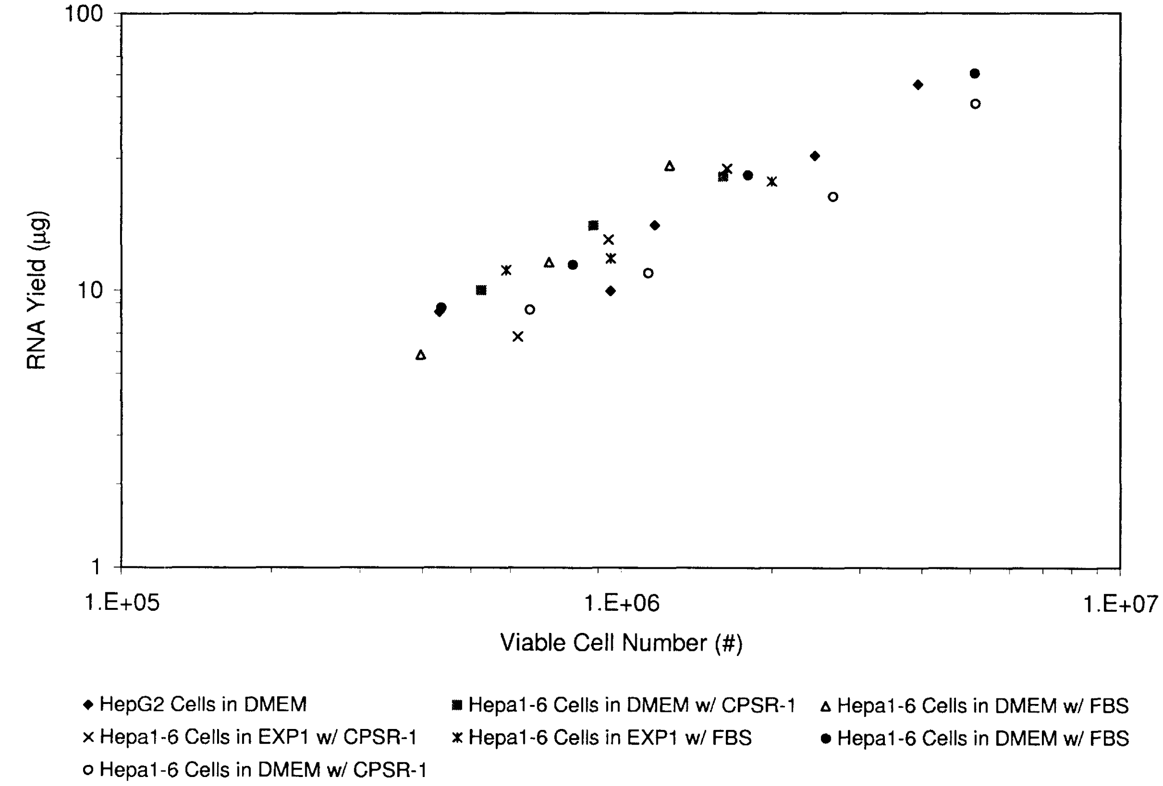


Figure 2-5: Yield of RNA as a function of cell number.

Sample labeling relies upon the direct incorporation of CyDye (Amersham Pharmacia) labeled nucleotides into cDNA during the synthesis reaction. This procedure has been used historically in the laboratory, and the main differences between the present and former methods are the amount of total RNA used, the primer used for first strand synthesis, and reaction cleaning protocol. The results showed that no changes were obtained in the fluorescent ratios when the initial amount of total RNA used in the assay was between 8 μg and 20 μg. Based on these tests and the yield of RNA from culture samples, 10 μg of RNA is used with most sample preparations. For mammalian RNA processing we have switched to using a mixture of poly-dT priming and random hexamers. This, in theory, should give a cleaner sample, with a lower potential for cross hybridization from cDNA derived from rRNA and tRNA present in the total RNA sample. Sample clean-up used the Nucleotide Removal kit (Qiagen) which removes the excess dye and oligonucleotides smaller than 17 base pairs in length. The binding capacity of this kit was shown to be more than adequate for our samples. In addition to these procedural changes, the RNA degradation step was also investigated. This work showed that completed degradation of the rRNA bands was accomplished using our procedure, indicating potent degradation of residual RNA following cDNA synthesis.

cDNA Hybridization, Scanning, and Data Acquisition

The hybridization protocol was investigated extensively, however, was not markedly altered from the traditional protocol developed in the lab. In this regard, the Corning hybridization chamber, hybridization temperature, and the hybridization time have been investigated. The other hybridization chamber tested, from Stratagene, lead to a greater degree of variation in the data and was therefore eliminated from further consideration. Hybridization time was investigated over one and two day time intervals. We selected single day hybridization because it resulted in higher fluorescence signals.

The effect of temperature was studied as a way of altering the stringency of the hybridization. Using a non–complementary set of control probes from *Arabidopsis thaliana*, it was found that hybridization at 55 °C lead to the least amount of cross hybridization as shown in Table 2.1. At 50 °C significant fluorescent signals were obtained from the non–complementary control probes, while at 55 °C or greater, signals from the non–complementary probes were eliminated.

Fifty–five degrees Celsius was chosen for the hybridization temperature because

		0.1 mM			**10 mM**			**40 mM**			**PCR Products**		
Hybridization				# Features			# Features			# Features			# Features
Temperature	Gene	Cy5	Cy3	Detected	Cy5	Cy3	Detected	Cy5	Cy3	Detected	Cy5	Cy3	Detected
50 C	*Cab1*	202	131	3	372	408	15	1695	1929	14	847	1120	15
	LTP4	196	242	5	1525	1815	16	2410	2906	15	-	-	0
	LTP6	-	-	0	-	-	0	-	-	0	-	-	0
	NAC1	-	-	0	-	-	0	142	124	1	-	-	0
	PRKase	-	-	0	-	-	0	185	240	2	-	-	0
	rbcL	239	296	2	492	731	4	8217	10680	16	-	-	0
	RCA	207	215	5	639	684	15	4573	5291	14	2337	3060	16
	RCP1	1696	1614	15	3882	4056	14	18857	16691	15	682	737	9
	TIM	116	141	1	246	288	8	1311	1563	16	-	-	0
	XCP2	-	-	0	-	-	0	207	296	1	809	1214	14
55 C	*Cab1*	-	-	0	-	-	0	-	-	0	-	-	0
	LTP4	-	-	0	-	-	0	-	-	0	-	-	0
	LTP6	-	-	0	-	-	0	-	-	0	755	449	4
	NAC1	-	-	0	-	-	0	-	-	0	-	-	0
	PRKase	-	-	0	-	-	0	-	-	0	-	-	0
	rbcL	-	-	0	-	-	0	933.8	980	4	-	-	0
	RCA	-	-	0	-	-	0	-	-	0	-	-	0
	RCP1	-	-	0	-	-	0	219.4	192.8	5	2342	1414	10
	TIM	-	-	0	-	-	0	-	-	0	-	-	0
	XCP2	-	-	0	-	-	0	-	-	0	-	-	0
60 C	*Cab1*	-	-	0	-	-	0	-	-	0	-	-	0
	LTP4	-	-	0	-	-	0	-	-	0	-	-	0
	LTP6	-	-	0	-	-	0	-	-	0	-	-	0
	NAC1	-	-	0	-	-	0	-	-	0	-	-	0
	PRKase	-	-	0	-	-	0	-	-	0	-	-	0
	rbcL	-	-	0	-	-	0	-	-	0	-	-	0
	RCA	-	-	0	136	27	1	-	-	0	-	-	0
	RCP1	-	-	0	-	-	0	-	-	0	-	-	0
	TIM	-	-	0	-	-	0	-	-	0	-	-	0
	XCP2	-	-	0	-	-	0	-	-	0	-	-	0

Table 2.1: *Arabidopsis thaliana* gene probes. These gene probes have no known homology to any mammalian genes and therefore should not bind labeled cDNA from mammalian samples. Conditions are selected that eliminate signals from these probes to optimize the hybridization stringency. The probe printing concentration (0.1 mM, 10 mM, or 40 mM) is shown at the top of each set of columns; PCR products were probes generated using PCR fragments for these genes. Numbers in the Cy5 or Cy3 columns represent the fluorescent signal intensity for that gene.

at temperatures of 60 °C or greater, the signal intensity of the Cy5 channel was substantially decreased as shown in Figure 2-6. In contrast fluorescent signals from the Cy5 channel routinely ranged from 10's to 10,000's, enabling a stringent assay that could detect signals over three orders of magnitude.

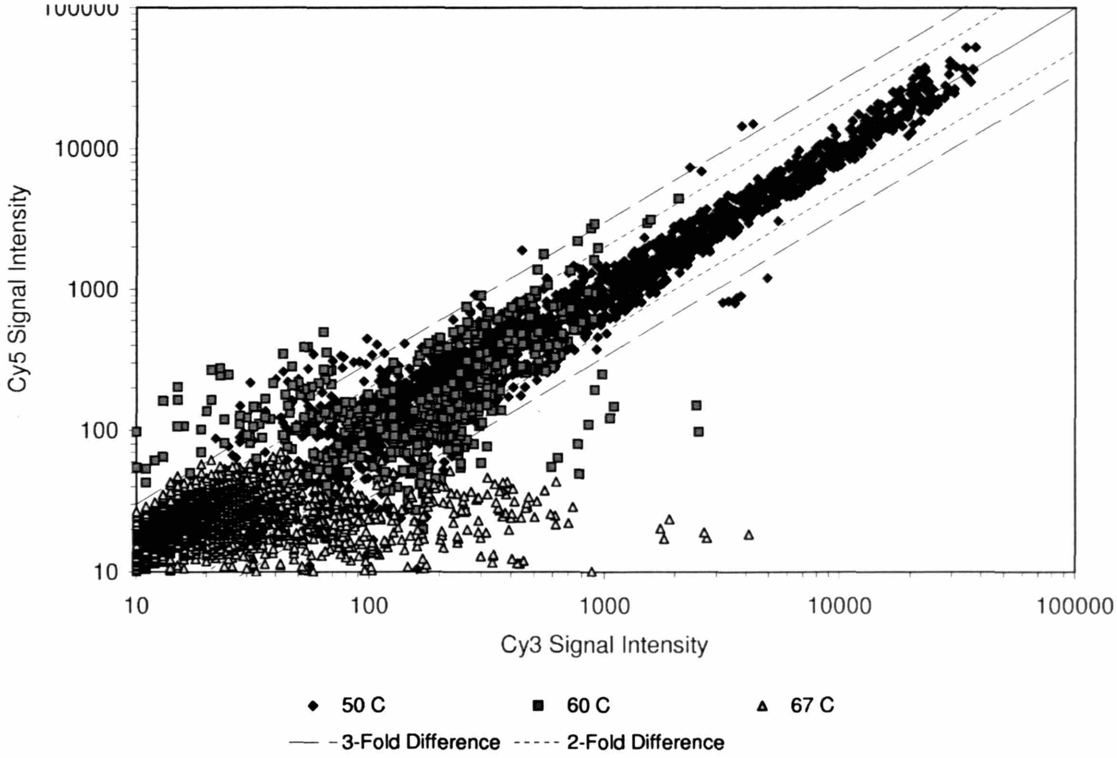


Figure 2-6: Effect of hybridization temperature on fluorescence signals. This figure shows the same sample labeled with both Cy3 and Cy5, hybridized at different temperatures. In theory the probe signals should be equal in both channels, which is closely approximated by the 50 °C samples. However, as temperature increases, gradual loss of the Cy5 signal is observed.

Once the hybridization is complete, the microarrays were washed at high agitation. For washing we used a commercially available solution from Clontech (Mountain View, CA) and 3X SSC. Microarrays were dried by low speed centrifugation.

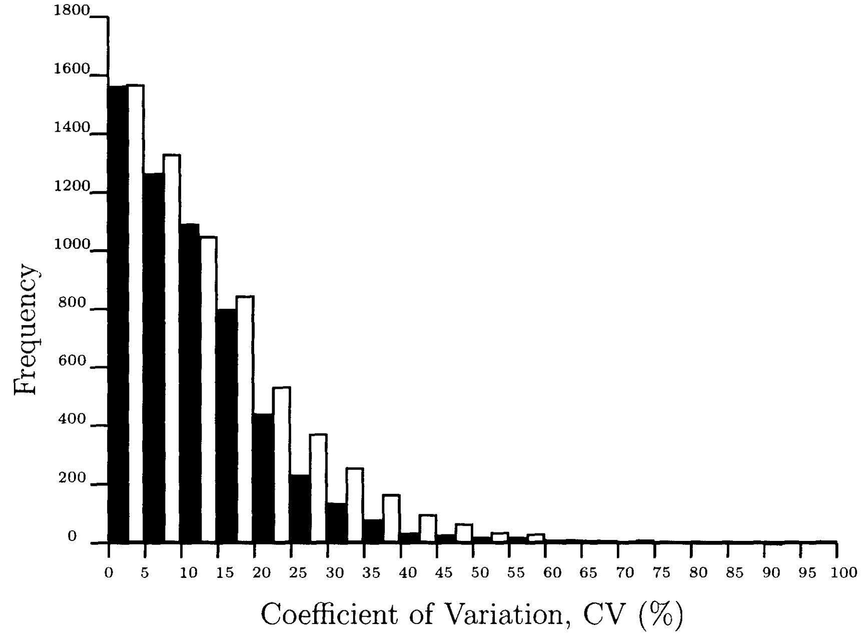
After drying, microarrays were scanned with a Genepix Scanner (Axon Instruments) to obtain the data. It was found that the spot intensity, as well as the background intensity, were increased at higher laser PMT voltages and that repetitive scanning of the arrays could affect the ratio data. Therefore all scans were conducted as close to 600 V as possible for both laser channels. It was also found that signals tended to fade with time following the hybridization, however the ratios largely remained unchanged (indicating that the fluorescence signal for both dyes degraded at approximately the same rate). This degradation of signal intensity with time occurred over days and is insignificant on the time scales required to scan the microarrays following array washing.

Data analysis required as input the gene identifiers from the Genepix (scanner) data output, the ratios of means of the Cy5/Cy3 pixel fluorescence corresponding to each gene identifier, the ratio of means normalization factor (used to correct for the difference in fluorescence and incorporation of the Cy5 and Cy3 fluorophores), and the adjusted flags corresponding to each gene identifier. The adjusted flags marked features that either did not possess basic fluorescence requirements sought by the imaging software or had less than 60% of their pixels' intensities greater than two standard deviations of the background intensity over the background intensity. Using these data as input, `Matlab` returned the gene identifier, the mean of replicate spots that were not flagged by our filters, the standard deviation of the unflagged replicates, the coefficient of variation among replicates, and the number of observations used in calculating the statistics for each gene.

DNA Microarray Validation

The developed DNA microarray protocols were extensively validated in the laboratory as described previously [28]. For validation, we prepared arrays containing an approximately 13,000 gene sub–set of our oligonucleotide mouse library, printed in triplicate. Total RNA from skeletal muscle and brain tissue were used for validation comparisons, and each sample was analyzed in duplicate and prepared using our standard protocols. `Matlab` was used to calculate basic statistics.

The coefficient of variation (CV) was calculated for each replicated gene expression and the distribution across all genes is plotted in Figure 2-7. For the muscle versus muscle control arrays, the median CV across all probes was 10.2%. For the muscle versus brain arrays the median coefficient of variation across all probes was 9.8%. This indicates that for a gene transcription ratio of 1, we might expect the true value to lie between 0.9 and 1.1; similarly for a gene transcription ratio of 3, we might expect the true value to lie between 2.7 and 3.3.



Black Bars = CV for Muscle versus Muscle Comparisons
White Bars = CV for Muscle versus Brain Comparisons

Figure 2-7: Distribution of the coefficient of variation for DNA microarrays. The coefficient of variation was calculated for every gene on the microarray and plotted for the muscle versus muscle and muscle versus brain arrays.

Although the median CV across all probes for the muscle versus muscle control arrays was 10.2%, some genes had a greater CV. The 314 genes on the muscle versus muscle arrays with a fold difference greater than two common had a median CV of 24.7%. Because of their increased CV and high fold change, these 314 genes were eliminated from subsequent analysis.

The arrays' ability to detect differential transcription between muscle and brain RNA was evaluated by two different methods. In the first, we examined the number of genes that were up- or down-regulated by a factor greater than two (that is, whose mean ratio was either greater than two, or less than 0.5) in the muscle versus muscle and the muscle versus brain RNA comparisons. This criterion has been used as a basis for assessing differential transcription in a number of studies [119, 208, 256].

In the second method, we defined a threshold for differential expression by using the 95% confidence interval determined from the muscle versus muscle control arrays. Table 2.2 summarizes the results, where the p-values reported were from two–tailed student t-tests.

Array Condition	# of Probes Detected	# of Genes >2-Fold Different	Differentially Expressed Genes at the 95% Confidence Level
Muscle vs. Muscle	7574	438	429
Muscle vs. Muscle	6417	314	302
Average	6996	376	366
Muscle vs. Brain	7143	1201	1161
Muscle vs. Brain	8318	981	931
Average	7731	1091	1046
P-value	0.47	0.03	0.03

Table 2.2: Differential gene transcription validation data. This table summarizes results of the array validation with respect to the study of differential expression.

Although in Table 2.2 there are only about 370 genes exceeding the threshold in the muscle versus muscle arrays, more than 1000 genes were differentially expressed in the muscle versus brain arrays. This result supports the assertion that the microarray assay method and selection criterion are significantly more likely to identify differentially expressed genes.

In duplicate arrays, 76% of the genes observed on one muscle versus muscle array were also observed on the duplicate; likewise 77% of the genes found on one muscle versus brain array were conserved on the duplicate. These data demonstrate the inter-array reproducibility by showing the majority of genes are reproducibly found in multiple replicate arrays.

In addition we were able to demonstrate the specificity of our microarrays by patterning specific probes on the array, and then labeling the complementary targets with either Cy3-dCTP or Cy5-dCTP. As shown in Figure 2-8 the labeled target cDNA very specifically binds its respective probes, creating the observed pattern. In Figure 2-8, the "red" probes surround the pattern bind cDNA target labeled with Cy3, while the "green" probes making up the letters MIT are labeled with Cy5. The

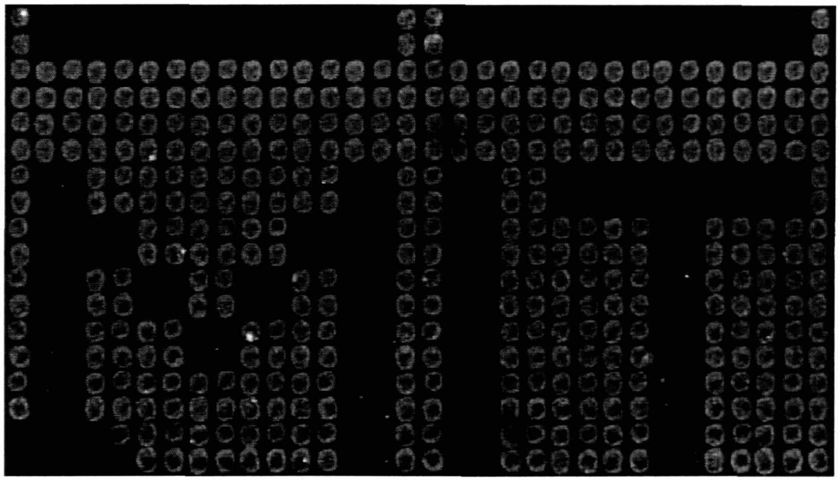


Figure 2-8: Test of DNA microarrasy specificity using patterned probes and purified RNA.

orange spots at the top of the figure are a mixture of the two probes and give signals of approximately equal intensity in both channels.

RT–PCR was also used to verify some of the array results. For each of the genes investigated the variation in the ratios of the mRNA levels between the array results and RT–PCR was less than 30% as shown in Table 2.3.

Genes	Assay	High–Fat vs. Control	F/ WR vs. Control
IL6st	Array	154 ± 21% ⋆	144 ± 21% ⋆ †
	RT–PCR	167 ± 19% ⋆	185 ± 15% ⋆ †
PTP4a2	Array	71 ± 4% ⋆	89 ± 3% ⋆
	RT–PCR	75 ± 16%	94 ± 18%
RGS3	Array	35 ± 5% ⋆	54 ± 8% ⋆
	RT–PCR	38 ± 9% ⋆	59 ± 8% ⋆

Table 2.3: Comparison of array results and RT–PCR results for selected genes. Data are expressed as a percent of the control expression levels. F/ WR: Fasting/ weight reduced mice

⋆Indicates that the measurements were significantly different from control values at $P < 0.01$.
†Indicates that the measurements made on the microarray were significantly different from the RT–PCR measurements at $P < 0.05$.

2.3 Data Analysis

Valuable information can be extracted from microarray data by using statistical and data mining methods. Statistical methods rigorously quantify the reliability of differences in the microarray data [171] and can objectively evaluate changes in gene transcription ratios and derivative quantities [245]. Data mining is particularly useful for uncovering patterns and structure in microarray data that might have otherwise been difficult to detect through manual inspection and intuition alone [143, 200]. Applying statistics and data mining methods to microarray data in unison enables rapid and reliable analysis without *a priori* assumptions.

Selection of a particular analysis method depends largely on the experimental design. While we explored the use of each method introduced in this chapter, subsequent chapters rely primarily on Fisher Discriminant Analysis and Principle Components Analysis because of our interest in classifying samples based upon their experimental treatments. Each of these methods is described in the literature and only a brief overview is given here.

2.3.1 Statistics

Many statistical methods can be used to analyze the gene transcription data [20, 130, 131, 271]. Use of any particular method is highly dependent upon the experimental design and type of microarray technology used.

To assess differential gene expression, a gene by gene t-test [124, 241, 264] can be applied to evaluate statistically significant expression differences in pairwise comparisons between the control and experimental samples. Another useful method is Wilks-λ based ranking [54, 115, 121]. This technique is particularly appropriate for *multi-class* comparisons, ranking genes on the basis of their within group, and between group variances. Thus, a gene exhibiting a small variation within each of several groups, but large variation between groups would rank highly; conversely a gene that had a high level of variation within a group, and a low level of variation among the groups would be ranked low. The Wilks-λ score can be transformed into an F statistic, which can be compared with the F distribution to assess the statistical significance of the observation.

2.3.2 Multivariate Analyses

Multivariate analyses are data mining techniques used to extract information from data with many variables. Thus, as opposed to statistical techniques that often focus on the mean and variance of one variable, or differences in pairwise comparisons, multivariate techniques focus on covariances or correlations [54, 121]. In a sense, they attempt to uncover structure in the data set and identify what are the most important variables. In microarray analysis, where numerous genes can be simultaneously measured, these techniques provide a way of quickly finding important relationships among the samples and genes. There are many different methods, however, Fisher Discriminant Analysis, Principle Component Analysis, and Partial Least Squares were the three most commonly used and explored in this work.

Fisher Discriminant Analysis

Fisher Discriminant Analysis [54, 121, 123, 230] (FDA) is a method that can be used to determine combinations of genes that are capable of correctly classifying samples. Thus if RNA samples were taken from normal mice, diabetic mice, and diabetic mice receiving some treatment, FDA could be used to find genes whose expression classifies the samples according to their collective gene transcription profiles. In this regard, FDA is considered a *supervised* data analysis method, in that it is told at the outset which samples belong to which classes. Conversely, a subset of the samples can be used as a training set to develop a model that predicts the membership or other samples.

The results of FDA are based on *linear combinations of gene expressions* that consider the discriminatory power of gene groups as opposed to individual genes. Samples are scored based on the weighted contributions of each gene to a newly defined metric called a canonical variable. Because each gene's contribution to a sample's score is weighted by a coefficient called a "loading," genes with very small loadings do not significantly contribute to the sample's score and classification, and can therefore be eliminated from further consideration. Samples are scored according to

$$S = \sum \lambda_1 g_1 + \lambda_2 g_2 + \ldots + \lambda_i g_i + \ldots + \lambda_n g_n \tag{2.1}$$

where S is the sample score, λ_i represents a gene's loading, g_i represents a gene transcription level, and the sum occurs over all discriminatory genes, n.

This technique can be used as a powerful tool to visualize microarray results in a lower dimensional space defined by the canonical variables. The canonical variables are one dimensional metrics calculated as a weighted linear sum of the other variables, in this case gene expressions. The underlying principle is that if the scores accurately classify the samples, then the genes selected to determine the scores are discriminatory of the treatments examined when sample classification is used as a criterion.

In FDA the canonical variables, $\mathbf{V}$, are selected so as to maximize class separation [123]. These variables are determined as the eigenvectors of the inter–group variance, $\mathbf{B}$, scaled by the intra–group variance, $\mathbf{W}$, as

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{V} = \mathbf{V}\Lambda \tag{2.2}$$

where

$$\mathbf{B} = \mathbf{T} - \mathbf{W} \tag{2.3}$$

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}\bar{X}^T)^T(\mathbf{X} - \mathbf{1}\bar{X}^T) \tag{2.4}$$

$$\mathbf{W} = \sum(\mathbf{X}_j - \mathbf{1}\bar{X}_j^T)^T(\mathbf{X}_j - \mathbf{1}\bar{X}_j^T) \tag{2.5}$$

and the sum occurs over all of the sample classes. In this formulation $\mathbf{X}$ represents the (n samples x g genes) data matrix, $\mathbf{T}$ represents the total variation among all the data and the eigenvalues, Λ, indicate the discriminatory power of the canonical variables.

Principle Components Analysis

Principle component analysis (PCA) is similar to FDA in that it can be used as a data reduction technique or to find structure in a data matrix. PCA reduces the original set of variables (in this case genes) into a smaller, orthogonal set that is composed of linear combinations of the variables of the original set. Unlike FDA, PCA is *unsupervised*, that is, it does not require the *a priori* assignment of samples to a specific class. Instead the coordinates of the smaller variable set are chosen such that they capture as much of the total variance as possible in the original data. In this way, it may be possible to identify groups of genes or samples that show similar behavior. Like FDA, using PCA can help identify groups both computationally and visually.

The procedure for using PCA is straightforward [3, 54, 164]. For a given data matrix composed of n samples and g genes, the data may be scaled and is usually transformed into a covariance or correlation matrix. If we let $\mathbf{X}$ represent the original n x g data matrix, then the covariance matrix, $\mathbf{C}$, is defined as:

$$\mathbf{C} = \left[\frac{1}{n-1}\right]\left[\mathbf{X}^T\mathbf{X} - \left(\frac{1}{n}\right)(\mathbf{X}^T\underline{1})(\underline{1}^T\mathbf{X})\right] \tag{2.6}$$

where matrices have been denoted in bold, vectors are underlined, and the transpose is indicated by the superscript T. Likewise the correlation matrix, $\mathbf{R}$, can be calculated as:

$$\mathbf{R} = \left[\frac{1}{n-1}\right]\left[\mathbf{D}^{-\frac{1}{2}}\left[\left(\frac{1}{n-1}\right)\left(\mathbf{X}^T\mathbf{X} - \frac{1}{n}(\mathbf{X}^T\underline{1})(\underline{1}^T\mathbf{X})\right)\right]\mathbf{D}^{-\frac{1}{2}}\right] \tag{2.7}$$

where $\mathbf{D}^{-1/2}$ is defined as the matrix whose main diagonal elements are the reciprocals of the standard deviations of the g genes in $\mathbf{X}$.

To identify the principle components, the set of vectors of coefficients, $\underline{\mathrm{y}}_1$, $\underline{\mathrm{y}}_2$, $\ldots$, $\underline{\mathrm{y}}_i$, $\ldots$, $\underline{\mathrm{y}}_{m-1}$, $\underline{\mathrm{y}}_m$, such that y^T $\mathbf{X}$ is maximized over all linear combinations of $\mathbf{X}$ with the constraint y^T $\mathrm{y} = 1$ for all coefficient vectors, is sought. To find this set of vectors, it has been shown that they must satisfy g simultaneous equations of the form

$$(\mathbf{C} - \lambda_i\mathbf{I})\underline{y}_i = 0 \tag{2.8}$$

or equivalently, depending upon the input matrix

$$(\mathbf{R} - \lambda_i\mathbf{I})\underline{y}_i = 0 \tag{2.9}$$

This is the common eigenvalue, eigenvector problem. Nontrivial solutions for the eigenvectors, $\underline{y}_i$, can be found by solving for the eigenvalues, λ_i, of the determinant

$$|\mathbf{C} - \lambda_i\mathbf{I}| = \mathbf{0} \tag{2.10}$$

or

$$|\mathbf{R} - \lambda_i\mathbf{I}| = \mathbf{0} \tag{2.11}$$

The determinant of these equations results in a polynomial of order g; hence the g roots associated with the polynomial are the eigenvalues. From this set, the first principle component can be identified by choosing the largest eigenvalue (root of the polynomial) and then solving for the corresponding eigenvector. This eigenvector gives the coefficients of the variables, genes in this case, of the first principle component. The procedure is then repeated for each of the subsequent g eigenvectors with the constraint that the principle components must be mutually orthogonal. Other methods of calculating the principle components are possible such as orthogonal decomposition of the input matrix or by using nonlinear iterative partial least squares [85, 207].

Because PCA is not scale invariant, using either the covariance or correlation matrix will affect the solution obtained, and there is no way of relating the solutions from the two different matrix transformations. For this reason it's prudent to conduct both transformations and run the analyses in parallel.

Partial Least Squares

In quantitative genetics, the relationship between the environment, genotype, and resulting phenotype is called the "*norm of reaction*" [51]. The norm of reaction dictates how a given distribution of environments maps onto a distribution of phenotypes for a specific genotype. For example, in the study of diabetes it would be particularly interesting to determine why certain ethnic groups (representing the genotype being studied) within a given environment (representing the independent variables) have elevated glycemia or increased risk for Type II diabetes (representing the dependent variables that describe the phenotype of interest). To deal with these kinds of characteristics, geneticists turned to basic statistical concepts that describe populations in terms of their means and variation [129]. Unfortunately, the difficulty in finding homozygous populations and controlling the environment, have relegated norm of reaction studies to a few easily manipulated organisms such as the fruit fly *Drosophila melanogaster* [44] and *Arabidopsis thaliana* [193]. The results of these studies have generally yielded only small differences between genotypes and are not consistent over a wide range of environmental perturbations [44, 89]. Furthermore these studies focus on how the variation in the phenotype distributes between the environment, the genotype, and interactions thereof, and thus do not identify specific genes.

Microarray data can be used to identify genes, but some method of analysis is required to link identified genes to environmental perturbations or changes in phenotype. Because both microarray and physiological data can possess many dimensions, a regression method reduces the dimensionality to a significantly smaller set of variables is highly desirable.

One way to investigate these types of multivariate problems, where it is desired to correlate multiple inputs, represented by an "X-Block," (**X**), with multiple outputs, represented by a "Y-Block," (**Y**), is to use a regression method called partial least squares (PLS) [83]. PLS considers the *collective contributions* of the inputs to the outputs, and thus utilizes multidimensional data as opposed to other regression techniques that use data with a single dimension. It is advantageous for large systems because both **X** and **Y** are decomposed into a lower dimensional space where their relationship is explored.

As an example, we explored the application of PLS to microarray data by investigating how a model cell line, Hepal-6[5], alters its gene expression to control glycolytic flux in response to oscillating glutamine concentrations between 0 mM and 4 mM. It has been previously reported that glutamine affects glucose up–take and glycolytic flux [163, 196], and can serve as a carbon source for gluconeogenesis [206] and *de novo* lipogenesis [112]. In these experiments, total RNA was isolated at each time point and the microarray data was used for **X**; at the same time the forward flux through phosphohexose isomerase was measured and used for **Y**. This flux indicator is derived using tritiated glucose (2-^{3}H-glucose), which generates labeled water [269]. Based on the experimental results a PLS model was created, where the transcription data (11 samples x 3,185 genes) was related to the flux measurements (11 samples x 1 flux measurement).

To create the PLS model, both the transcription data, **X**, and the flux data, **Y** were autoscaled[6]. The purpose is to remove distortion that may arise from very large ratios, or variables that show dramatic swings across their timepoints. In this way, the analysis proceeds with each gene on a normalized basis, such that the profiles become more important than the absolute values. It should also be noted that autoscaling the matrices makes them poorly conditioned, and nearly singular, as the determinant of any autoscaled matrix is close to zero.

After autoscaling the data matrix, and selecting an initial set of genes based on a signal-to-noise filter, PLS was run to construct the model. PLS decomposes the original data matrices into a lower dimensional space and then builds a correlation between the reduced matrices. The decomposition of the original matrices is defined

[5] Hepal-6 cells are a murine hepatic carcinoma cell line.

[6] To autoscale the data, each gene measurement has it's mean value (calculated across all samples) subtracted, and the difference is divided by the corresponding standard deviation. Autoscaling converts the data in each row to mean zero and unit variance, and results in the correlation matrix.

by their "outer" relations, given by:

$$\mathbf{X} = \mathbf{T}\ \mathbf{P}^T + \mathbf{E} = \sum \underline{t}_h \underline{p}_h^T + \mathbf{E} \tag{2.12}$$

$$\mathbf{Y} = \mathbf{U}\ \mathbf{Q}^T + \mathbf{F} = \sum \underline{u}_h \underline{q}_h^T + \mathbf{F} \tag{2.13}$$

Because it is possible to let the matrices **T** and **U** (referred to as the "score" matrices) represent the variable matrices **X** and **Y**, a mixed inner relation can be established through:

$$\mathbf{Y} = \mathbf{T}\ \mathbf{B}\ \mathbf{Q}^T + \mathbf{E} \tag{2.14}$$

The resulting model is shown below in Figure 2-9.

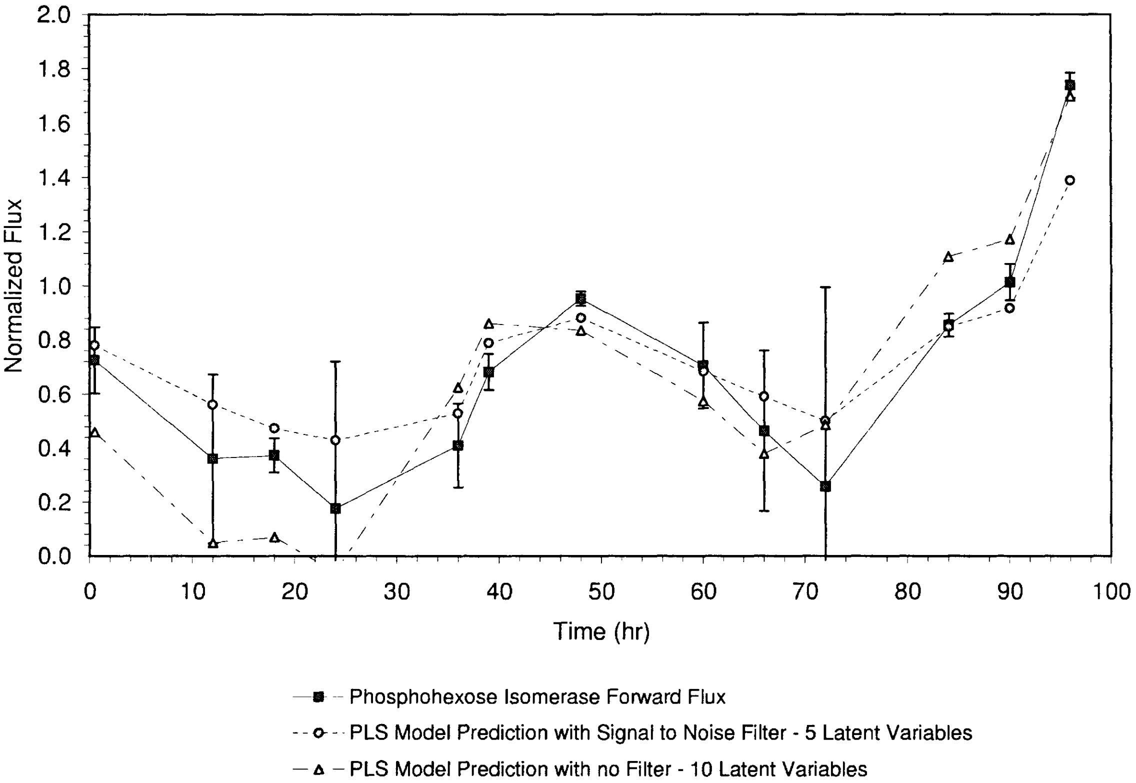


Figure 2-9: Partial Least Squares model prediction of glycolytic flux based upon gene transcription values.

In Figure 2-9, the signal to noise filter does yield an improved fit to the data, even when fewer latent variables are included in the model (all things being equal, the more latent variables included in the model, the more variance captured by the model). Indeed, when the model was recreated using random sets of genes, excluding the 132 used in the filtered model, none of the resulting models predicted the data as well as the filtered model, nor did any of the resulting models capture as much of the variance as the filtered model. This seemingly validated the model with respect to the original **X** and any irrelevant genes contained therein.

In an attempt to further validate the method, the entire X-Block was randomized, and the model reconstructed. In this case, the model based on random data made a prediction that was more highly correlated than the prediction based on the actual data. That is, when random data from a distribution centered at zero with a unit variance, was substituted for each element in the original **X**, the predicted fit was even better than that obtained from the experimental data. This implies that random occurrences, or noise in the data, could provide a model prediction as closely correlated as the actual data. This result sets a new constraint for using the algorithm because if PLS can generate relationships from random, irrelevant data, that have as good predictability as those generated from the actual data, then we can have little confidence in the significance of the relationships found. In terms of the algorithm used, this result implies that the model is grossly overfit, as the inclusion of noise in the data improves the model prediction.

Because of the failure of the actual data to provide a model whose prediction was better than random data, a series of simulations was undertaken (see Appendix A on page 148). These simulations were conducted to determine under what conditions the actual data do a better job of prediction than random data.

We tested whether PLS derived models could arise by chance using completely randomized data matrices that were not related to a pre-determined Y-block. Simulations were run by varying the number of "relevant" genes, "irrelevant" genes, samples, and Y-block variables. In these simulations the Y-block is determined as a linear function of the "relevant" genes, while the "irrelevant" genes are added to the data matrix, but do not contribute to determining the Y-block *a priori*. In this way it is possible to compare PLS models that relate the X-block to the Y-block when the algorithm is given perfect information (that is, only the relevant genes are contained in the X-block), as well as when the X-block contains varying levels of information (when a mixture of relevant and irrelevant genes are contained in the X-block). As described in Appendix A on page 148, it was found that if the number of genes is much greater than the number of samples, as typically occurs in a microarray experiment, the model prediction based on actual, relevant data, could very easily arise by chance from random data. Thus to have relevant models, the number of samples used must make the data matrix closer to full rank, than is typical in most microarray experiments. For full genome arrays this would require thousands of samples, which is prohibitive in most experiments. Given these circumstances, PLS may not be a suitable tool for discovering new relationships between gene transcription data and other biochemical data contained within the Y-block. This does not preclude the useful application of PLS to either discovery, or for modeling biological systems where full rank data may be obtained. It does necessitate careful planning in the prudent use of the technique.

Although there are usually many more genes than samples in microarray experiments, depending upon the experiment there may be effective ways to limit the gene domain. Most of these rely upon either rigorous computational methods (for example, tests for reliable signals or differential expression), or biological hypotheses that can be used to study a sub-set of the genes with respect to the desired outputs. In these cases the researcher is either assuming that most of the relevant genes are in the model, or is testing the model to try and find the relevant subset.

If microarray data were perfect and there was no variance in the measurements–if it accurately determined the state of each gene under a given condition at a specific time, then the models constructed using PLS would at least represent the "best fit" linear approximation of how the transcription data in the X-block relates to the biochemical data in the Y-block. Unfortunately the microarray data can have a significant degree of variation and is susceptible to both random and systematic errors that may result during experimentation [171]. Under these circumstances it is important to critically evaluate model predictions, and correlations.

2.3.3 Computational and Other Analyses

In addition to statistics and multivariate methods, many other methods of analysis have been developed for microarray data. Among these are other techniques for pattern discovery [60, 210, 269], as well as techniques for systems or network discovery [126, 147, 217, 269]. Below is a brief introduction to two other techniques that we explored in model systems.

Cluster Analysis

Cluster analysis is used to find genes that are potentially co-regulated. The idea is that if one gene is induced or repressed in the same manner as another gene, across many samples (either conditions or timepoints), then the two genes may be related. While the biological significance of such a relation would still have to be assessed, cluster analysis provides targets for the discovery of new transcriptional regulatory elements and mechanisms.

Most clustering algorithms use the following process:

- Data acquisition;
- Data normalization;
- Data filtering;
- Data clustering.

Data acquisition was discussed largely in Section 2.2.2. Data normalization is used to correct for artifacts that may influence the data, such as differing dye incorporation rates, and has been reviewed substantially in the literature [199, 245, 270]. The most commonly used normalization methods are mean-centering and autoscaling. Mean centering reduces the mean transcriptional value of any gene across all samples to zero by subtracting the gene's mean transcriptional value from each sample value (across all samples in the data set). This causes the clustering algorithm to focus on the variance in each gene about its mean as opposed to the absolute level of transcription for any given gene. Autoscaling transforms the data into a set that is mean centered and has unit variance. This helps identify established patterns that are independent of the mean and are well conserved across the samples. Data filtering is usually used to remove noise in the data set. Many different types of filters exist and the choice of any given filter depends partially on the experimental design. In this work, the most common filter used was to remove genes that either did not have reliable values across all samples, or genes that were not statistically different in one or more samples of the data set.

Once the data processing is complete, the final step is to cluster the data. There are many different algorithms, such as K-means clustering [48, 81], nearest neighbor [240], self organizing maps [236], and hierarchial [50]. In the Hepal–6 experiments (see above subsection on Partial Least Squares) the concentration of glutamine was oscillated in the cells' medium causing changes in gene transcription and glycolytic flux. To identify genes that were either correlated or anti-correlated with the flux measurement, we used Pearson correlation [143] and the Teiresias [269], which is a pattern discovery algorithm. Teiresias discretizes the expression data by categorizing each transcription value into one of several predefined bins. It then finds patterns in the discrete profiles. Figure 2-10 shows the result of using Teiresias to cluster genes based on their relation to the glycolytic flux determined in the experiment.

The clustering results in the hepatoma experiment bring into focus the fact that increased transcription of some genes was required to allow cells to respond to the new environmental conditions (changing glutamine concentration). Most of the genes found to be activated or anti-correlated with flux are not known to be directly connected to intermediary metabolism, thus highlighting other genes or systems that are perturbed as a result of glutamine changes in the medium. Teiresias sought out genes that had a predefined pattern, but were not necessarily highly correlated with the flux signal, because the algorithm allows variation within the pattern at certain positions[7]

[7] Teiresias can search transcriptional data for any predefined pattern. If the gene expression data is discretized into bins of high (H), medium (M), and low (L) expression, then for an expression profile with five samples, Teiresias can find full patterns (such as "`M L M H M`") or partially full patterns (such as "`(M,H) L (L,M) H M`" or "`M . M H .`" where either value is permissable within the parentheses, and the period allows any value, H, M or L.).

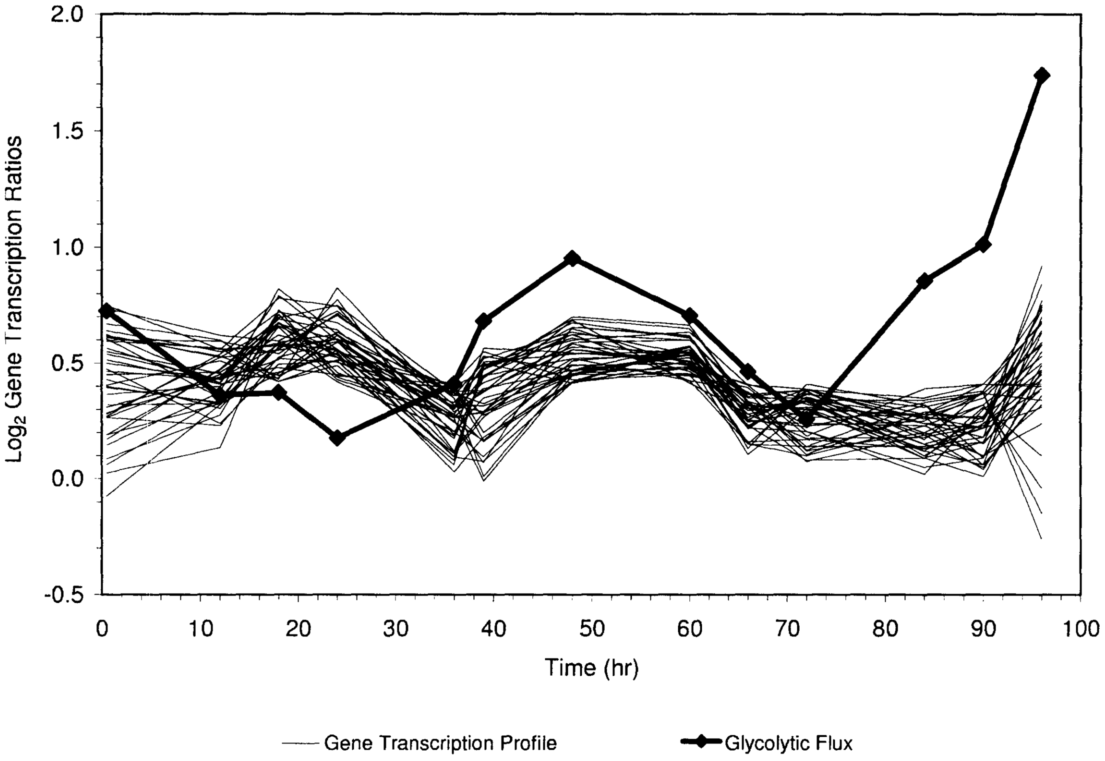


Figure 2-10: Clustering of glycolytic related genes from the Teiresias algorithm.

While the role of the genes detected here in modulating flux has not been resolved, the ability of this model to examine the relationship between genes and fluxes may be an important tool for future studies. Among the genes identified, it was interesting to note that the analysis did not detect significant changes in expression for genes encoding glycolytic enzymes, highlighting the relationships of other genes to glycolytic flux.

Time Lagged Correlations

The various forms of clustering [3, 60, 113] employed to date have produced valuable information, including potential gene relationships and the identity of transcription factor binding motifs. These methods, however, are limited in their ability to infer causality or directional relationships between genes. The results of clustering algorithms yield relations such as "gene A is a good predictor of gene B," which is an equivalent statement to "gene B is a good predictor of gene A." Neither Bayesian networks [77], nor information theory based approaches [227] have made use of the sequential nature of time-series data in current applications. Conversely, when enough time points are available to prevent over fitting the data and find statistically significant correlations, a discovery method to uncover potential causal relationships among genes may be attempted. Directionality can be added to these probabilistic networks by determining the temporal order in which gene expression patterns are affected in a sequence.

A more complete picture of transcriptional regulatory behavior should be possible by probing the transcriptional *dynamics* of carefully designed experiments covering a wide range of conditions. Dynamic experiments that sequentially vary external parameters offer insights into how cellular physiology depends on changing environmental conditions. Time-lagged correlation analysis is one method that can be applied to infer putative causal relationships between system perturbations and system responses. Linear Pearson correlations have been used to identify genes that are co-expressed or anti-expressed for clustering purposes [143]. Time-lagged correlations extend this technique by determining the best correlations among profiles shifted in time. For a transcription profile represented by a series of n measurements taken at equally spaced time points, the correlation between genes i and j with a time lag, τ, is $\mathbf{R}(\tau)$ = $(\mathrm{r}_{ij}(\tau))$, defined by

$$S_{ij}(\tau) = \langle (x_i(t) - \bar{x}_i)(x_j(t+\tau) - \bar{x}_j)\rangle \tag{2.15}$$

$$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}(\tau)S_{jj}(\tau)}} \tag{2.16}$$

where $\mathrm{x}_i(t)$ denotes the expression of gene i at time t, $\bar{x}_i$ is the expression value of gene i averaged across all time points, and the angled brackets represent the inner product between the time-shifted profiles. The matrix of lagged correlations $\mathbf{R}(\tau)$ can be used to rank the correlation and anticorrelation between genes through conversion to a Euclidean distance metric, d_{ij}:

$$d_{ij} = (c_{ij} - 2c_{ij} + c_{jj})^{1/2} = \sqrt{2}(1.0 - c_{ij})^{1/2} \tag{2.17}$$

$$c_{ij} = max|r_{ij}(\tau)| \tag{2.18}$$

where, c_{ij} is the maximum absolute value of the correlation between two genes with a time lag τ. If the value of τ that gives the maximum correlation is zero, then the two genes are best correlated with no time lag. The matrix $\mathbf{D} = (d_{ij})$ describes the correlation between two genes, i and j, in terms of "distance" by making genes that are least correlated (for any τ) the "farthest" apart [6]. Thus by transforming the correlation matrix, **R**, into a distance matrix, **D**, we are able to include highly anti–correlated genes, in addition to correlated genes, in the network. By finding genes that are closely related and then examining the corresponding value of τ, an underlying network of potential cause and effect relationships can be elucidated. Some caution is needed to ensure genes with high correlation have been chosen using enough data points to give statistical significance, otherwise all of the τ values used will merely overfit the data. Such errors may be obvious if values for τ are unreasonably long from a biological standpoint.

Time lagged correlations were used to analyze metabolism in a model system, where the photosynthetic bacterium, *Synechocystis* sp., was exposed to different light conditions [217]. Dynamic light perturbations were induced in this system to drive the transcriptional response of the bacteria, which was measured using DNA microarrays. The gene transcription responses were then placed into a network based upon their time lagged correlations to either the input light signal or another gene cluster, providing a set of putative causal relationships that could be tested in subsequent experiments. After collecting transcriptional data from over 47 time points, the network shown in Figure 2-11 was constructed.

In Figure 2-11 solid lines represent gene groups with correlation at the indicated time lag, while broken lines represent gene groups that are anticorrelated at the indicated time lag. The resulting network is composed of 50 groups containing 259 genes. The genes within the network include known genes that have demonstrated light–induced regulation, as well as unannotated genes whose functions have yet to be assessed. This suggests that dynamic studies of transcriptional behavior with

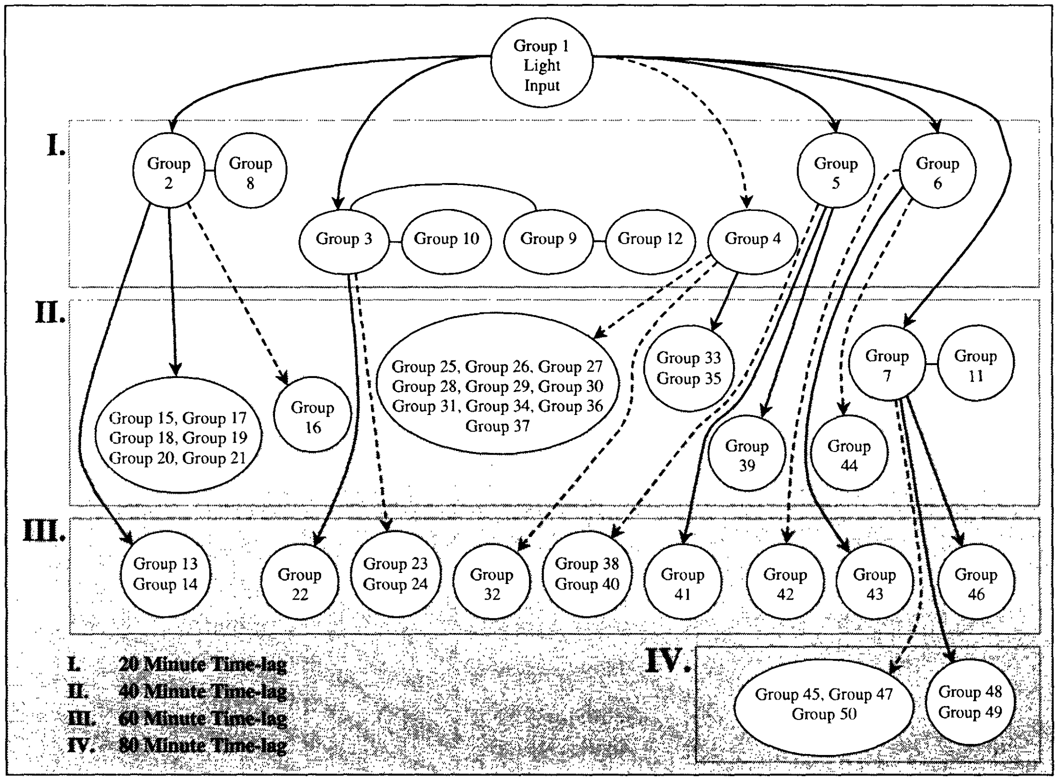


Figure 2-11: Gene interaction network derived from time lagged correlation analysis using gene transcription data.

significant numbers of time–points can play a key role in understanding cellular regulation. As other measurements such as protein and metabolite data become available, time–lagged correlation studies should allow for the creation of hypothetical networks similar to that in Figure 2-11, but with greater degrees of mechanistic information. Such approaches will hold new insights into the regulation of biological systems.

Chapter 3

Methods for Gene Characterization

Following on Chapter 2, this chapter discusses various methods for gene characterization. Because genome sequencing projects have catalyzed the development of high-through-put technologies that help identify an increasing number of genetic targets [40, 78, 161] finding equally efficient methods to characterize these genes is important.

Historically, studying loss of function phenotypes in cell culture or whole animals has been a critical aspect to determining a gene's *in vivo* regulation and biochemistry. RNA interference, RNAi, is a recently discovered phenomenon that can be used to specifically silence genes in a complementary, high-through-put manner [9, 154] enabling the use of gene silencing to create loss of function phenotypes in greater eukaryotes. This chapter summarizes our development of RNAi based gene silencing methods in tissue culture.

3.1 Model Systems

To study diseases or quantitative traits in higher eukaryotes, an adequate model system is required. From a research perspective, human studies in diabetes have helped identify important risk factors (See Section 1.1), however they are limited to clinical investigations that require appropriate training and usually cannot by themselves delineate the fine biochemical processes that define disease pathogenesis.

For mechanistic studies a greater degree of control is required and several different models are possible.

Animal Models

Many different animal models exist for studying disease pathogenesis. Such animal models include rats, chimpanzees, zebra fish, frogs, pigs, and dogs. One of the most commonly used for studying diabetes and obesity is the mouse. The advantages of this experimental model in gene discovery are numerous:

- Mice are mammals and therefore in the same infraclass[1] (*Eutheria*) as humans;
- Mice are readily available and easy to maintain in colonies because of their small size and moderate living requirements;
- Mice reproduce relatively quickly and have an adequate life span for many types of experiments;
- Mouse genes show a high degree of similarity to human genes [42, 178];
- There is a rich research history in the mouse that includes copious mutant strains and well developed experimental techniques for transgenics.

In addition to these reasons, mice are a particularly good animal model for metabolic diseases because they have an endocrine system that is very similar to the human endocrine system. Thus the relevant tissues and organs that are affected in diabetic patients, including the pancreas, white adipose tissue, muscle, and liver, are found in mice and behave in a similar fashion. Glucose homeostasis, to our knowledge, is largely regulated in a similar manner in all known mammals and therefore can be studied in controlled and statistically rigorous detail in mice.

While mice have been extremely valuable in gene discovery and biochemical analysis, they are not ideal for some types of experiments. For example, knocking genes

[1] In taxonomic hierarchy, the "infraclass" lies below Kingdom, Phylum, Subphylum, Class, and Sublass [183].

out in mice is very difficult, sampling may be limited depending upon the amount and type of sample required, and selectively perturbing specific tissues organs or tissues is often impossible. For experiments where these factors are limiting other models may be used.

Primary Cells

In lieu of a relevant animal model, the next best model system is often primary cells isolated directly from the animal. These cells are as readily available as the animal itself, can be isolated in relatively large quantities from different organs or tissues, are readily cultured and can be sampled with ease, and importantly they maintain their differentiated tissue phenotype better than cell lines. Therefore it's possible to isolate these cells and study how they react to imposed and well controlled perturbations. The detail with which this is possible far exceeds that of the mouse model, enabling the use of intricate imaging techniques, isotope labeling experiments, binding experiments, and direct isolated use of agonists and antagonists.

The primary disadvantages to using primary cells is that they're no longer connected to the other organs (and therefore there's an important loss of information that *in vivo* studies provide), and the cells cannot be expanded or used for prolonged experiments. Thus very large or long experiments, as one might design for use with time lagged correlations, may be difficult to perform using primary cells.

Cell Lines

While cell lines are immortalized and therefore may display aberrant behavior that does not reflect phenotypes observed *in vivo*, they posses other advantages that periodically make them a preferential system. Among the advantages of using cell lines as a model system are

- Cell lines are easily stored and do not need to be isolated from an animal;
- Cell lines are usually viable over a longer period of time than primary cells;

- Cell lines can be expanded to permit very large scale experiments.

Cell lines have been isolated from many different organs and animal models and sometimes have specific properties that make them good model systems. For example, some cell lines over express hormones [228], glucose [88], or proliferate in response to cytokine or hormonal treatment [145]. Such cells are often used for directed biochemical studies of cellular signaling, metabolism, gene transcription, protein interactions, or other specific aspects of cellular physiology. There are even a few cell lines that are relatively well accepted models for certain tissue types [153, 208].

Despite their "off-the-shelf" ease of use, cell lines usually lose their differentiated tissue characteristics during the immortalization process used to establish their lineage. For example, while primary hepatocytes can readily produce glucose in culture, hepatic cell lines, such as Hepal–6 and HepG2, are predominantly glycolytic and cannot be induced to produce glucose. For this reason *in vivo* data from an animal model, or data from primary cells, is generally preferred in studies of cellular physiology.

3.2 Gene Knock–out and Over–expression

Once a gene has been implicated in disease pathogenesis, either through QTL analysis, microarray studies, or some other method, the molecular mechanism through which it contributes to the observed phenotype is sought. This work entails a host of biochemical studies which often include measuring transcription of the gene in various tissues and under different conditions, measuring levels of the corresponding protein, determining the gene product's function, mutation analysis, and finally understanding how the gene participates in a molecular pathway that explains observed phenotypes.

Many of the *in vitro* experiments (such as determining enzymatic activity and kinetics, *in situ* hybridization, affinity, etc.) required for gene characterization are straightforward, and may be conducted in bacteria (for gene cloning and protein production), cell culture, primary cells, or tissues taken from an animal. One of the difficulties, however, is assessing the gene's cellular function in greater eukaryotes. This is critical for understanding the physiological role of the gene and it's product.

Both over–expression studies [23, 57, 100, 152, 181] and mutation studies [58, 139, 166, 235, 237, 251] have been successfully employed to elucidate a gene's effect on physiology, however, the through–put of these experiments is much lower than those used for gene identification.

In bacteria and in some lower eukaryotes such as yeast, the techniques for assessing gene function *in vivo* is well developed and almost trivial to perform; however, in higher eukaryotic cells, tissues, or animals, determining gene function *in vivo* can be an enormous task. The reason for the difficulty in conducting *in vivo* studies in higher eukaryotes is the intricate experiments required to disrupt gene function (gene "knock–out") and over–express genes. These problems include specifically targeting genes for disruption to obtain a viable transgenic animal, or targeting specific organs in transient over–expression studies.

Although systems have been developed to create a variety of specific mutations (chromosome rearrangements, deletions, insertions, point mutations, tissues specific mutations, inducible mutations, etc. [23, 29, 168]), the process is time consuming [67] and the experimental manipulations are not trivial. The overall process for knocking–out a gene follows these steps:

- The targeted gene is isolated (preferably from the same genetic background as the intended blastocyst) from a genomic library to produce a clone.
- A construct is engineered to disrupt the gene by homologous recombination that allows for double selection.
- Embryonic Stem (ES) cells are electroporated with the engineered DNA construct.
- Transgenic ES cells are selected for through resistance to a toxic chemical (such as gancyclovir). Individual clones are then selected.
- Clones are screened by Southern analysis and polymerase chain reaction (PCR) for the appropriate homologous recombination event.

- Positive recombinant ES clones are injected into blastocysts and implanted into the uterus of a foster mother.
- Pups are born and chimeric animals can be selected by their coat color. The recombinant event is then validated again by Southern analysis (and PCR).
- Appropriate chimeric animals are then crossed to generate heterozygous and homozygous mice.

Depending upon the success of each of these steps, generating new, validated knock–out animals with the correct mutation takes at least one and a half years, and more typically two to three years [67, 168]. A similar process is used to introduce a new gene into an animal for expression studies [67]. Both of these processes are time consuming, cumbersome because of the required level of screening, and not trivial to perform.

In addition to transgenic techniques, it is also possible to conduct transient gene expression experiments by transfecting cells with an expression vector. While this is relatively easy with primary cells or cell lines[2], transient over–expression in animals can be more difficult and requires some degree of optimization. In most cases either a viral vector or plasmid DNA are injected directly into the animal, and the phenotype is observed post–injection [37, 41, 135, 255]. Potential issues can arise in targeting specific tissues depending upon the mode of delivery and vector used, as well as unintended vector effects. Vector effects can be controlled for in experiments, but may still generate artifacts if they interfere with the animal's physiology. Because of these reasons the ability to test a gene's function *in vivo* has not kept pace with the ability to identify potentially important disease related genes [75].

Until recently generating knock–out cell lines was also difficult, requiring a high degree of screening, and many cell lines were derived from the knock–out animal itself. Further, cell lines suffer from being a less relevant animal model. Fortunately

[2] Transfection experiments in cell culture can be performed simply by exposing the cells to a transfection mixture containing the DNA vector of interest and a chemical carrier. Some optimization is required, however, the results may be adequate for screening gene effects or to study particular molecular or cellular phenomena such as protein binding, protein phosphorylation, or metabolism.

this situation changed with the discovery of RNA interference (RNAi), which allows the creation of "functional knock–outs" through post–transcriptional gene silencing. Thus for the first time, it is now possible to efficiently decrease gene expression in primary cells and cell lines, and to a lessor degree animals.

3.3 RNA Interference

RNA interference (RNAi) refers to a highly conserved biological pathway that can be exploited for post–transcriptional gene silencing in many different cell types [98, 160]. First discovered in *Caenorhabditis elegans* [73],its efficacy has been subsequently demonstrated in *Drosophila melanogaster* [188], plants [249], a variety of cell lines [63, 138], and even whole animals [106, 158]. Interestingly, among these distantly related organisms, the different mechanistic aspects of the RNAi pathway have diverged, making the use of RNAi for gene silencing species dependent. The core pathway that appears to be well conserved is illustrated schematically in Figure 3-1.

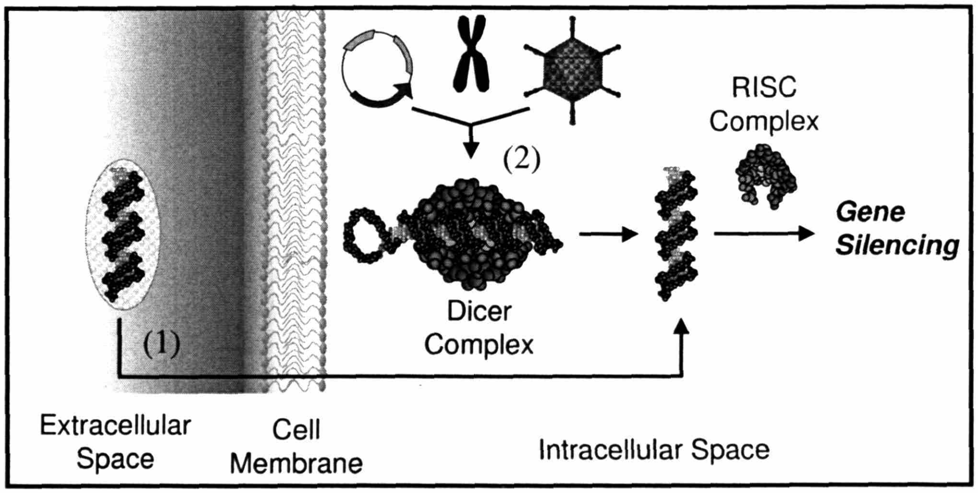


Figure 3-1: The core pathways of RNA interference (RNAi). RNA may be delivered by transfection to the RISC complex to activate the pathway (1); or it can be produced endogenously from a transfected plasmid, expressed as a micro–RNA from the genome, or expressed from an engineered virus (2).

In the core pathway, double-stranded RNA (dsRNA), enters the cell cytoplasm, either through the transcription of endogenous genes encoding microRNAs (miRNA) [90], viral delivery [209, 216], or complex formation with a lipid carrier and endocytosis [55, 63], where it encounters the Dicer enzyme complex. This complex has two RNase III motifs, an RNA helicase domain, and a dsRNA binding domain [176, 198]. Dicer cuts the dsRNA into 19 - 21 base pair (bp) fragments, typically with a 1 or 2 nucleotide overhang [170]. In most cell types the RNA Interference Silencing Complex [97], or RISC complex, binds these small interfering RNAs (siRNA), unwinds them, and uses their hybridization ability to target and degrade longer, complementary transcripts. In plants there is an amplification mechanism [108], and in some cell types there is also a mechanism that inhibits translation [275]. Notably, if longer dsRNA oligonucleotides (>30 bp) are used to transfect mammalian cells, they will elicit a response entailing interferon synthesis and protein kinase (PKR) activation that stalls translation by phosphorylating the translation initiation factor eIF2a [49]. This response is non-specific, inhibiting all of translation. Because longer dsRNA can stop translation in mammalian cells, Dicer's role appears to be in processing miRNAs. miRNAs are endogenously expressed RNAs that form hairpin loop and stem structures that are cleaved to their active form. Although translation inhibition can occur in the presence of long dsRNA, researchers have shown that the silencing pathway can be activated downstream of Dicer, by supplying short interfering RNAs (siRNAs) to RISC [242].

Thus far most research using RNAi has focused on reducing the expression levels of single genes, and observing the resulting phenotype at some subsequent point in time. These experiments have primarily investigated the effects of specific genes on predefined dependent variables [25, 157], modulation of the host system response to infection [165, 118], and preliminary findings in animals [155, 158, 209].

The RNAi pathway provides a new tool for exploring gene function. Use of the RNAi pathway enables a simple and efficient way to modulate gene expression by creating effective loss-of-function phenotypes. This assumes that the gene silencing pathway itself functions autonomously and is truly gene specific, a hypothesis yet to be fully proven [117, 214]. Furthermore, the technique relies on the selection of efficacious targeting sequences, which has been studied extensively [62, 96, 204, 247].

Because the silencing mechanism of RNAi occurs via a protein-catalyzed pathway, the dynamics of the pathway must be considered for optimal application of this method. In each RNAi application the efficiency of silencing depends on a number of factors, some of which are under the control of the experimenter. To conduct an RNAi–based silencing experiment, gene specific sequences must be selected and delivered to the cells at a predetermined time. It is therefore important to understand how an experimental protocol affects gene silencing experiments. To understand this relationship, one must account for a number of key variables including the efficiency of transfection, the dynamics of gene activation and repression, the level of mRNA transcription, the stability of the mRNA transcript, the rate of protein translation, and the protein's stability, among other factors.

In this work we developed RNAi–based gene silencing protocols in hepatoma cells, and then extended those results to primary hepatocytes. Gene silencing provides an efficient way for us to screen genes identified in other studies in a relevant cell culture model, which serves as the first level of characterization preceding future physiological studies in animals. While developing these methods, we also investigated some of the primary experimental parameters that control RNAi–based gene silencing and formulated a model to describe the silencing pathway. We studied the effect of RNA concentration, complex exposure time, and the relative timing of transfection on the dynamics of gene silencing in cells transiently expressing green fluorescent protein (GFP). RT–PCR was used to measure the GFP transcript levels, while the amount of protein was determined by measuring the fluorescence emitted from washed cells. We found that the level of gene silencing can be controlled between 0% and 100% by altering the experimental parameters used. In addition, we also developed a simple model that is useful to understand how these experimental parameters affect the degree of gene silencing and help plan more complex experiments where multiple genes are involved.

3.3.1 RNAi Experimental Development

An important determinant of gene silencing experiments is the conditions of the siRNA transfection. In both cell culture and animal experiments [160] the amount of siRNA used and method of delivery are critical parameters. Depending upon the duration and desired level of gene silencing required a particular experiment, the siRNA dose will be influenced by the siRNA sequence, the level of gene transcription, the stability of the mRNA transcript, and the stability of the siRNA. Many different delivery methods have been demonstrated in the literature [222], however, in this work we focus on non–viral delivery of synthesized siRNAs.

The efficiency of RNA transfection into Hepal–6 cells was measured using fluorescein–labeled, non-silencing RNA. Figure 3-2 on the next page shows a titration of the RNA–dependent fluorescence measured one day after transfection for different amounts of labeled siRNA in the transfection mixture. As shown, a monotonically increasing response is obtained. Besides dose, the efficiency of transfection also depends on the length of time that the cells are exposed to the transfection mixture. Figure 3-3 shows the dynamic change in RNA–dependent fluorescence on subsequent days following exposure of Hepal–6 cells to 1 μg of fluorescein–labeled non–silencing RNA for different lengths of time. The results indicate that different amounts of siRNA may be delivered by varying the dose of RNA in the complex mixture and the time of exposure. The latter effect, however, saturates above two hours when 1 μg of RNA is used.

Next we examined the silencing of a GFP expressing plasmid by co–transfection with GFP siRNA. Figure 3-4 on page 72 shows that fluorescence decreases with increasing amounts of siRNA. In this system, control cells were treated with plasmid and (unlabeled) non–silencing siRNA. In contrast, 1 μg of siGFP was sufficient for almost 100% silencing and the effect is stable for at least four days.

To study the effect of transfection time on the production of GFP in Hepal–6 cells, we transfected the cells with a DNA plasmid encoding GFP. Positive control cells were co–transfected with non–silencing siRNA, mock-transfected negative control cells were

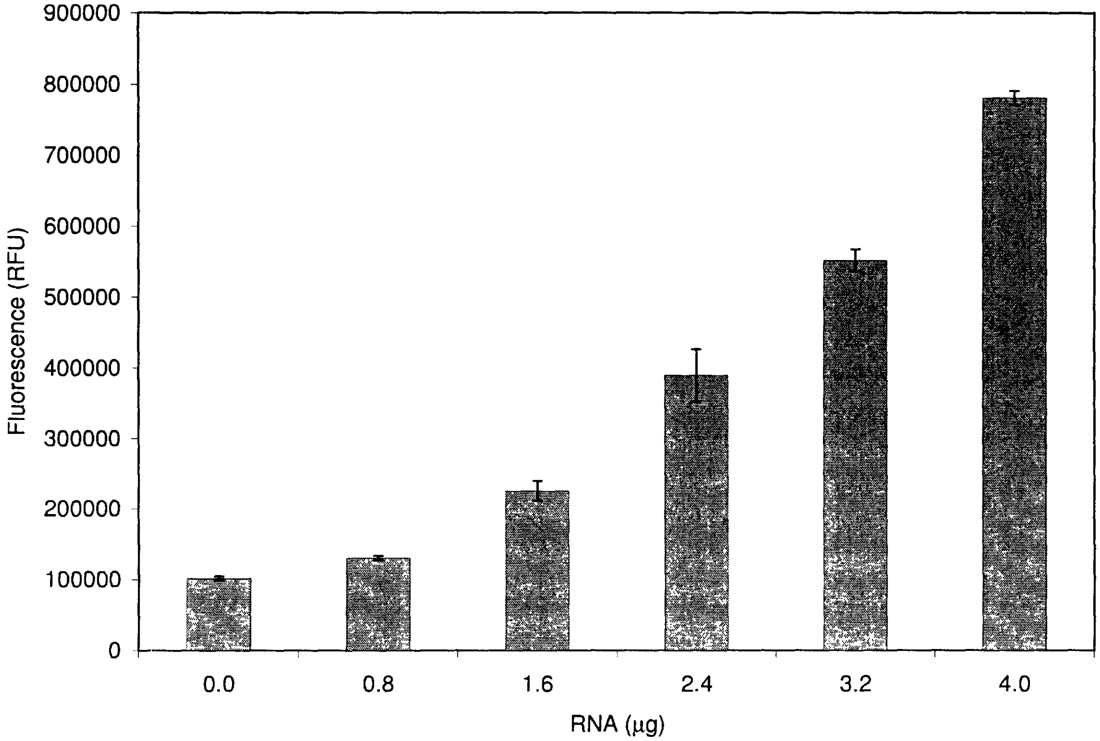


Figure 3-2: A titration on the amount of RNA used during transfection.

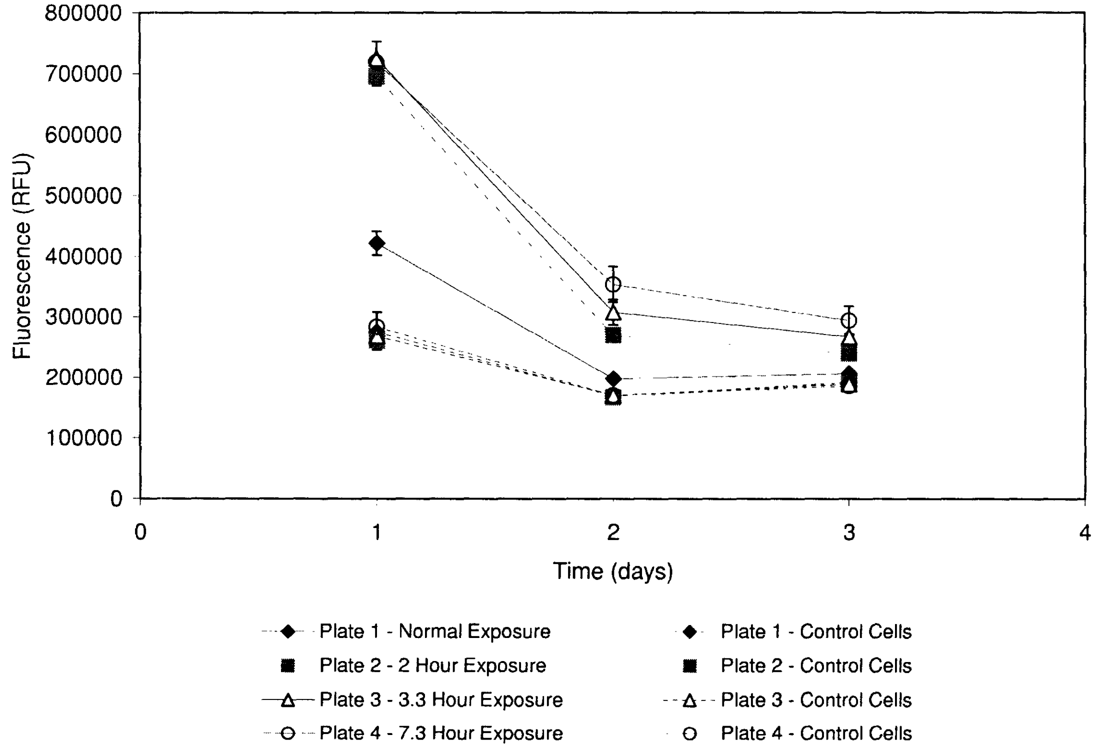


Figure 3-3: Effect of RNA transfection complex exposure time on RNA delivery to, and retention within, cells.

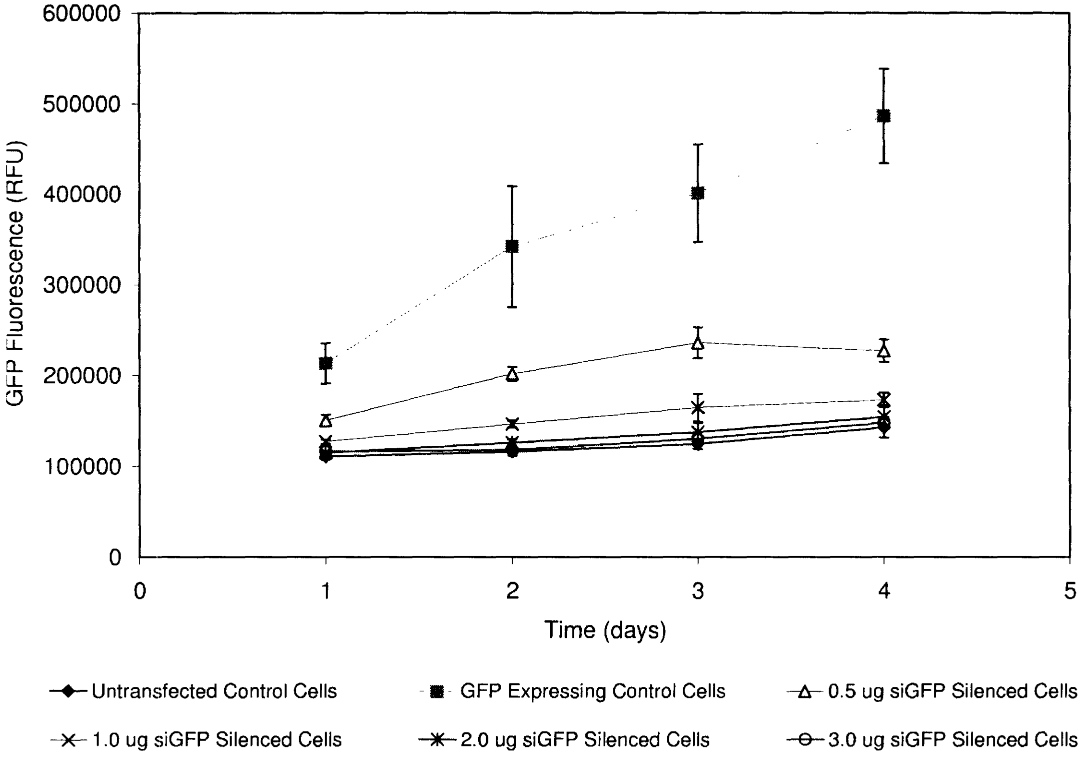


Figure 3-4: Silencing of Green Fluorescent Protein in Hepa1–6 cells as a function of siRNA concentration. Fluorescence measurements started one day after transfection of the siRNA.

exposed to the transfection conditions but received no plasmid or siRNA, and GFP siRNA was used to treat experimental cells at three different times. One treatment of the experimental cells used co–transfection, in the second treatment the siRNA was added 24 hours after the DNA transfection, and in the third treatment cells were transfected with the siRNA 48 hours after the DNA transfection. The results of this experiment are shown in Figure 3-5. It can be seen that only the co–transfection condition was effective in silencing the expression of GFP protein as determined by the fluorescence level. Transfecting 24, or 48 hours later had no significant effect.

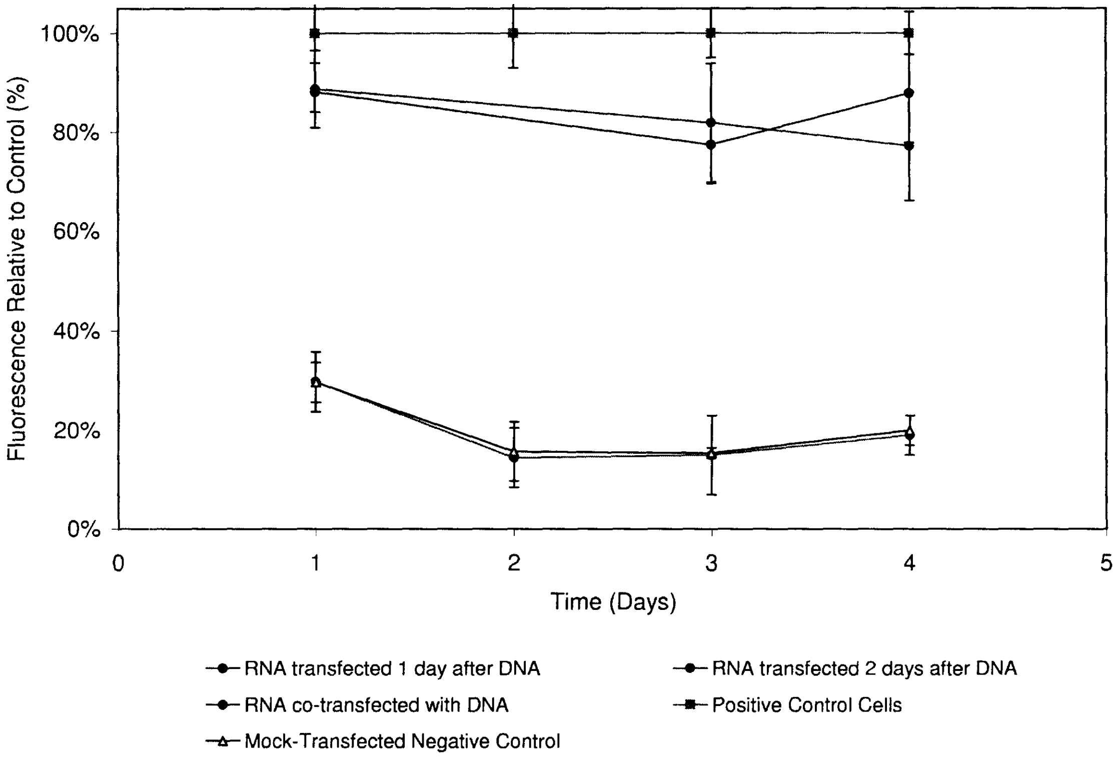


Figure 3-5: Effect of transfection time on GFP Silencing.

To understand the dynamics of RNAi–based gene silencing in this system, we repeated the co–transfection experiments, this time measuring both the protein levels, as indicated by the amount of fluorescence observed, and the mRNA levels, using RT–PCR. Figure 3-6 on the following page shows the obtained data. The mock–transfected cells show no increase in mRNA or fluorescence throughout the experiment, demonstrating the background levels of the assays. The positive control (cells transfected with plasmid DNA and non–silencing siRNA) shows a very rapid increase in the level of mRNA over the first 24 hours, followed by a subsequent increase in the protein levels 24 hours later. Interestingly, there's a large decrease in the mRNA levels on day two, indicative of plasmid loss from the cells. In the cells co–transfected with GFP siRNA, there was some accumulation of mRNA, but much lower than the level of the positive control, and the protein level never exceeded that of the negative control.

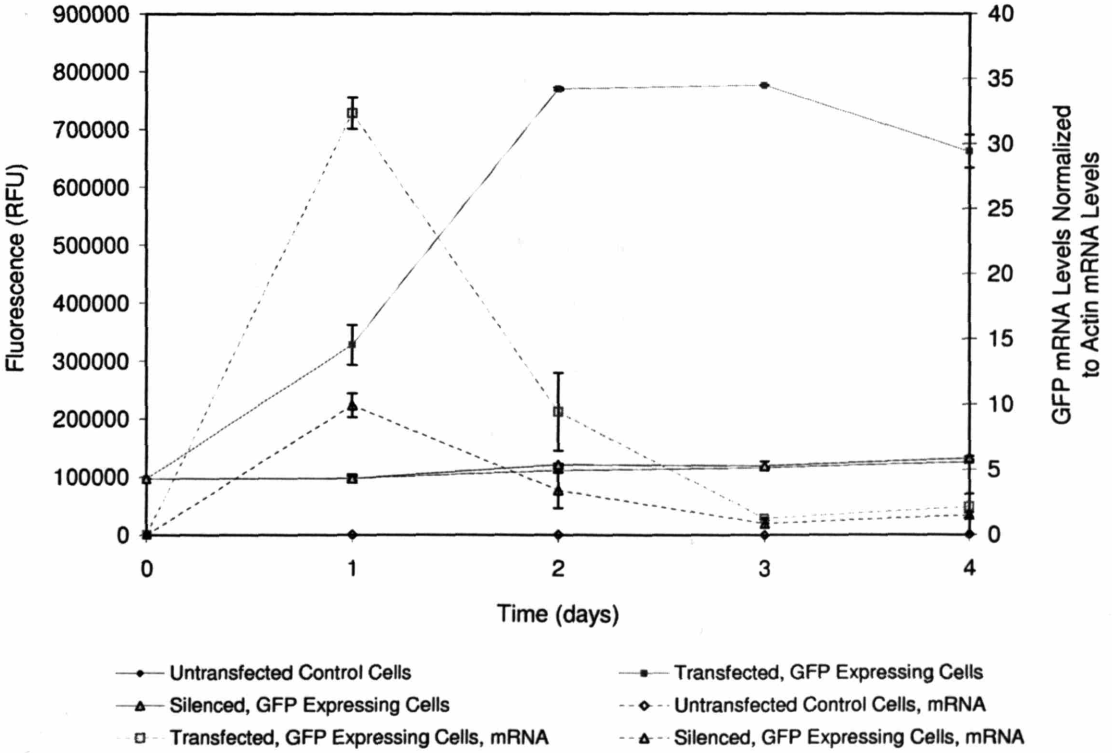


Figure 3-6: Dynamics of GFP mRNA and protein levels during gene silencing with siRNA.

Figure 3-6 illustrates the importance of monitoring both the RNA, and the protein levels in this system. For example, although there were only small changes in the RNA levels on days 2, 3, or 4 there are actually large differences in the protein levels. In Figure 3-6 the shallow rise of the GFP mRNA level in the GFP siRNA co–transfected treatment demonstrates two things. First, although the observed fluorescence never increased over the levels of the negative control, the mRNA level did increase. Thus to suppress the level of transcripts in the co–transfected cells to that of the negative control cells more siRNA must be delivered. Second, because there was no observed increase in the fluorescence level in the co–transfected treatment, despite a small increase in mRNA levels, the fluorescence assay may not be sensitive enough to adequately quantify the amount of GFP protein at low concentrations.

Because of the steep decline in the GFP mRNA levels of the positive control, it is clear that any transfection time other than co–transfection will not result in high levels of silencing. For some experiments, only partial silencing may be desirable, and controlling the time of RNA transfection is one way in which partial silencing may be obtained.

The principal finding illustrated by Figure 3-6 is that the dynamics of gene expression is a key factor to effectively manipulate the RNAi silencing pathway. By knowing when a gene is activated and how quickly it is repressed, the rate of transcription and translation, the mRNA and protein stability, and the rate of transcript degradation through the RNAi pathway, one can design robust experimental protocols to modulate the intracellular concentration of RNA and protein to desired levels.

RNAi Gene Silencing Model

The experimental system in which the cells are co–transfected with plasmid DNA expressing a reporter gene and the corresponding specific siRNAs has been investigated by a number of researchers [63, 187], but the system dynamics have not been studied in detail. The results of Figures 3-2 — 3-6 were used to construct a model capturing the current understanding of the RNAi mechanism in fully confluent, adherent mammalian cells. The model describes the dynamics of mRNA transcript (T), protein (P), plasmid (template DNA, D), and transfected siRNA (R) concentrations during RNAi–based gene silencing from a transfected plasmid by the following set of differential equations:

$$\frac{dD}{dt} = -k_{dD}D \tag{3.1}$$

$$\frac{dR}{dt} = -k_{dR}R \tag{3.2}$$

$$\frac{dT}{dt} = k_{Ts}D - k_{Td}T - k_{TR}RT \tag{3.3}$$

$$\frac{dP}{dt} = k_{Ps}T - k_{pd}P \tag{3.4}$$

Concentrations are expressed per unit cell mass. The loss of plasmid is represented as a first–order process, but this could also be thought of simply as the loss of transcribable template, due to gene repression. A similar first–order process is assumed for the loss of siRNA, R. Transcripts are generated at a rate proportional to the concentration of plasmid present and similarly destroyed in a first order process. The silencing mechanism reflected in the last term of equation 3.3 is assumed second order with respect to the concentration of the silencing RNA and transcript. Finally, the level of protein production is proportional to the concentration of transcript, and degrades according to a first order process.

The rate constants were determined by minimizing the squared difference between the numerically integrated model predictions and the actual data for both the mRNA and protein levels of the co–transfected and positive control treatments. The determined rate constants agree well with those measured previously in the literature [211] and are given in Table 3.1.

Figure 3-8(a) compares the model results with the data for the mRNA concentrations under the different conditions. The model accurately describes the kinetics of the mRNA pool for the positive control, silenced, and negative control treatments. It predicts a maximum in the pool size approximately 0.5 days following transfection, followed by a steady decrease in mRNA as the plasmid is lost from the culture. Figure 3-8(b) shows the model predictions for the mRNA levels when siRNA transfection occurs 24, or 48, hours after DNA transfection, while Figure 3-8(c) shows the model predictions for the protein levels under all conditions. Replotting the predictions from Figure 3-8(c) as a percent fluorescence of the positive control allows us to compare it with the original data obtained. Figure 3-8(d) shows the resulting comparisons. The trends agree qualitatively well with the actual data, illustrating the utility of the model.

Although models like the one designed here are system specific, they are valuable

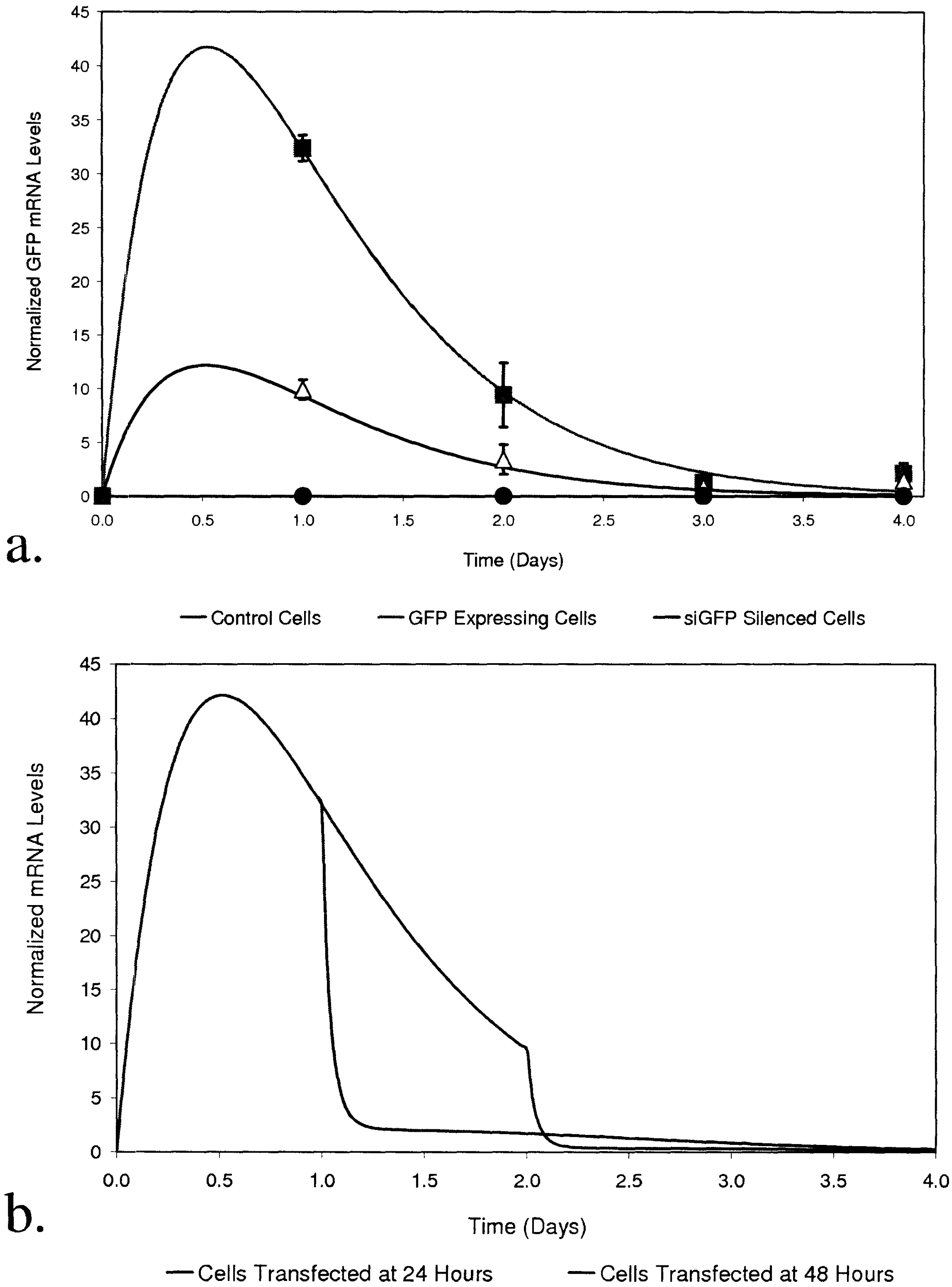


Figure 3-7: Model predictions and data for GFP gene silencing in Hepal–6 cells. (a) mRNA Model Predictions and RT-PCR Data. The model prediction of the mRNA levels agress closely with the actual data. (b) Model Predictions of the mRNA Levels when siRNA Transfection Occurs 24– or 48–hours After DNA Transfection.

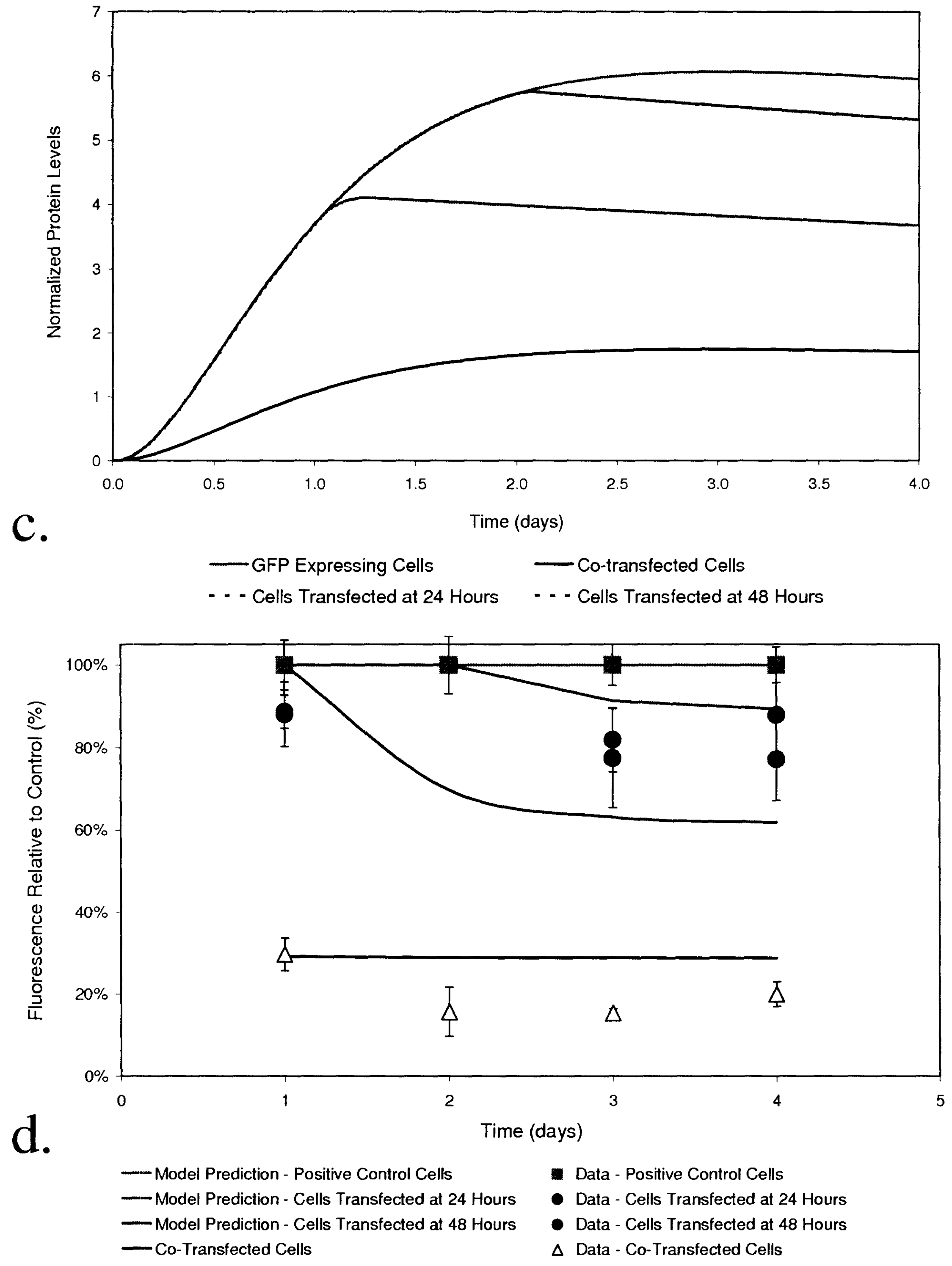


Figure 3-8: Model predictions and data for GFP gene silencing in Hepal–6 cells. (c) Model Predictions of the Protein Levels when siRNA is Transfected at Different Times Following DNA Transfection. (d) Comparison of the Model Predictions and Actual Data. Data from GFP silencing experiments is plotted as a percentage of the positive control. The pink squares represent the GFP expressing positive control, while the blue and purple circles represent cells transfected with GFP siRNA 24– and 48–hours after DNA transfection, respectively. The yellow triangles represent cells co-transfected with DNA and RNA. As shown transfecting with RNA after DNA results in less potent gene silencing (as determined by protein levels) and is in agreement

Rate Constant	Value	Description
k_{dD} (day^{-1})	1.68	Gene repression, or loss of plasmid rate constant. Directly contributes to the decreasing concentration of transcript over time.
k_{dR} (day^{-1})	1.93	siRNA degradation rate constant. Directly contributes to the reduction in mRNA via the RNAi silencing pathway.
k_{ts} (transcripts/ template/ day)	301.69	Transcript synthesis rate constant.
k_{td} (day^{-1})	2.11	Transcript degradation rate constant.
k_{tR} (day^{-1} siRNA^{-1})	16.67	Transcript degradation rate constant from the RNAi silencing pathway. Directly contributes to the reduction in mRNA via the silencing pathway.
k_{PS} (protein/ transcripts/ day)	0.11	Protein synthesis rate constant.
k_{Pd} (day^{-1})	0.04	Protein degradation rate constant.
R_0 (siRNA/ cell)	1.0	Initial concentration of siRNA in cell.

Table 3.1: Rate Constants for RNAi–Based Gene Silencing Model.

tools in exploring the outcome of different gene silencing protocols and the interplay between key experimental parameters. To illustrate, the model was applied to a particularly interesting case arising from multi-gene interactions. Consider the relatively simple two gene system in which one gene, A, induces a second gene, B, and the second gene in turn represses A. This situation can arise in a number of biological contexts [102, 226] and can result in oscillatory behavior depending upon the rate constant values. The model equations are:

$$\frac{dR_A}{dt} = -k_{RA}R_A \tag{3.5}$$

$$\frac{dT_A}{dt} = k_{TsA} - k_{Td}T_A - k_{TRA}R_AT_A - k_{TAP}P_B \tag{3.6}$$

$$\frac{dP_A}{dt} = k_{PAs}T_A - k_{PAd}P_A \tag{3.7}$$

$$\frac{dR_B}{dt} = -k_{RB}R_B \tag{3.8}$$

$$\frac{dT_B}{dt} = k_{TsB}P_A - k_{TBd}T_B - k_{TRB}R_BT_B \tag{3.9}$$

$$\frac{dP_B}{dt} = k_{PBs}T_B - k_{PBd}P_B \tag{3.10}$$

Concentrations are expressed per unit cell mass, and the units and processes are similar to those described in the previous model. Gene A is constitutively transcribed, resulting in a zero–order generation process for transcript A, but repressed (or equivalently, degraded) by a first order process controlled by protein B. Gene B is induced by the protein product of gene A, resulting in a first–order accumulation.

The behavior of the native system using predefined rate constants is shown in Figure 3-10(a). Here we see the rising rate of transcript A leads to the formation of protein A, followed by the formation of protein B, which eventually rises to a level that suppresses gene A transcription. This type of behavior could easily be identified in microarray experiments where, depending upon the sampling frequency and method of autoscaling, these profiles would have a high probability of clustering together or being linked through time lagged correlation analysis.

In this system the timing of RNA transfection is crucial and can have a variety of results from no effect to prolonged silencing. Likewise, depending upon how much knowledge is known *a priori*, the time at which RNA transfection is initiated can be exploited to control or study the system. For example, if genes A and B were identified in a cluster and we wanted to study their effects, then one approach would be to silence each gene individually. In contrast to the previous system, if we transfected at the perceived time of gene induction in an attempt to completely silence gene B for an extended period of time, virtually no change in the mRNA and protein levels will be observed as shown in Figures 3-10(b) and 3-10(c), respectively. Conversely, if gene B siRNA is transfected 0.4 days after gene A induction, then accentuated silencing can be attained and much larger differences in the mRNA and protein levels between the silenced and control cultures observed (Figures 3-10(b) and 3-10(c)). If repeated doses are used, then gene B can be effectively silenced, eliminating its repressive effect on gene A to a greater extent, and allowing a more in–depth study of gene A at low levels of gene B expression (Figure 3-10(d)).

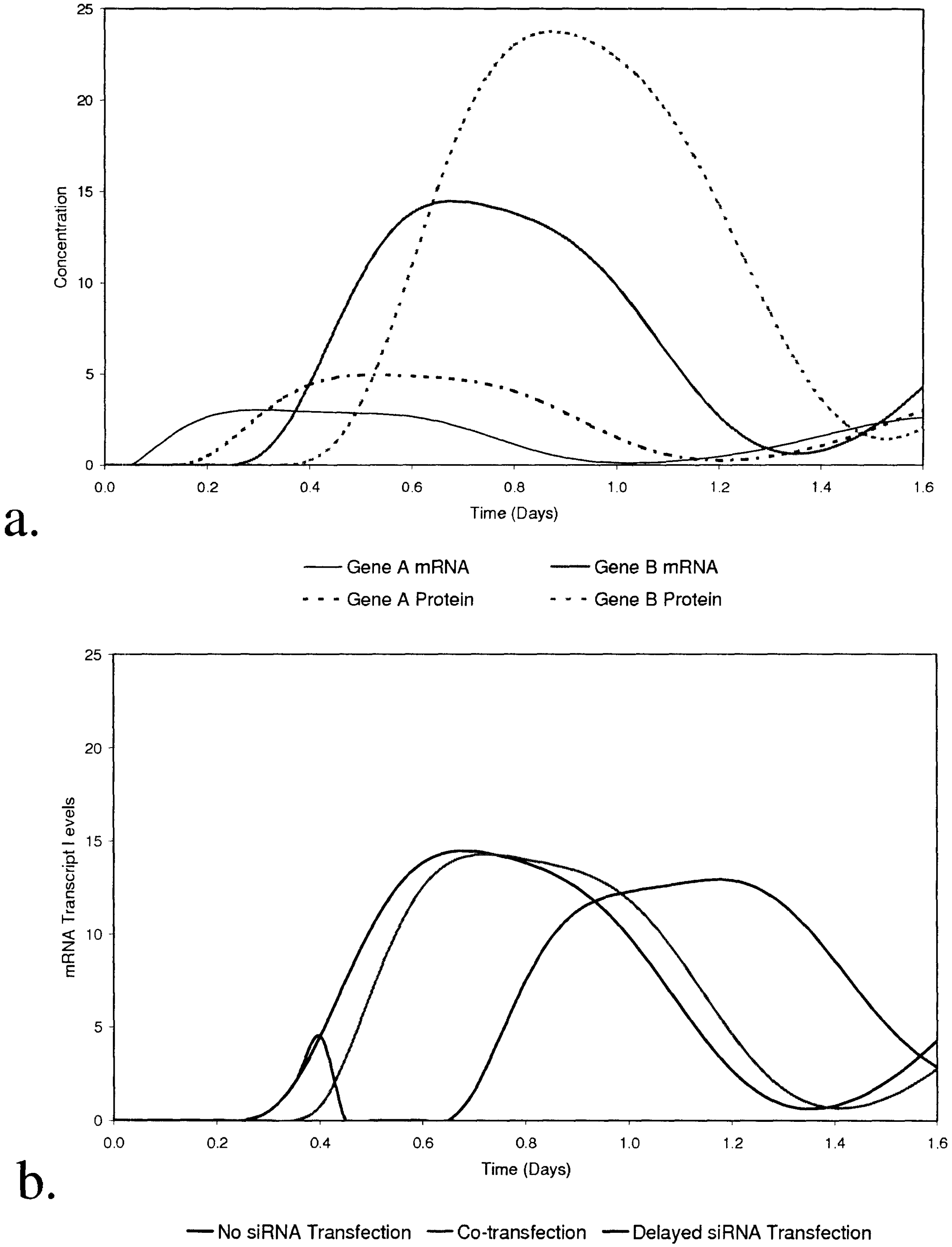


Figure 3-9: Model predictions of RNAi–based gene silencing in a two gene system. (a) Model predictions for native expression of genes A and B with no silencing. (b) Comparison of transcript levels between gene B silencing treatments. Only small differences are observed between the untransfected negative control and co-transfection conditions. However, the delayed transfection condition becomes 0.5 days out of phase with the negative control.

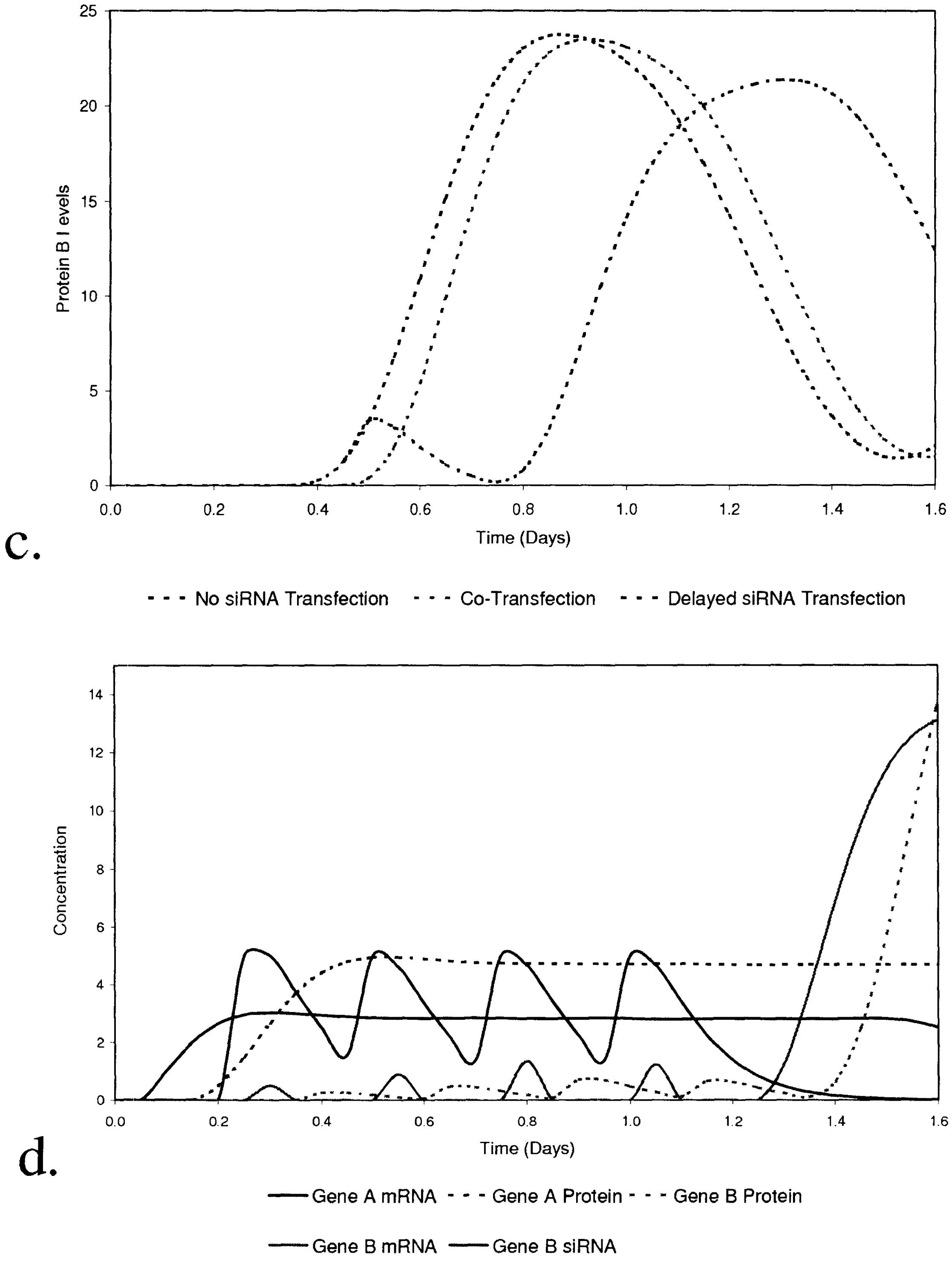


Figure 3-10: Model predictions of RNAi–based gene silencing in a two gene system. (c) Comparison of gene B protein levels between the treatments. While relatively small differences are observed between the untransfected control and co-transfection, delayed transfection results in protein expression that is 0.4 days out of phase. (d) A gene B silencing protocol in which gene B is consistently silenced, allowing the study of gene A in the absence of gene B and oscillatory behavior.

The model used to describe our co–transfection experiment also has utility in designing optimal experiments for endogenous genes that can be activated and repressed. Generally gene silencing will be most effective if the siRNA is transfected at the time of gene induction. However, in some cases we investigated, depending upon the rate constants governing gene expression, gene silencing may require unreasonably large amounts of siRNA. Such procedures are expensive and may be toxic to cells that can only tolerate lower amounts of transfection reagents. These situations can usually be mitigated, maintaining effective silencing for the desired duration, by using much smaller doses with repeated applications. The protocol specifications will be set by the experimental hypothesis and the anticipated variance in the dependent variable, whether RNA, protein, or some other molecule. Additionally, within any desired protocol, other effects should be considered to tailor the regimen for the specific system including the tolerance of the cells to the transfection conditions and length of time over which the silencing must occur. For example, because most RNA transfections are performed in serum free medium (to avoid RNA degradation caused by RNases within the serum), using higher levels of siRNA with lower complex exposure time may be useful when transfecting cells that are sensitive to the absence of serum or other medium constraints.

3.3.2 Use of the RNAi Pathway for Gene Silencing

In rare cases, the use of RNAi may not be feasible to silence a desired gene. If the level of gene transcription is high, but the efficiency of transfection is low, then stable transfection using a retroviral vector, followed by cell selection, may be required, or the use of transgenics to generate a knock–out animal may be unavoidable. Because cell lines can vary greatly in their transfection efficiency, it is prudent to optimize the transfection prior to beginning a set of silencing experiments. In this way, an optimized procedure can be used to expedite the experiments and any constraints that the transfection may create will be recognized and accounted for during experimental design.

Despite RNAi's apparent ease of use, the results can easily be misinterpreted if careful experimental planning and monitoring are not observed. For example, had the GFP siRNA only been transfected 24 or 48 hours after the plasmid DNA, virtually no silencing would have been observed, as shown in Figure 3-5 and 3-8(d). In other cases, even if dynamic experiments are not being performed, if the rate of protein turn–over is slow, then even when transcript levels fall to zero adequate amounts of protein may reside in the cell, performing its normal function, and resulting in no observed phenotypic difference. For these reasons it is important to monitor the gene's products, not only at the transcriptional level, but also at the protein level when possible.

In order to experimentally decompose cascades of interacting genes within a network, knowing the right time to transfect during transient experiments is critical. In this regard, we have shown the utility of a simple model that can help determine when to transfect and how much RNA to use. The model takes into account the rate of transcription, transcript stability, translation efficiency, protein stability, siRNA stability, and rate of degradation due to RNAi–based silencing. In Figure 3-8(d) we see that the model predictions agree qualitatively well with the data, indicating low levels of silencing would be observed if the cells were transfected 24 to 48 hours after the DNA transfection. Conversely, our model of a two gene system described by equations 3.5 – 3.10, transfecting initially with siRNA for gene B has little effect, while transfecting at a later time results in more effective silencing of gene B. Indeed, for a given amount of siRNA, this model could be used to predict the appropriate transfection time for any desired transcript or protein level.

The ability to decrease gene expression, or partially silence a gene, at just the right time, may not initially seem enormously valuable, until one starts to consider networks of interacting genes that often underlie quantitative traits. Baltimore, *et. al.*, recently dissected the interactions between nuclear factor kappa-B (NFkB) and its small family of inhibitors, IkB-a,b,e [110]. In this work, the researchers relied upon knock–out cell lines to finally establish the correct interactions between the components, which showed oscillatory behavior. Not only could this have been done with much less effort using RNAi, but by partially silencing certain components at precise times, the researchers could have potentially shown cycle dampening and other phenomena, assuming the mathematical model they proposed is correct.

It's important to note, that although our model of differential equations describes these processes in terms of "rate constants," none of these model parameters are actually constant, and all need to be assessed for a given set of experimental conditions and genes. For example, depending upon the cellular environment, transcription may be enhanced or repressed through the binding of regulatory elements upstream of the gene [24, 244]; likewise transcripts can possess differential stability depending in part upon their UTR sequences [268]; translation efficiency can also be affected by UTR sequences [191]; and the distribution of enzymes involved in the proteasome degradation pathway may affect protein stability [84]. Certainly for specific genes, different mechanisms may influence any of these steps. The point is that in order to utilize the model, the parameters should be determined under the desired set of conditions so that the model represents as closely as possible the *in vivo* situation.

The rate constants for all of the equations are easily determined using a variety of experimental techniques. If the gene is known to be inducible, then simply measuring the transcript abundance following induction will lead to the transcription rate constant. Conversely, spike and chase assays [79], or the nuclear run–on assay [239], can determine the amount of new synthesis, which over time can also determine the rate. Transcript degradation can be determined by exposing cells containing a transcript pool to the RNA polymerase II inhibitor, actinomycin, and then tracking the rate of pool degradation [203]. With this information, one can compare the rates in the presence and absence of siRNA, to determine the degradation rate due to the RNAi–gene silencing pathway. Protein synthesis can be measured in similar ways, but often relies on the use of labeled essential amino acids [5]. Protein degradation can be measured by exposing cells containing a protein pool to the translation inhibitor, cycloheximide, and then tracking the fall in protein levels over time. One can use either fluorescently labeled, or radio–labeled, siRNA in transfection experiments to determine the rate of degradation of the RNA *in vivo*.

RNA interference has the potential to greatly expedite our understanding of gene function. Experimentally it is easier to implement than transgenic technologies, however still requires careful planning and monitoring to be effective in a given experiment. Here we have shown that both the amount of RNA used in the transfection, as well as the complex exposure time during the transfection, can be altered to facilitate the experimental design. These can be important factors in planning more complex experiments, where the silencing of multiple genes is involved.

We have also shown that the dynamics of gene transcription are important to consider when using RNAi–based gene silencing. Whether conducting static or dynamic experiments, one must know the appropriate time to transfect when using siRNA. Using a GFP–expressing plasmid as our model system, we have shown that vastly different results can be obtained depending upon when the siRNA is transfected relative to gene induction. Because many phenomena of interest rely on gene induction, we feel having a good understanding of the transient nature of RNAi, and the associated implications is important to adequately using the technique. To help investigate and understand our model system, we created a mathematical model that could track the concentration of all key components. The RNAi system fits into this framework well, suggesting that the model will have utility in planning, and analyzing, future experiments.

3.3.3 Methods

Cell Culture

Hepa1–6 mouse hepatoma cells (ATCC) were expanded in T25 flasks (Corning) containing 10 mL of DMEM medium (Gibco, 25 mM glucose, 4 mM glutamine, phenol red; formulated to 1% with penicillin-streptomycin and 10% with fetal bovine serum) until confluent at approximately 4 x 10^6 cells per flask. These cells were then trypsinized and used to inoculate 24–well plates (BD Falcon), which were monitored for confluency (typically 2 - 3 days post–inoculation). All cells were cultured in an incubator at 37 °C with a 5% CO_2 atmosphere. All experimental treatments were conducted in either triplicate or quadruplicate.

DNA Transfection

Hepal–6 cells expressed green fluorescent protein (GFP) following transfection with pTracer plasmid DNA (Invitrogen), which encodes the GFPuv gene, expressed from a cytomegalovirus (CMV) promoter. For each well, 0.7 μg of DNA was diluted into a final volume of 50 μL of serum free Opti-MEM medium(Gibco). In a separate tube, 1.4 μL of Lipofectamine 2000 (Invitrogen) was mixed with 48.6 μL of Opti-MEM medium per well. The two mixtures were allowed to incubate for five minutes, were mixed, and then allowed to incubate for another 20 minutes to promote the formation DNA–carrier complexes. While the mixture incubated, the medium from each well of the 24–well plate was removed, and the cells were washed with sterile, pre-warmed, PBS (Gibco). Following the incubation, each well received 100 μL of the complex mixture, and was allowed to incubate. One milliliter of fresh medium was added to each well after the incubation was complete. Mock-transfected control cells were treated in the exact same manner, however, the mixture used did not contain any pTracer DNA.

RNA Transfection

Hepal–6 cells were transfected with synthesized siRNA. In these experiments two types of siRNA were used: siGFP RNA (Qiagen, catalog # 1022064) and non-silencing, siNS, RNA (Qiagen, catalog # 1022076, catalog #1022079). Our transfection protocol used Lipofectamine 2000 as the RNA carrier because a previous comparison using Transmessenger (Qiagen) or Oligofectamine (Invitrogen) carriers, demonstrated that Lipofectamine 2000 resulted in the highest transfection efficiency as shown in Figure 3-11.

The amount of siRNA used per well varied depending upon the experiment, however, the general procedure used was always consistent. Typically the siRNA was diluted into a final volume of 50 μL of serum free Opti–MEM (Gibco) medium. Serum free medium is necessary to avoid RNase activity that may be present in serum and degrade the siRNA. In a separate tube, Lipofectamine 2000 (Invitrogen) was mixed

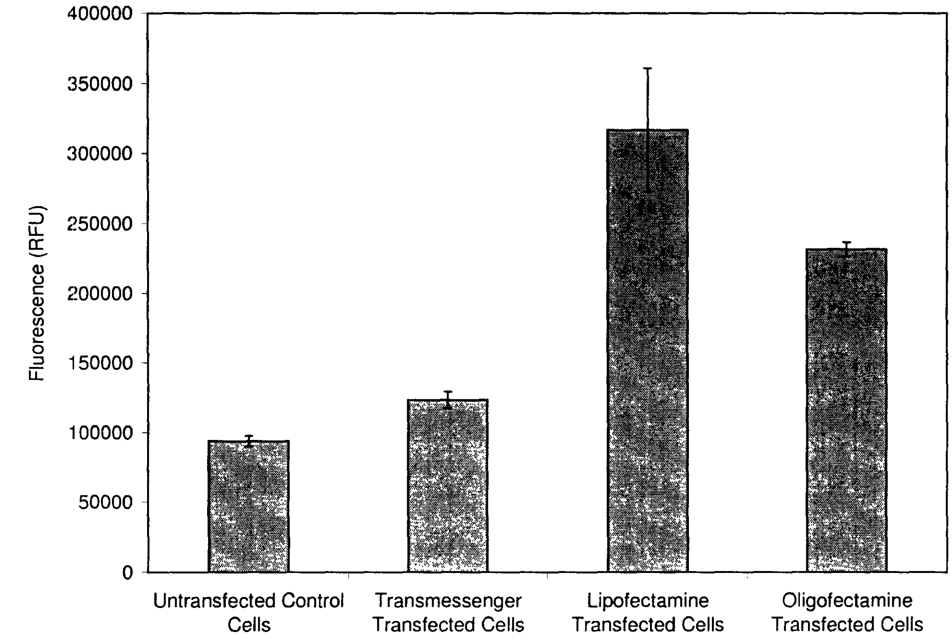


Figure 3-11: Comparison of RNA transfection reagents.

into a final volume of 50 μL of serum free Opti–MEM medium per well. Twice the amount of Lipofectamine 2000, as measured in μL, was used for a given amount of siRNA, as measured in μg. Thus for a mixture containing 1 μg per well of siRNA, 2 μL per well of Lipofectamine 2000 was used. Previous experiments showed that there was no significant difference whether 2 μL or 3 μL of Lipofectamine per μg of siRNA were used (data not shown). The two mixtures incubated for five minutes, were mixed, and then allowed to incubate for another 20 minutes. While the mixture incubated, the medium from each well of the 24–well plate was removed, and the cells were washed with sterile, pre–warmed, PBS (Gibco). Following the incubation, each well received 100 μL of the siRNA complex mixture, and was allowed to incubate for at least two hours unless otherwise noted.

Fluorescence Measurements

Both GFP and fluorescein fluorescence measurements were made by transferring the washed cells to black 94–well plates and using a plate reader with appropriate optical filters (Packard). Before transferring the cells, the growth medium was removed and the cells were washed with 1 mL of prewarmed, sterile PBS. After washing, the cells were trypsinized using 100 μL per well of trypsin (Gibco). The entire cell slurry was then transferred to the 94–well plate and fluorescence measured on the plate reader.

RT–PCR Measurements of GFP mRNA Levels

Total RNA was isolated from individual wells using the RNeasy kit (Qiagen) according to the manufacturer's instructions. Briefly, the cells from each well were treated with 500 μL of buffer RLT formulated with β-mercaptoethanol. The resulting lysate was transferred to a QIAshredder column (Qiagen) and homogenized using centrifugation for 2 minutes at maximum speed. One volume of 70% ethanol was added, and the ethanol–lysate mixture was loaded onto RNeasy mini spin columns provided with the kit. The columns were washed once with 700 μL of buffer RW1, and then twice with 500 μL of buffer RPE, and finally eluted using 30 μL of RNase free water (Ambion). Typical yields were between five and 10 μg of total RNA per well, and were stored at -30 °C until processed.

Our RT–PCR assay uses a two-step protocol in which complementary DNA (cDNA) is first synthesized, then diluted and used in a polymerase chain reaction (PCR). For cDNA synthesis, 1 μg of total RNA was mixed with 1 μg of oligo-dT$_{18-20}$ (Invitrogen), heated at 70 °C for ten minutes, and then mixed with 2 μL of 5 mM dNTPs (Invitrogen), 2 μL of 100 mM DTT (Invitrogen), 4 μL of 5X First Strand Buffer (Invitrogen), and 200 U/ mg of Superscript II reverse transcriptase. The final volume of each reaction was 20 μL, with the remainder consisting of RNase free water. The reverse transcription reaction proceeded for two hours at 42 °C. Once complete, the remaining RNA was degraded by addition of 1.5 μL of 1 N NaOH and incubation at 65 °C for 10 minutes. The NaOH was then neutralized by the addition of 1.5 μL of 1 N HCl, and 3 μL of the final mixture was diluted into 297 μL of RNase free water.

RT–PCR was conducted in 94–well plates using the iCycler RT–PCR machine (Bio-Rad). Briefly, 1 μL of the final, diluted cDNA template was mixed with 19 μL of RNase free water, 25 μL of Bio–Rad RT–PCR Supermix (Bio-Rad), 2 μL of sense and antisense primers, and 1 μL of 12.5 mM dNTPs. The final primer concentration was 0.25 mM. The sense primer sequence was 5'- GGTGTTCAATGCTTTTCCCG - 3', and the antisense primer sequence was 5' - CGCGTCTTGTAGTTCCCGTC - 3'. The resulting PCR fragment is 128 nucleotides long and has been verified using gel electrophoresis. As an internal control we used a similar procedure to monitor the levels of β-Actin mRNA, which are assumed to be constant over most experimental conditions. The PCR cycle used a single three minute hot-start at 95 °C, followed by 50 cycles of 30 seconds at 95 °C, one minute at 60 °C, and two minutes at 72 °C, during which time the reaction fluorescence was measured.

GFP standards were developed by amplifying the entire GFP gene by PCR, cleaning the resulting mixture, and then diluting it to concentrations from 10^{-4} μg/ μL to 10^{-9} μg/ μL. The R^2 value of the standard curve, relating the threshold cycle to the amount of GFP standard, was always greater than 0.97.

Chapter 4

Hepatic Gene Identification

Having developed methods of gene identification described in Chapter 2, and tools to accommodate gene characterization described in Chapter 3, we began our investigations into Type II diabetes using a relevant mouse model. Initially our work focused on discovering genes that were important to mediating the liver's response to diet–induced obesity, insulin resistance, and insulin sensitivity from fasting and weight reduction [201].

4.1 Gene Identification Strategy

Obesity is a growing concern in the industrialized world. It is estimated that over 61% of adult Americans are overweight or obese [128] and an alarming number of children and adolescents are following suit [86]. Of primary concern are the associated complications stemming from obesity's growing prevalence, among which type II diabetes is reaching epidemic proportions.

The liver plays a critical role in glucose homeostasis by secreting glucose into the blood during the postabsorptive state. During insulin resistance, hepatic glucose output (HGO) increases and several key molecules contributing to this phenotype have been widely studied [38, 132, 182, 273]. Despite these extensive efforts, the genes identified thus far do not alone account for all of the variability in HGO, which is a complex, quantitative phenotype. Further insight may be obtained by conducting genome wide transcriptional studies during diet induced obesity (DIO) and its associated insulin resistant physiological state. This approach is a critical step towards further defining the molecular processes that regulate the phenotype and thereby augment the discovery of potential therapeutic targets.

C57/BL/6J mice fed a high–fat diet become obese, hyperglycemic, and hyperinsulinemic, reflecting an insulin resistant metabolic state [61, 76, 148, 189, 234] that resembles the human condition. Although it has been demonstrated that short–term caloric restriction can improve insulin resistance [13], the regulatory pathways that control hepatic metabolism during DIO and associated insulin resistance, and the improvement of insulin resistance with caloric restriction, are the focus of intense research efforts. The molecular mechanisms underlying these pathways rely upon alterations in gene transcription [180], which can be monitored using DNA microarrays [122, 243].

To investigate hepatic gene regulation in response to DIO and insulin resistance, whole genome microarrays containing 17,280 gene probes were used to examine transcription in two groups of C57/BL/6J mice : 1) the "control mice" received a normal diet for 10 weeks, 2) the "high–fat mice" received a high–fat diet for 10 weeks. In addition, to assess hepatic gene regulation in response to caloric restriction, which is a commonly recommended treatment for DIO and insulin resistance, a third group of mice was used, the "fasted/ weight reduced mice," which was fed the same high–fat diet for ten weeks followed immediately by 48 hours of fasting, returning their weights to baseline levels prior to tissue harvest. Fasting/ weight reduction data provides further differentiation among genes that not only respond to DIO and insulin resistance, but are also normalized by caloric restriction.

An extensive bioinformatics analysis led to the identification of 41 discriminatory genes participating in key molecular pathways in DIO, insulin resistance, and fasting/ weight reduction. The implicated pathways involve signal transduction and protein metabolism and secretion. In addition, the 41 genes identified can accurately classify the three groups of mice ("control," "high–fat," and "fasted/ weight reduce"), and importantly, they represent a set of candidate genes that may influence hepatic function during periods of insulin resistance and sensitivity.

4.2 Experimental Results

The effect of 10 weeks of high–fat feeding and 48 hours of caloric restriction on body weight in C57/BL/6J mice

C57/BL/6J mice significantly increased their body weight by 32% after 10 weeks of high–fat feeding ($p < 0.001$; Table 4.1). After 48 hours of fasting, their weights returned to baseline levels and were not significantly different from the control mice, but were significantly less than mice maintained on the high-fat diet ($p < 0.001$; Table 4.1).

Diet	Feeding Regimen	Weight 48 hours Prior to Harvest (Average ± St. Dev., n)	Weight at Harvest (Average ± St. Dev., n)
Normal Chow	Control	35.6 ± 1.8, 9	35.6 ± 1.5, 9
High-Fat	Control	$47.1 \pm 5.8^{\star}$, 9	$51.7 \pm 4.4^{\star\dagger}$, 5
High-Fat	Fasted		37.3 ± 2.6, 4

Table 4.1: Experimental treatments and mouse weights.

*Indicates that the weight was statistically different from the control at P< 0.001.
†Indicates that the weight of the high-fat and fasted mice was different at P< 0.001.

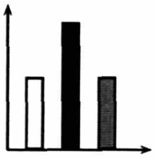
Microarray analysis of hepatic genes after 10 weeks of high–fat feeding and 48 hours of fasting/ weight reduction in C57/BL/6J mice

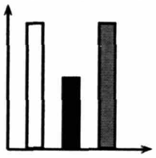
To determine hepatic gene transcription levels, total RNA was isolated from liver tissue of control, DIO–C57/BL/6J mice, and DIO–C57/BL/6J mice fasted for 48 hours. The RNA was fluorescently labeled during a reverse transcription reaction and hybridized to DNA microarrays that were used to measure the transcript abundance of each gene.

Employing statistical and data mining methods we searched the transcription data set for hepatic genes that show statistically significant responses during DIO, associated insulin resistance, and fasting/ weight reduction. We used the t–test to determine the statistical significance of every pairwise gene difference between the treatments. The t–test showed that 1981 genes had at least one statistically significant ($p < 0.05$) change between the treatments. Within this gene set, 113 genes were significantly changed between the high–fat fed mice and the control mice, 169 genes were significantly changed between the fasting/ weight reduced mice and the control mice, and 260 genes were significantly changed between the high–fat fed and fasting/ weight reduced mice, all at $p < 0.01$. From the 1981 genes selected by the $p < 0.05$ cutoff, we retained the 1169 genes that had a Wilks-λ value below our cutoff criterion of 0.47, which is equivalent to a p–value of less than 0.05 [115]. From these genes we selected those with the greatest Fisher Discriminant Analysis (FDA) and Principle Component Analysis (PCA) loading coefficients [54], resulting in the 41 genes reported in Table 4.2.

To show individual gene responses to the dietary treatments, the 41 genes in Table 4.2 were classified into six groups according to changes in the p-values from pairwise comparisons between the control mice, the high–fat fed mice, and the fasting/ weight reduced mice. This classification arranges the genes according to their transcript levels during the physiological states examined. For example, Group A in Table 4.2 comprises genes that were significantly elevated or repressed ($p < 0.05$) by high–fat feeding, but then normalized to (insignificant, $p > 0.05$) control levels by fasting and weight reduction. Similarly, group B genes were significantly elevated or repressed ($p < 0.05$) by high–fat feeding and partially normalized to control levels by fasting/ weight reduction: the expression differences are still significant ($p < 0.05$) when comparing both the high–fat and control mice with the fasted/ weight reduced mice. The genes of each group along with their normalized expression levels are given in Table 4.2.

Group A





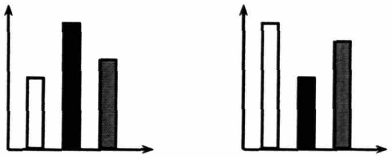
White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up- or down-regulated by the high-fat diet and normalized to control levels by fasting/ weight reduction.

Gene Name	GenBank #	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
Crym	NM_ 016669	1.01	0.47	1.27	μ-Crystallin
Cyp2c37	NM_ 010001	2.29	0.55	2.21	Cytochrome P450
Eva	BC015076	-0.70	0.00	-0.45	Epithelial V-like Antigen
Kcnk8	NM_ 010609	1.33	-0.08	0.95	Potassium Channel
Ndph	NM_ 010883	1.45	2.16	1.14	Neuron Function
Pmm1	AK004631	2.48	1.37	2.28	Protein Secretion
Serpina5	NM_ 008785	0.22	0.99	0.32	Serine Protease Inhibitor
Sh3kbp1	AK004636	0.16	1.43	0.26	Signaling
1110034G24Rik	AK004090	1.53	0.38	1.20	Unknown
1700019L13Rik	AK006130	1.84	0.98	2.03	Unknown
4930442L21Rik	NM_ 026253	0.15	1.29	0.36	Unknown
4930579D07Rik	AK016314	1.62	0.37	1.14	Unknown
4932422M17Rik	AK016534	0.70	0.22	1.19	Unknown

continued on next page

Group B

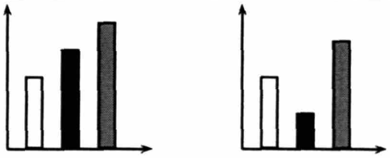


White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up- or down-regulated by the high-fat diet and partially normalized to control levels by fasting/ weight reduction.

Gene Name	GenBank #	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
Eif4a2	NM_ 013506	2.02	0.60	1.23	Translation
Fshβ-like EST	AK017593	3.00	1.68	2.67	Hormone
Mup4	NM_ 008648	3.23	1.82	2.58	Secreted Protein
PTP4a2	NM_ 008974	1.57	1.07	1.41	Signaling
RGS3	AF350047	2.50	0.97	1.62	Signaling
Tcam1	NM_ 029467	2.02	0.98	1.56	Adhesion Molecule

Group C



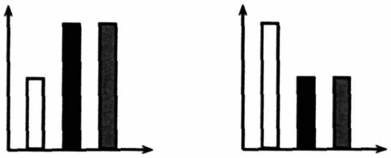
White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up-regulated by high-fat diet and fasting/ weight reduction, or genes down-regulated by high-fat diet and up-regulated by fasting/ weight reduction.

Gene Name	GenBank #	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
1500004A08Rik	AK005141	-0.73	-0.15	0.53	Unknown
2300009A05Rik	AK009046	0.49	0.15	1.16	Unknown
2810055C24Rik	AK012951	1.17	0.52	1.52	Protein Degradation

continued on next page

Group D



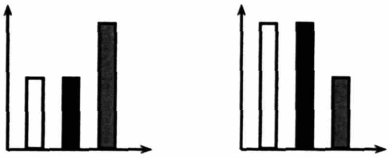
White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up- or down-regulated by the high-fat diet and fasting/ weight reduction.

Gene Name	GenBank #	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
Bmp2	NM_ 007553	1.79	2.40	2.86	Development
Copz2	NM_ 019877	3.47	4.88	5.85	Vesicle Trafficking
Fosb	NM_ 008036	0.68	1.89	2.25	Signalling
Gabrr1	NM_ 008075	0.49	1.78	2.34	Receptor
Has3	NM_ 008217	2.15	0.71	1.12	Hyaluronan Synthesis
IL6st	NM_ 010560	0.46	1.09	0.99	Signalling
Rab3c	NM_ 023852	1.01	1.69	2.39	Exocytosis
Ttr	NM_ 013697	2.90	2.23	1.70	Hormone Transport
1110007C24Rik	AK014449	0.78	2.13	2.98	Unknown
2700062B08Rik	NM_ 029838	2.17	3.58	4.19	Putative Collagen-like (CLAC)
3110004A18Rik	AK013988	0.81	2.18	3.02	Unknown
4833414G15Rik	AK019515	0.75	1.45	2.11	Putative Phosphatase
4933432M07Rik	AK017027	-0.63	0.53	0.30	Protein Degradation
5730458M16Rik	AK017674	-0.32	0.68	0.91	Unknown

continued on next page

Group E

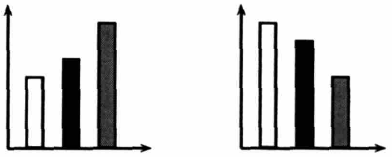


White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up- or down-regulated by fasting/ weight reduction.

Gene Name	GenBank \#	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
Ctrl	NM_ 023182	0.31	0.59	2.02	Protein Metabolism
Resp18	NM_ 009049	1.21	1.03	0.33	Hormone Secretion
Snrpg	NM_ 026506	0.68	0.51	1.67	Translation
2310034L21Rik	NM_ 025631	-0.84	-0.54	0.33	Unknown

Group F



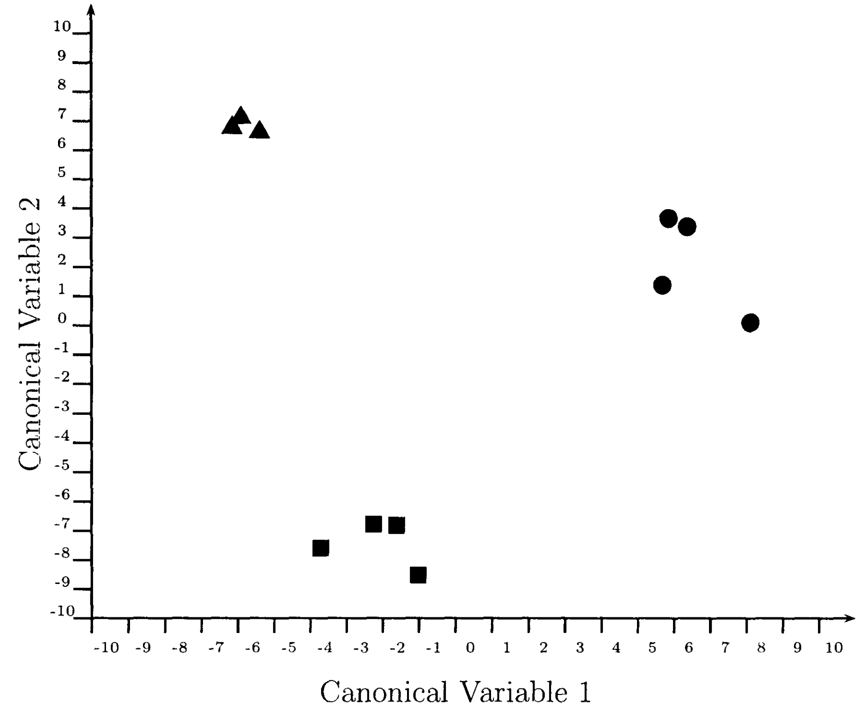
White Bars = Control Mice
Black Bars = High-Fat Mice
Gray Bars = F/ WR Mice

Genes up- or down-regulated by high-fat diet and fasting/ weight reduction.

Gene Name	GenBank \#	Control Mice	High-Fat Mice	F/ WR Mice	Biological Role
1700095D18Rik	AK007076	0.88	1.54	2.43	RNA-binding Region

Table 4.2: Log_2 ratios of genes found common to all analysis methods. Included are genes identified using t-test, Wilks-λ ranking, fisher discriminant analysis, and principle component analysis. These genes are organized by their pairwise t-test results, and the relation between their $\log_2$ ratios as shown by the corresponding bar charts. The $\log_2$ ratios of each group are measured relative to a standard reference RNA sample (see Methods). F/ WR: Fasting/ Weight Reduced.

The 41 discriminating genes contributed to the classification observed in Figure 4-1. In Figure 4-1, each sample is given two canonical variable (CV) scores, based on weighted sums of its gene expression values. The genes with the largest contributions to CV1 and CV2 are given in Table 4.2, suggesting these genes underlie the biological differences between the samples. Figure 4-1 shows that 10 weeks of high–fat feeding altered the transcriptional levels of genes so as to separate the control and high–fat mice in the CV1 and CV2 space. However, while 48 hours of fasting/ weight reduction normalized many of the genes contributing to CV2, resulting in a return to control levels for that variable, the genes contributing to CV1 remained perturbed, resulting in the observed separation between the fasted/ weight reduced mice and control mice. This suggests that while some genes, and their associated pathways that differentiate DIO and insulin resistance from normal physiology, return to control levels as weight is reduced, other genes remain perturbed, reflecting further physiological adaptations that occur during these treatments.



Circles = Control mice Squares = High–Fat mice Triangles = F/WR mice

Figure 4-1: Fisher discriminant analysis plot of mouse liver samples. Samples were scored according to the canonical variables determined by Fisher Discriminant Analysis (FDA). Each canonical variable is defined as a weighted sum of 100 specific genes, including each of the 41 genes contained in Table 4.2. To score a sample, the gene expression value is multiplied by an FDA coefficient, called a loading, and the products from the 100 genes used in the analysis are summed to give the canonical variable score for the sample. F/ WR: Fasting/ Weight Reduced.

Among the 41 discriminatory genes identified in this study, interleukin 6 signal transducer (*IL6st*), protein tyrosine phosphatase 4a2 (*PTP4a2*), SH3-domain kinase binding protein 1 (*Shk3bp1*), and regulator of g–protein signaling 3 (*RGS3*) are of special interest because based on known biology they may contribute to the physiological changes that accompany DIO, insulin resistance, and increased insulin sensitivity due to fasting/ weight reduction. Both *IL6st* and *Sh3kbp1* are significantly upregulated after 10 weeks of high–fat feeding ($p < 0.001$), but only *Sh3kbp1* is normalized to baseline levels after 48 hours of fasting and weight reduction (Table 4.2). Both *PTP4a2* and *RGS3* are significantly downregulated after 10 weeks of high–fat feeding ($p < 0.01$), and both are partially normalized after 48 hours of fasting/ weight reduction ($p < 0.01$ for fasted/ weight reduced versus high–fat and fasted/ weight reduced versus control; Table 4.2).

RT–PCR analysis of IL6st, PTP4a2, RGS3, G6P, PCK1, and malic enzyme

We compared the transcript levels measured by RT–PCR with the ratios measured using DNA microarrays by dividing RT–PCR expression values observed in high–fat fed mice and fasted/ weight reduced mice by the expression values measured in the control mice. Liver mRNA levels for each mouse in the study were determined by RT–PCR for *IL6st, PTP4a2*, and *RGS3*. The values measured by RT–PCR were not significantly different from the results observed by hepatic microarray analysis ($p > 0.05$; see Table 2.3 on page 43) for all genes except *IL6st* between the fasting/ weight reduced mice and control mice. Notably, in this single case, both microarray analysis and RT–PCR show significant increases ($p < 0.001$) in the levels of *IL6st* mRNA, demonstrating similar qualitative changes between the measurement methods. The close agreement between the micoarray results and RT–PCR results validates the specificity and accuracy of our microarray measurements. The difference in the ratios between the values determined by RT–PCR and those determined by microarray analysis was less than 30% for each of these genes (see Table 2.3 on page 43).

Although several commonly studied genes, such as glucose–6–phosphatase (*G6P*), phosphoenolpyruvate carboxykinase (*PCK1*), and malic enzyme, were eliminated by our bioinformatics analysis, we evaluated their expression by RT–PCR because of the considerable attention they have received in the literature in connection with hepatic glucose output. *G6P* and *PCK1* were upregulated following 10 weeks of high–fat feeding, but only the change observed in *G6P* achieved statistical significance ($p = 0.09$ for *PCK1* and $p < 0.01$ for *G6P* in the high–fat versus control comparison; Table 4.3). Fasting/ weight reduction resulted in even larger increases in mRNA levels for both *G6P* and *PCK1* ($p < 0.01$ versus controls; Table 4.3). In contrast, malic enzyme exhibited significant underexpression following 10 weeks of high–fat feeding, with further down–regulation following fasting/ weight reduction (Table 4.3).

Genes	Assay	High Fat vs. Control	F/ WR vs. Control
G6P	RT–PCR	476 ± 72% *	769 ± 216% *
PCK1	RT–PCR	132 ± 28%	217 ± 80% *
Malic Enzyme	RT–PCR	9.1 ± 1.5% *	0.1 ± 0.1% *

Table 4.3: RT–PCR results for *G6P*, *PCK1*, and malic enzyme. Expression levels are represented as a percent of expression in the control mice.

*Indicates that the measurements were significantly different from control values at $P < 0.01$.

4.3 Discussion of Gene Identification Findings

Diet induced obesity (DIO) in C57/BL/6J mice is a commonly used animal model for the development of insulin resistance in humans [61, 76, 148, 189, 234], which results in simultaneous hyperglycemia and hyperinsulinemia. Although short–term caloric restriction and weight loss can improve insulin resistance [13, 65, 104], the regulatory mechanisms in the liver that lead to insulin resistance in response to DIO, as well as the improvement of insulin sensitivity in response to short–term caloric restriction and weight reduction, remain largely unknown. To identify genes involved in hepatic physiology during DIO and short–term caloric restriction, we used DNA microarrays to measure genome–wide transcript abundance.

The 41 most discriminating genes determined by our bioinformatics analysis lie essentially within two large groups (see Table 4.2 on page 98): 1) Genes that are significantly induced or repressed by 10 weeks of high–fat feeding and completely (Group A) or partially (Group B) normalized by 48 hours of fasting/ weight reduction, 2) Genes that are significantly induced or repressed by 10 weeks of high–fat feeding, but are not normalized by 48 hours of fasting/ weight reduction (Group D). Both of these groups contain genes involved in signal transduction pathways, as well as protein metabolism and secretion, highlighting the importance of these molecular pathways in the hepatic response to DIO and fasting/ weight reduction.

Because genes in Group A and B (Table 4.2) were perturbed by DIO, their expression levels correlate with observed physiological differences that develop during this condition. These differences include elevated concentrations of serum triglycerides, leptin, and tumor necrosis factor-α, as well as changes in the levels of other factors that have been previously demonstrated to play a physiological role during DIO in C57/BL/6J mice [1, 76, 148, 189, 234]. Notably, Group A and B genes are either completely (Group A, Table 4.2) or partially (Group B, Table 4.2) normalized following 48 hours of fasting/ weight reduction, when insulin sensitivity has increased, suggesting they may be important to the development of hepatic insulin resistance during DIO.

Several relevant signal transduction pathways are influenced by the genes within Group A and B (Table 4.2), particularly *Sh3kbp1*, *PTP4a2*, and *RGS3*. While *Sh3kbp1* and *PTP4a2* may be directly involved with insulin signaling, by respectively binding PI-3-kinase and dephosphorylating protein tyrosine residues, *RGS3* interacts directly with G–proteins and some evidence suggests RGS family members may also indirectly affect proteins in the MAPK signal transduction pathways [136] as well as certain tyrosine phosphatases [140].

Sh3kbp1 (SH3–domain kinase binding protein, also called Ruk) belongs to the CD2AP/ CMS family of adapter-type proteins, which mediate a number of different cellular mechanisms including signal transduction [250]. Insulin signaling occurs via phosphorylation of insulin receptor substrates (IRSs) that interact with signal transduction molecules including PI-3-kinase, Grb2, nck, and SHP2 [267]. Sh3kbp1 has been shown to directly inhibit PI-3-kinase signaling by binding the p85α regulatory subunit *in vivo* and *in vitro*, and interacts with Grb2 *in vitro* [87]. Therefore, increased levels of *Sh3kbp1* mRNA in the high–fat fed mice relative to both the control and fasted/ weight reduced mice, suggests that Sh3kbpl may mediate DIO associated insulin resistance in hepatocytes via a mechanism described in Figure 4-2.

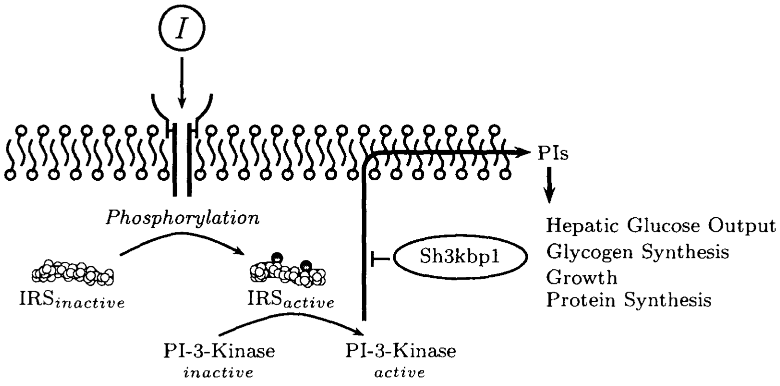


Figure 4-2: Inhibition of PI-3-Kinase signaling by Sh3kbpl. In this figure, insulin, *I*, binds to its receptor, activating the receptor's tyrosine kinase activity. Insulin receptor substrates, IRS, are activated by phosphorylation. IRS phosphorylates PI-3-kinase, which migrates to the cell membrane where it generates phosphatidylinositol, PI, second messengers, which alters physiological processes. Shown here, Sh3kbpl is capable of binding the regulatory subunit of PI-3-kinase, inhibiting its ability to generate PI second messengers, and thereby attenuating insulin signaling.

PTP4a2 (Protein tyrosine phosphatase 4a2) dephosphorylates tyrosine residues in proteins. When insulin binds its receptor it activates the receptor's tyrosine kinase activity [36], leading to autophosphorylation and subsequent tyrosine phosphorylation of molecules containing Src homology 2 (SH2) or phosphotyrosine binding (PTB) domains, such as insulin receptor substrates (IRSs). Therefore PTPs can influence insulin signaling by dephosphorylating protein tyrosine residues. Although it would be anticipated that PTPs would attenuate insulin signaling, they have been implicated in both positive and negative regulation of this pathway [8]. A definitive role for many PTPs in glucose homeostasis and insulin signaling has not been established, however, *PTP1B* knock–out mice have enhanced insulin sensitivity and are resistant to DIO [64]. Therefore if PTP4a2 also negatively regulates insulin signaling, its significant downregulation ($p < 0.01$) following 10 weeks of high–fat feeding may be a physiological adaptation that helps protect hepatocytes against insulin resistance, which is normalized by fasting/ weight reduction.

RGS3 (Regulator of G-protein coupled receptor (GPCR) signaling 3) has been primarily studied in neurons [7, 120, 225] and cells in culture [31, 253]. RGS proteins bind Gα subunits and generally increase the GTPase activity [224]. We found that hepatic *RGS3* mRNA levels are significantly decreased ($p < 0.01$) after 10 weeks of high–fat feeding, but partially normalized by fasting/ weight reduction. These findings are particularly relevant because hepatocytes express a truncated form of RGS3 that has been shown to directly inhibit $G_s\alpha$ stimulated cAMP production and $G_q\alpha$ stimulated IP production [33], in addition to interacting with, $G_i\alpha$ [175]. Glucagon signals via a GPCR that stimulates adenyl cyclase and increases cAMP levels [14]. Because the truncated form of RGS3 inhibits cAMP production, lowering RGS3 concentration may augment basal cAMP levels and thereby promote hepatic glucose output resulting from cAMP induced phosphoenolpyruvate carboxykinase (PCK1) expression and cAMP repressed glucokinase transcription. Although glucokinase expression levels were not measured, *PCK1* mRNA levels were increased by both 10 weeks of high–fat feeding and fasting/ weight reduction (see Table 4.3 on page 101).

While genes in Group D (see Table 4.2 on page 98) were also significantly induced or repressed following 10 weeks of high–fat feeding, unlike genes in Group A and B, they do not respond to 48 hours of fasting/ weight reduction. Therefore hepatic regulation of Group D genes may not be as directly linked to changes resulting from DIO and insulin resistance and sensitivity. Despite this, it is interesting that a number of Group D genes are also implicated in several signal transduction pathways that may be activated by DIO. These genes include *BMP2*, *Fosb*, *Gabrr1*, *IL6st*, and *4833414G15Rik*.

BMP2 (Bone morphogenetic protein 2), is a highly conserved member of the transforming growth factor-β (TGF-β) gene family. BMP2 is related to BMP9, which was the first reported hepatic factor shown to decrease blood glucose levels by increasing insulin release and decreasing food intake [34]. While these mechanisms may be a compensating response to DIO, they oppose the physiological adaptations that accompany 48 hours of fasting/ weight reduction, and therefore additional studies are required to determine the effects of BMP2 upregulation in mice following these dietary treatments.

FosB is a member of the AP-1 family of transcription factors [4]. These molecules are considered immediate early genes, because they initiate responses to environmental stimuli [218]. The Fos family of transcription factors form either homodimers with one another, or heterodimers with the Jun family of transcription factors, which then bind DNA to alter gene transcription [172]. Because insulin affects the expression of members of the AP-1 family of transcription factors [179], it is not surprising that during DIO and fasting/ weight reduction, conditions that perturb insulin signaling, significantly increase transcription of *FosB*.

IL6st (Interleukin 6 signal transducing subunit, also called gp130) is a key component in cytokine signal transduction that occurs during inflammation through the JAK (Janus kinase)/ STAT (signal transducers and activators of transcription) pathway. IL6st forms homo- and heterodimers with other signal transducing subunits in response to binding by an assortment of ligands including IL–6, IL–11, LIF, CT–1, CNTF, and OSM [105]. Among these, IL–6 knockout mice develop mature–onset obesity [254], and treatment of hepatocytes with IL–6 reduces the expression of PCK1 [38], thus implicating IL–6 in the regulation of hepatic glucose output. There are at least four different Jaks (Jak1, Jak2, Jak3, and Tyk2) and seven different STAT factors (STAT1, 2, 3, 4, 5a, 5b, and 6) that can interact with IL6st. Of particular relevance to DIO and insulin resistance is STAT3. The liver–specific STAT3 knockout mouse is insulin resistant and develops glucose intolerance when fed a high–fat diet, due in part to increased expression of PCK1 and G6P [116]. Adenoviral mediated reconstitution of STAT3 signaling ameliorated glucose intolerance in both L–ST3KO and *Lepr-/-* mice [116] by lowering PCK1 and G6P levels, demonstrating the importance of STAT3 signalling to hepatic glucose output. Because *IL6st* is significantly upregulated ($p < 0.001$) by 10 weeks of high–fat feeding and 48 hours of fasting/ weight reduction, when *PCK1* and *G6P* were also induced relative to control levels (see Table 2.3 on page 43), it may be that IL6st performs a sensitizing function that contributes to feedback control of hepatic glucose output via IL6 and STAT3 signaling.

In addition to the cellular signaling pathways that contained differentially expressed genes identified in this study, a number of genes involved in protein metabolism and secretion were also identified. Although a direct link between protein metabolism/ secretion and DIO/ insulin resistance is not as well established, in other insulin sensitive tissues the release of hormones and trafficking of receptors clearly plays a role in regulating tissue specific responses to insulin and glucose. Group A and B genes involved in protein metabolism and secretion pathways include *Kcnk8*, *Pmm1*, *Serpina5*, and *Eif4a2*. Group D genes that were identified include *Copz2*, *Rab3c*, and *4933432M07Rik*.

Serpina5, encodes a serine protease inhibitor. Serine protease inhibitors represent a family of glycoproteins that are known to inactivate serine proteases by forming stoichiometric enzyme–inhibitor complexes. Among the proteases known to be inhibited by Serpins are trypsin, chymotrypsin, the sperm protease acrosin, and a variety of proteases involved in hemostasis [274]. *Copz2* encodes a vesicle coating protein that helps to mediate vesicle trafficking, while *Rab3c* is a member of the Ras oncogene family that encodes a monomeric GTP-binding protein that is implicated in regulated exocytosis and vesicle transport, and has been suggested to play a role in GLUT4 translocation in rat cardiac muscle cells [248]. Hence, Copz2 and Rab3c may synergistically influence protein trafficking in response to 10 weeks of high–fat feeding and 48 hours of fasting/ weight reduction.

Using DNA microarrays we have investigated the effects of DIO and fasting/ weight reduction on liver gene transcription. We have analyzed this data set using four computational methods that represent a rigorous approach to analysis requiring no *a priori* assumptions about the data. This has enabled us to infer the importance of any given gene change among a multitude of gene differences resulting from DIO and fasting/ weight reduction. Our results lead us to focus on 41, out of an initial 1981 genes.

Although many of the genes resulting from our analysis have not yet been studied extensively in the context of energy homeostasis, several are related to important molecular pathways that have been previously identified in the literature. Those pathways include different signal transduction cascades, as well as pathways involved in protein metabolism and secretion. Given the diverse functions of the liver, identifying genes involved in signaling and protein metabolism pathways in response to DIO and fasting/ weight reduction is not surprising. Among the genes involved in signaling are *Sh3kbp1*, *Rgs3*, *PTP4a2*, *BMP2*, *IL6st*, *Fosb*, *Gabrr1*, and possibly *Rab3c*. Genes implicated in protein metabolism and secretion pathways include *Crym*, *Serpina5*, *Eif4a2*, *Ctrl*, *Snrpg*, *Kcnk8*, *Copz2*, and *Rab3c*.

While the link between many of these genes and DIO will require further investigations, their identification here is an important contribution to understanding how the hepatic response to DIO and fasting/ weight reduction is mediated through a variety of molecular pathways. These genes all share a consistent set of attributes that made them stand out in the data set. They demonstrate significant differences between the dietary treatments, are individually discriminatory of each treatment, and are members of a set that classifies each sample using both supervised and unsupervised algorithms. Genes that satisfy all of these criteria represent good candidates for influencing the liver's response to DIO and fasting/ weight reduction, and therefore warrant more detailed investigations.

4.4 Methods

Animals

Three to five week old C57/BL/6J mice were obtained from Jackson Laboratories (Bar Harbor, ME). All animals were allotted a seven day acclimation period with access to food and water *ad libitum*, and were maintained at 25 °C with a 12-hour light/ dark cycle (lights on from 06:30–18:30) for the duration of the study. A normal chow (Purina Rodent Chow; Harlan Teklad #5008; 6.5% fat, 49% carbohydrate, 23%

protein, 3.5 kcal/ g) and high–fat diet (Harlan Teklad #TD88137, 42.16% fat, 42.81% carbohydrate, 15.02% protein, 4.53 kcal/ g) were fed to respective mice, as outlined below.

This report explored alterations in hepatic gene mRNA levels in C57/BL/6J mice fed either a control or high–fat diet for 10 weeks, as well as alterations in mRNA levels of C57/BL/6J mice fasted for 48 hours following 10 weeks of high–fat feeding. Fasted animals were allowed access to water during the fasting period. All animals were sacrificed by CO_2 asphyxiation, followed by immediate collection of liver tissue, which were stored at -80 °C as previously described [278].

The control group consisted of C57/BL/6J mice fed normal chow diet for 10 weeks. The experimental group consisted of C57/BL/6J mice fed a high–fat diet for 10 weeks (n = 9/ group). The ten week high–fat dietary treatment has been demonstrated to be long enough for C57/BL/6J mice to develop insulin resistance and a condition that resembles type 2 diabetes [189, 234]. Two days before tissue harvest, the C57/BL/6J mice on the high–fat diet were divided into two groups, with one group remaining on the high–fat diet (n = 5) and one group fasting for the final 48 hours (n = 4). Mouse weights were recorded two days prior to, and on the day of tissue harvest.

All animals were handled in accordance with the principles and guidelines established by the National Institutes of Health. The protocol was approved by the Institutional Review Board at Beth Israel Deaconess Medical Center, Boston, MA.

Preparation of total RNA and cDNA for microarray hybridization

Total RNA was purified from liver tissue samples using STAT-60 (Tel-Test, Inc., Friendswood, TX) according to the manufacturer's instructions, and stored at -80 °C. Labeled control cDNA was made from Total RNA control samples (Universal Mouse Reference RNA, catalog # 740100, Stratagene) using Cy3 dCTP (Perkin-Elmer), and labeled liver cDNA was made from total RNA experimental samples using Cy5 dCTP (Perkin-Elmer) during reverse transcription, as described previously [28].

Microarrays were prepared using GAPS glass slides (Corning) and a Virtek arrayer (Bio-Rad). Arrays contained 17,280 features, printed from a synthesized oligonucleotide mouse library (Operon) as described previously (see Section 2.2.2) [28].

RT-PCR analysis of IL6st, PTP4a2, G6P, PCK1, and malic enzyme

A two-step RT-PCR protocol was performed to confirm the mRNA levels of several genes. In this procedure the cDNA synthesis was performed as detailed previously [28] except the Cy–labeled nucleotides were replaced with unlabeled nucleotides such that all dNTPs were at the same final concentration during the reaction. PCR was conducted in 94-well plates using the iQ SYBR Green Supermix Kit (Bio–Rad), according to the manufacturer's instructions on an iCycler RT-PCR machine (Bio-Rad). Briefly, 1 μL of the final, diluted cDNA template was mixed with 19 μL of RNase free water, 25 μL of Bio-Rad RT-PCR Supermix (Bio-Rad), 2 μL of sense and antisense primers, and 1 μL of 12.5 mM dNTPs. The final primer concentration was 0.25 μM. The PCR cycle used a single three minute hot-start at 95 °C, followed by 50 cycles of 30 seconds at 95 °C, one minute at 60 °C, and two minutes at 72 °C during which time the reaction fluorescence was measured. Each mouse sample was measured in either triplicate or quadruplicate.

The sense and antisense primer sequences were: for interleukin 6 signal transducer (*IL6st*) 5'- GCGGCTCGAACTTCACTGC - 3', and 5' - CACGATGTAGCTGGCATTCACG - 3'; for protein tyrosine phosphatase 4a2 (*PTP4a2*) 5'- TTTCTGCTGCGGAACATTTCAAG - 3', and 5' - GCGTGCGTGTGTGAGTGTG - 3'; for regulator of g–protein signalling 3 (*RGS3*) 5'- GCACATCCCGCATTCCAGTTAC - 3', and 5' - AGGGAACACCAGGACTTTAGGG - 3'; for glucose–6–phosphatase (*G6P*) 5'- GTGATTGCTGACCTGAGGAACG - 3', and 5' - TGCCACCCAGAGGAGATTGATG - 3'; for phosphoenolpyruvate carboxykinase (*PCK1*) 5'- CAGAGAGACACAGTGCCCATCC - 3', and 5' - AAGTCCTCTTCCGACATCCAGC - 3'; for malic enzyme 5'- GCCAGAGGATGTCGTCAAGG - 3', and 5' - ATTACAGCCAAG-

GTCTCCCAAG - 3', respectively. These primers each gave specific fragments of the correct length when viewed upon a 4% agarose gel (data not shown). As an internal control β-Actin mRNA levels were also measured. The sense and antisense sequences were 5' - AATAAGTGGTTACAGGAAGTC - 3' and 5' - ATGAAGTATTAAGGCG-GAAG - 3', respectively.

Gene specific standards were developed by amplifying the entire mRNA coding sequence of each gene by PCR from a cDNA library, gel purifying the resulting band, and then diluting it to concentrations from 10^{-4} μg/ μL to 10^{-9} μg/ μL. The R^2 value of the standard curve, relating the threshold cycle to the amount of standard template, was always greater than 0.97. The mRNA levels of β-actin measured were not significantly ($p > 0.05$) different between the dietary treatments for any of the groups.

Computational methods

A combination of statistical and data mining methods were used to extract information from the microarray data (see Section 2.3 on page 44 for more information). Statistical methods rigorously quantify the reliability of differences in the microarray data [171] and can objectively evaluate changes in gene transcription ratios and derivative quantities. Data mining is particularly useful for uncovering patterns and structure in microarray data that might have otherwise been difficult to detect through manual inspection and intuition alone [143, 200].

A t-test [241] was used to evaluate whether a gene exhibited statistically significant expression differences in pairwise comparisons between the control, high–fat, and fasting/ weight reduced groups (see Section 2.3.1). The t-test results showed that 1981 genes had at least one statistically significant ($p < 0.05$) change between the treatments.

Wilks-λ based ranking [115] was used to identify discriminatory genes that differentiated the three groups (see Section 2.3.1 on page 44 for more information). In this analysis a Wilks-λ threshold value of 0.47 was used, which is equivalent to a p–value of 0.05. From the 1981 genes selected by the $p < 0.05$ cutoff, we retained the 1169 genes that had a Wilks-λ value below 0.47.

Fischer Discriminant Analysis [230] (FDA) was used to identify not just individual genes, but combinations of genes whose expression levels are capable of correctly classifying the control mice, high–fat mice, and fasting/ weight reduced mice. FDA is based on *linear combinations of gene expressions* and considers the discriminatory power of gene groups as opposed to individual genes (see Section 2.3.2). As shown in Figure 4-1, using expression data of the selected gene combinations allows accurate classification of the dietary treatments suggesting that the genes in Table 4.2 are discriminatory of the conditions examined when sample classification is used as a criterion. On the basis of the successful classification afforded by the FDA projection, discriminatory genes were selected using the magnitude of the loading coefficients.

Principle Component Analysis [54] was used as an unsupervised classification procedure to complement FDA. The results of the PCA analysis largely mirrored the FDA results (data not shown).

Methods used here, along with the data set, are available for public use at our laboratory's web-site [19]. The entire data set is also available through the National Center for Biotechnology Information's Gene Expression Omnibus database [173, 59].

Chapter 5

Effect of Hepatic Gene Silencing on Metabolism

Forty–one genes were identified in Chapter 4 based upon their differential expression between the control, high–fat, and fasting dietary treatments. Because these genes demonstrate significant expression differences between the treatments, are individually discriminatory of each treatment, and are members of a set that classifies each sample using both supervised and unsupervised algorithms, they are good candidates that may influence the liver's response to diet–induced obesity and fasting/ weight reduction. In this chapter we apply the gene silencing methods developed in Chapter 3 to a subset of these genes to assess their effects on hepatic metabolism.

5.1 Hepatic Gene Characterization

Insulin resistance leads to excess hepatic glucose output (HGO), which is driven primarily by gluconeogenesis and glycogenlysis, and whose genetic basis is not understood. In Chapter 4, DNA microarrays were used to analyze hepatic gene transcription in a mouse model of diet–induced obesity (DIO). From this work, 41 genes were rigorously identified as good candidates for further study.

Unfortunately, many of the genes listed in Table 4.2 have not been studied in detail and are not known to be directly involved in glucose production. This is sometimes a problem when interpreting the results of microarray studies: many genes are differentially expressed and can be used to classify the experimental treatments, however, they have little or no support from other scientific studies in the literature. Furthermore, the information required to assess their biological relevance, or place them in a physiological context, cannot be gathered from only transcription studies which do not answer questions such as:

- Is the observed difference in phenotype the cause or the result of differences in transcript levels?
- Does the change in transcript levels correspond to a change in protein levels?
- Is the gene being actively induced, repressed, or is there a difference in transcript degradation rate?
- What is the function of the gene's RNA or protein product?

To answer these questions, other experiments are required to complement the microarray results.

A common approach is to select gene candidates from the set that have been studied previously and evaluate their regulation and function with respect to the new phenotype of interest. This might proceed by:

- Searching for gene mutations in animal models that either susceptible or resistant to a phenotype of interest, such as obesity or insulin resistance.
- Monitoring the gene's expression and splicing variants in other tissues.
- Studying the gene's tissue specific regulation.
- Determining the mRNA half–life.
- Cloning the mRNA coding sequence and expressing it in different cellular models.
- Performing biochemical studies on the protein product.

- Determining whether the protein is secreted or localized to specific portions of the cell.
- Identifying other proteins or molecules that bind the gene's protein product.
- Disrupting the gene in a mouse model or cell line.

While many different experiments can be conducted to assess a gene's physiological effects, from this list it is clear that proceeding directly from microarray experiments to the detailed task of gene characterization is a large undertaking. Even in the case where there are 41 strong genetic targets, performing this barrage of experiments on each gene is an enormous task. Thus, when studying specific molecular phenotypes, such as HGO, one would prefer to further decrease the number of gene candidates and find those that are most relevant to the phenotype from within the set.

One of the most effective ways to investigate how a gene influences a phenotype is to disrupt or eliminate the gene and then observe changes in the phenotype. Although *in vivo* manipulation of genes is very time consuming and high–through put evaluation is currently prohibitive for most laboratories (see Section 3.2 on page 64), as discussed in Chapter 3, RNA interference (RNAi) can be effectively used to silence genes and thereby create "functional" gene knock–outs in cellular models. RNAi can therefore be used to screen loss of function gene effects on phenotypes of interest.

Hepatocytes can produce glucose from proteins and amino acids, primarily alanine, as well as from lactate [174]. The chemical reaction network that converts substrate molecules into glucose is shown schematically in Figure 5-1, where each edge represents a chemical reaction between the node molecules. These reactions are catalyzed by enzymes, some of which are reversible and some of which are irreversible. The regulation of these enzymes, and therefore their intracellular catalytic activity, can be controlled at many levels including transcription of the enzymes' genes, translation of the transcript, post–translational modification of the enzyme, and enzyme inhibition.

Normally, during the postabsorptive period, when insulin levels are falling, the liver produces glucose first by breaking down glycogen via glycogenlysis, and then

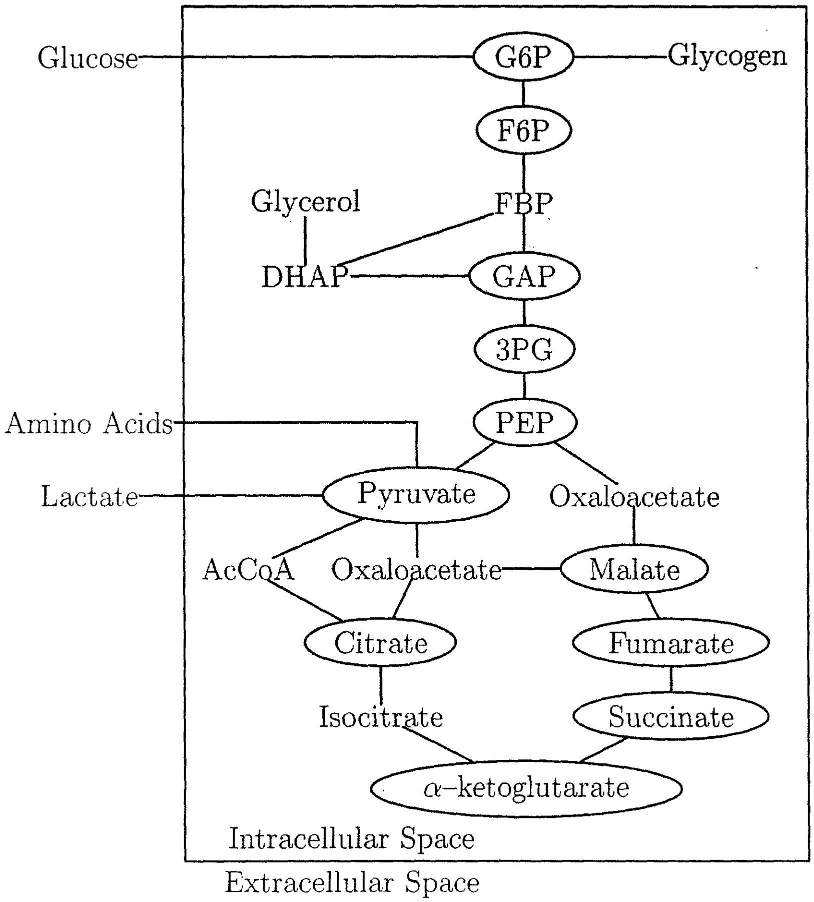


Figure 5-1: Metabolic network for hepatic glucose production. This figure shows key components of the metabolic network employed by hepatocytes to produce glucose from amino acids, primarily alanine, or lactate. Circled molecules can be measured using a combination of gas chromatography and mass spectrometry (GC–MS). Abbreviations: G6P (glucose–6–phosphate), F6P (fructose–6–phosphate), FBP (fructose–1,6–bisphosphate), DHAP (dihydroxyacetone phosphate), GAP (glyceraldehyde–3–phosphate), 3PG (3–phosphoglycerate), AcCoA (acetyl–coenzyme A), PEP (phosphoenolpyruvate).

from amino acids and lactate through gluconeogenesis, by catalyzing the reactions in Figure 5-1 that link these sets of molecules. Conversely, during the postprandial period (see Figure 1-1 on page 15) when blood glucose and insulin concentrations are high, hepatocytes take up glucose to store it as glycogen, a glucose polymer, and also stimulate glycolysis. These processes are mediated to a large extent by the binding of insulin to its receptor, which activates a series of cell signaling and regulatory events that control the intracellular metabolic network. For example, in Figure 5-1 the line connecting glucose to glucose–6–phosphate represents two reactions: movement of glucose through the glucose transporter, Glut-2, and phosphorylation of glucose by hepatic glucokinase (also called hexokinase). The second reaction through hexokinase is influenced by insulin receptor binding, which simultaneously inhibits the reverse reaction (conversion of glucose–6–phosphate to glucose) catalyzed by the enzyme glucose–6–phosphatase. The next line, connecting glucose–6–phosphate to glycogen, also represents a series of reactions that are catalyzed by phosphoglucomutase, UDP–glucose pyrophosphorylase, and glycogen synthase. The rate of reaction between glucose–6–phosphate and glycogen (or vice versa) is influenced by insulin, glucagon, and other hormones. Thus, this simple pathway, between extracellular glucose and intracellular glycogen, proceeds through five different enzymes, each of which is produced by gene transcription, mRNA translation, and some of which are regulated by post–translational modification.

Hepatic glucose output, which is a complex, quantitative trait that increases during insulin resistance, is not genetically defined and a variety of different gene mutations may present with the same phenotype (see Section 1.1), or alter the rate of HGO. The rate of HGO ultimately depends on the collective flux, or reaction rate, of glucose from all hepatocytes, which in turn depends upon the flux of the individual intracellular reactions that constitute the glycogenlysis and gluconeogenesis pathways. The flux through an enzymatic reaction is dictated by two variables: substrate concentration and enzyme activity. However, enzyme activity is a function of the concentration of the enzyme (and therefore transcriptional and translational control), mutations in the enzyme or other regulatory proteins, the concentration of products and other chemical species, and post–translational modifications. Thus even in a modestly sized network similar to that shown in Figure 5-1, the regulatory complexity and number of genes involved may be substantial.

The fluxes of the cell ultimately define its physiological state, which can be described in part by the distribution of intracellular metabolites. Like transcriptional profiling, metabolite profiling can be used to classify experimental treatments and differentiate among many physiological states. This is particularly important when studying complicated phenotypes like hepatic glucose output, which may present similar symptoms, but have diverse molecular causes.

From the DNA microarray results of Chapter 4, we selected 15 genes that were overexpressed during one of the dietary treatments and screened their effects on hepatic metabolism using RNA interference. The screening experiments were carried out in primary hepatocytes isolated from C57/BL/6J mice and metabolites were isolated from cultures of these cells treated individually, or combinatorially, with siRNA(s) to silence specific genes that were up–regulated in the feeding studies (see Chapter 4). The results demonstrate that certain genes significantly perturbed metabolite levels in ways that emulated conditions of decreased glucose production. In addition, while these genes individually had a small but significant impact on reducing hepatic glucose output, their combined effects were substantial, as measured in our assay. Thus by using combinatorial siRNA screening we were able to rapidly find genes that decreased hepatic glucose output in our primary cell model. Because they were initially discovered based upon their gene expression values in a relevant animal model, they may by important for the regulation of HGO in C57/BL/6J mice.

5.2 Gene Silencing Strategy for Studying Hepatic Metabolism

To study the effects of gene silencing on intracellular metabolite concentrations, a relevant model system was required that could induce both increased and decreased hepatic glucose output levels, and was still amenable to efficient gene silencing. We used primary hepatocytes isolated from C57/BL/6J mice as our model system. The advantages of these cells are that they can produce glucose (in contrast to hepatoma cell lines that generally only take up glucose), retain some of their differentiated physiological characteristics, and can be readily used in gene silencing experiments. The drawbacks to using these cells are that they must be isolated directly from the mice, which limits the size of experiments that can be performed, they are no longer connected to the other tissues *in vivo*, and that relatively large numbers of cells must be used to obtain reproducible signals from gas chromatography and mass spectrometry (GC–MS). Furthermore, because of the large number of cells required and their intermittent availability, the cost of these experiments (in terms of both time and money) is quite high. Thus to increase speed and decrease costs, a combinatorial approach was employed to conduct gene silencing experiments, once control experiments had validated the model system.

In the combinatorial approach, fifteen siRNAs were selected and split into three groups of five. Primary hepatocytes were transfected with each group of siRNAs, and groups that had a substantial effect on the metabolite profiles were split into smaller groups in subsequent experiments. In this way it's possible to screen the effects of all 15 siRNAs and determine the primary effect of a single siRNA within a maximum of seven experiments using a binary search. Figure 5-2 illustrates the screening approach.

There are a number of caveats to investigating the phenotypic effect of different genes using combinations of siRNAs. First, we are assuming that a single gene has the primary effect observed, which may or may not be true. Because these genes all share common selection criteria (see Chapter 4), being identified under the same experimental conditions, they may actually work together or have some additive or interactive effects. Second, it is known that the RNAi gene silencing pathway can become saturated [22, 221]. Thus, there is a limit on the number of genes, and the corresponding degree of silencing, that can be silenced within a single experiment; this number is dependent to some extent on the absolute levels of gene expression

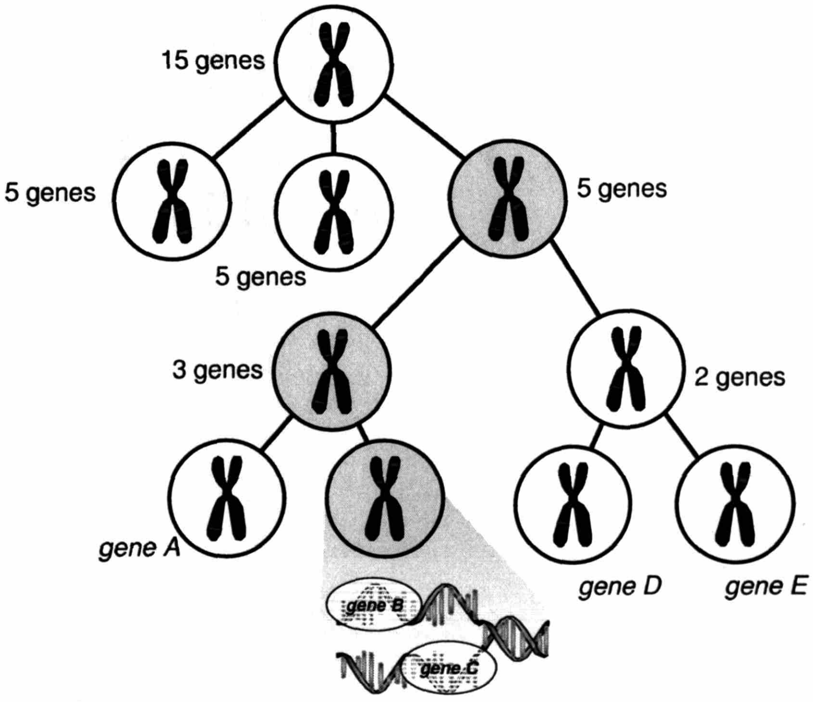


Figure 5-2: Combinatorial siRNA screening strategy for 15 genes. Using this approach the primary effects of a single gene could be discovered in only seven experiments. In this case siRNA groups highlighted in gray have the primary effect on the phenotype, which can be mapped down to the single gene, either *gene B* or *gene C*, using a binary search.

because only a finite amount of dsRNA that can be delivered and utilized by the RISC complex for gene silencing. Third, because this is a screening approach, the expression values for all silenced genes are not measured in the early stages of the screen. It is possible that the degree of silencing for some genes within the screen may be less than adequate to illicit an observable effect, which would increase the number of false negatives. It is unknown what the likelihood is *a priori* of incomplete silencing affecting the experimental results. Given the large number of candidate genes, however, this is an acceptable risk in conducting these experiments.

5.3 Results of Control Experiments

The primary cell culture model was tested for glucose production under gluconeogenic, control conditions, and three conditions of low glucose production: actinomycin treatment, treatment with siRNA specific to glucose–6–phosphatase (*G6P*), and treatment with siRNA specific to phosphoenolpyruvate carboxykinase (*PCK1*). Actinomycin is a non–specific inhibitor of transcription [232, 269] that should lower the levels of all transcripts in the cell. This acts as a positive control, which provides insight to the degree of glucose reduction that we might anticipate from silencing only key genes. The caveat to using actinomycin is that it is a transcription inhibitor, while RNAi based gene silencing relies upon a protein catalyzed pathway (See Section 3.3 on page 67) and therefore is likely to have different kinetics. Nonetheless, actinomycin treatment affects the transcription of all genes and therefore is believed to be a representative positive control. Silencing of *G6P* and *PCK1* act as positive controls for the effects on two key gluconeogenic genes. Both of these genes are known to be important for hepatic glucose output and therefore lower levels of glucose production are anticipated when silencing these genes. The results of these control treatments are show in Figure 5-3.

In Figure 5-3 the actinomycin treatment has the largest effect on HGO, decreasing the amount by over 20%. Likewise, both *G6P* and *PCK1* siRNA treatment decrease HGO by approximately 15%. The corresponding levels of *G6P* and *PCK1* gene

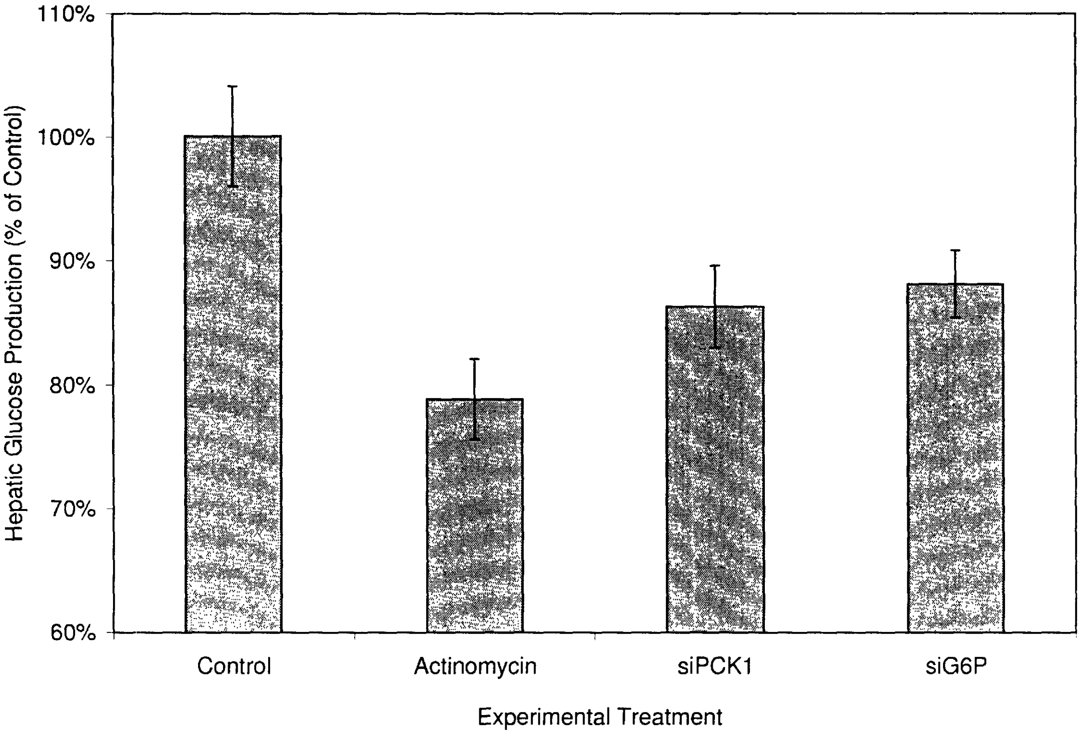


Figure 5-3: Production of glucose by primary hepatocytes. The graph shows hepatic glucose production under high producing control conditions, control conditions treated with actinomycin, and control conditions treated with either *G6P* or *PCK1* siRNA.

expression during these treatments were measured by RT–PCR. Figure 5-4 shows that at the time of sampling, transcript levels of *G6P* and *PCK1* had been substantially reduced by actinomycin and siRNA treatments, relative to control levels.

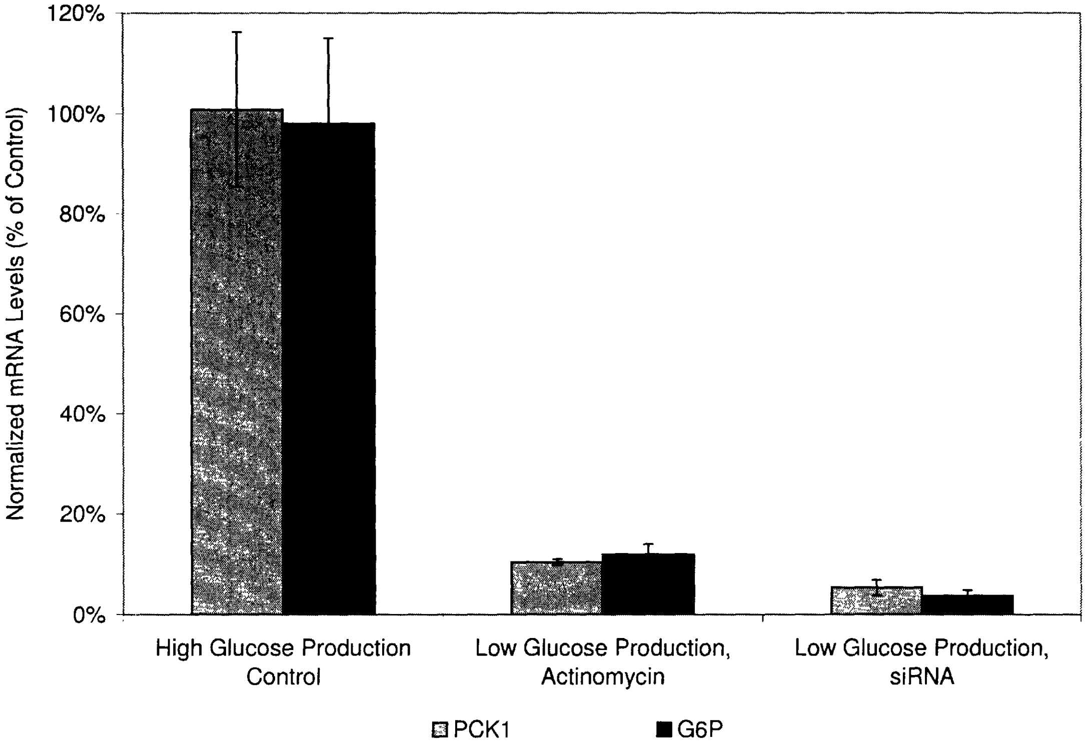


Figure 5-4: Measurement of *G6P* and *PCK1* gene transcription levels from hepatocytes during control, actinomycin, and siRNA treatments.

While these measurements are based upon the amount of glucose produced by hepatocytes over a 24–hour period, it was surprising that neither actinomycin treatment, nor silencing of key gluconeogenic genes resulted in complete suppression of HGO. This suggests that in addition to silencing of overexpressed genes, overexpressing repressed genes may play a key role in further reducing HGO.

In addition to measuring HGO during these treatments, intracellular levels of key metabolites were also measured using GC–MS. The concentrations of α–ketoglutarate (aKG), citrate, isocitrate, succinate, fumarate, malate, pyruvate, phosphoenolpyruvate (PEP), 3–phosphoglycerate (3pg), glyceraldehyde–3–phosphate (GAP), fructose–6–phosphate (F6P), glucose–6–phsophate (G6P), ribose–5–phosphate (R5P), and ribitol were measured within the samples and compared across treatments as shown in

Figure 5-5 and Figure 5-6.

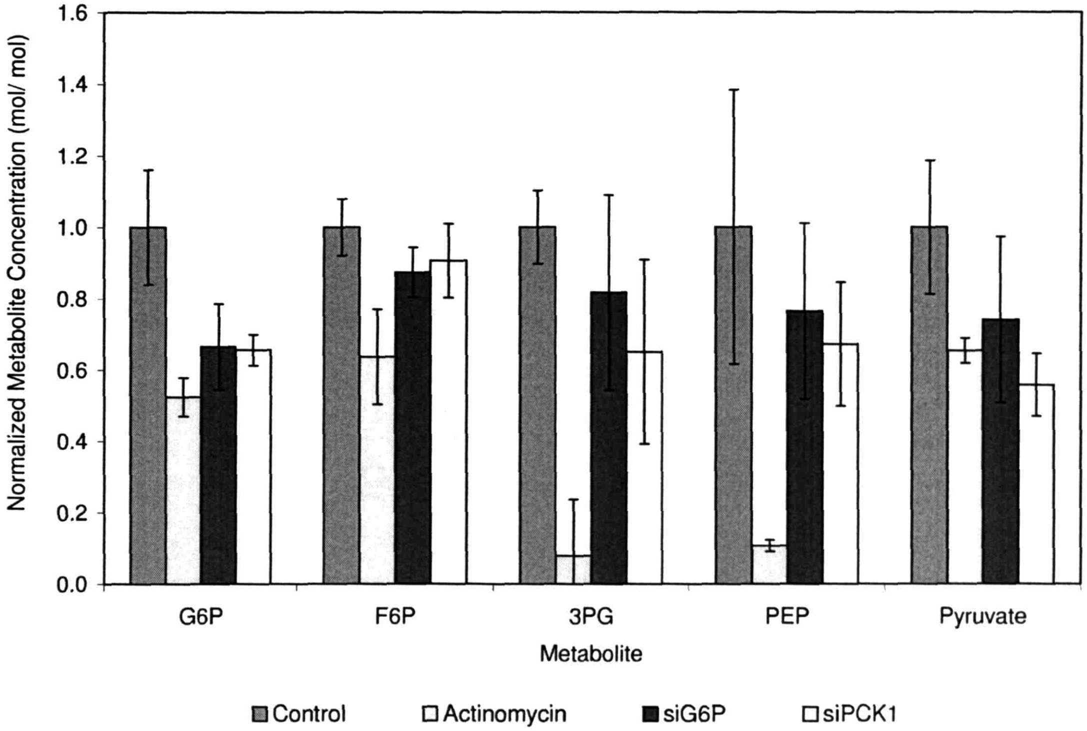


Figure 5-5: Measurement of glycolytic metabolite levels at different hepatic glucose output rates.

In Figure 5-5 actinomycin treatment reduces the concentration of all metabolites, but especially 3PG and PEP. In contrast, silencing of *G6P* and *PCK1* only appears to significantly affect G6P and perhaps pyruvate. Thus, because actinomycin treatment results in a greater decrease in glucose production than silencing of either *G6P* or *PCK1*, to gain further reduction in HGO, genes that affect the pool size of 3PG and PEP may be particularly important for further reducing glucose production. Likewise in Figure 5-6, while actinomycin treatment reduced the pool size of every metabolite in the TCA cycle, it had a particularly strong effect on citrate and α–ketoglutarate. This time, while both *G6P* and *PCK1* gene silencing significantly reduced the size of all metabolite pools, it may be important to find genes that regulate the levels of citrate and α–ketoglutarate to reduce HGO to the same extent as actinomycin treatment.

Although large reductions in several metabolites, such as citrate and α–ketoglutarate,

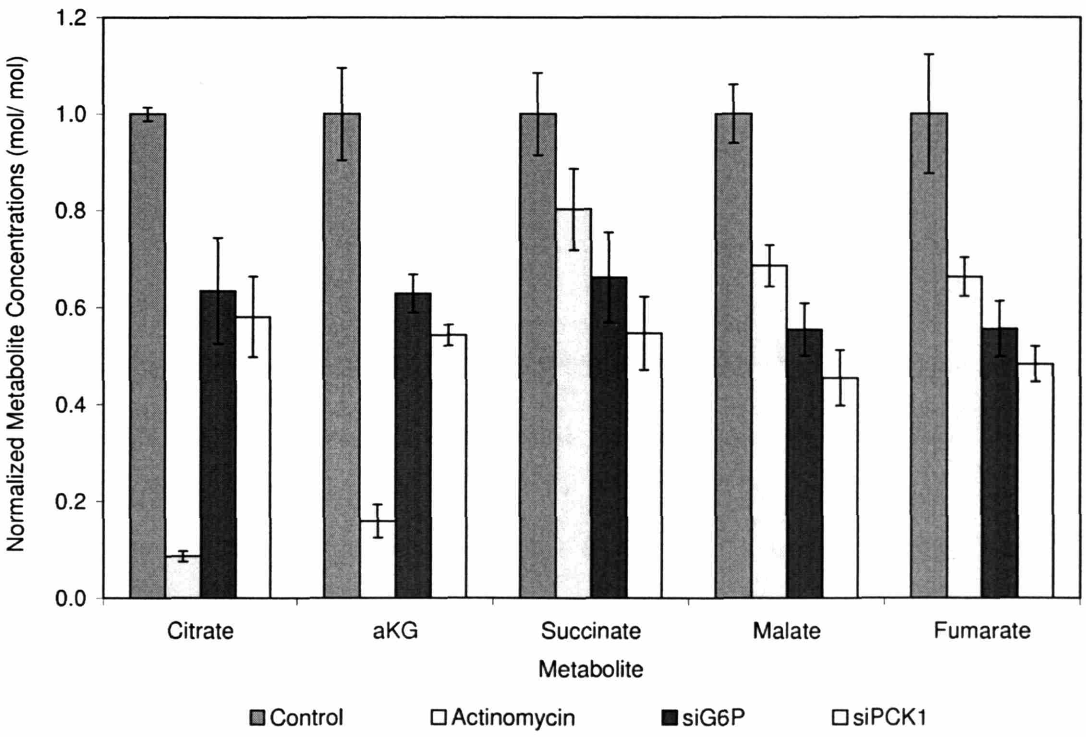


Figure 5-6: Measurement of metabolite levels for compounds within the TCA cycle at different hepatic glucose output rates.

occurred during the actinomycin treatment, which had the lowest amount of hepatic glucose output, the entire metabolite profiles can be used to quantify the metabolic state of the cell. Metabolites were chosen from central carbon metabolism (glycolytic and gluconeogenic pathways), the TCA cycle, and pentose phosphate pathway to monitor hepatic metabolism. The resulting profiles were analyzed with Fisher Discriminant Analysis (FDA) to generate lower dimensional metrics to classify the control and experimental treatments.

Using FDA the control, actinomycin treatment, and *G6P* and *PCK1* siRNA treatments, representing three different levels of HGO, can be separated based upon their metabolite profiles, as shown in Figure 5-7.

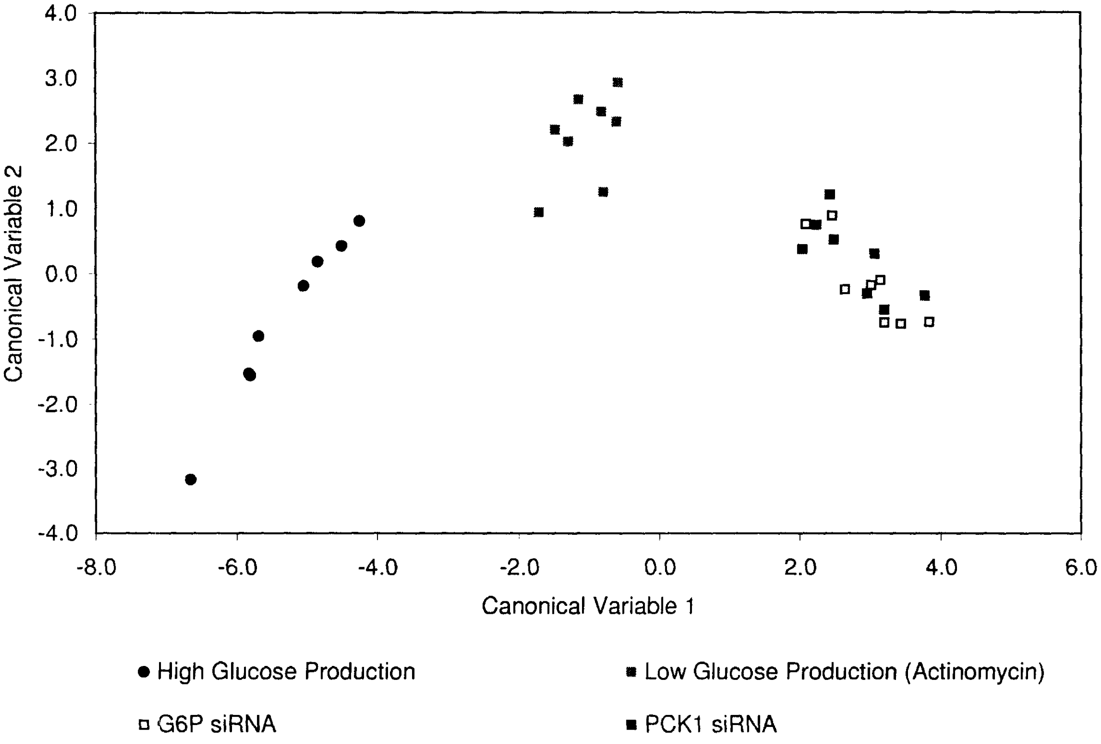


Figure 5-7: Fisher discriminant analysis of control metabolite samples, each with a different hepatic glucose output rate.

Because Figure 5-7 shows a clear separation resulting from the treatments, it indicates that the information required for sample classification is contained within the data set. In Figure 5-7 canonical variable 1 separates all three treatments, while canonical variable 2 separates the actinomycin treated samples from the control samples and samples treated with siRNA.

The loadings (see Section 2.3 for more information) used to create the projection in Figure 5-7 are given in Table 5.1. The selected loadings are used as coefficients to score the samples based on a subset of metabolite concentrations. The metabolites selected for sample classification include α–ketoglutarate, citrate, and phosphoenolpyruvate, which were identified by inspection of the control metabolite profiles as being key metabolites to low glucose producing conditions. Canonical variable 1 captured 71.54% of the variance in the data set, while canonical variable 2 captured the remaining 28.46% of the variance.

Canonical Variable	α-KG	Citrate	F6P	Fumarate	G6P	PEP	R5P
1	0.859	0.683	-0.802	-0.622	-0.701	0.869	0.796
2	-0.513	-0.642	-0.458	-0.327	-0.605	-0.405	-0.634

Table 5.1: FDA loadings for the projection of control samples in the CV1 and CV2 space.

Using the loadings of the canonical variables we can classify experimental samples based upon their metabolite distributions and determine if they reside in high or low glucose producing regions. In this way it is possible to determine which genes and groups of genes from our set impact hepatic metabolism.

5.4 Results of Endogenous Silencing Experiments

Genes were selected from Table 4.2 based upon their expression levels and screened using gene silencing to determine their effect on hepatic metabolism. To expedite the silencing experiments, genes were assembled into groups, shown in Table 5.2, and hepatocytes were treated with each group of siRNAs to determine their effect on the pool sizes of intracellular metabolites.

Figure 5-8 shows the FDA results from hepatocytes treated with either Group I, Group II, or Group III siRNAs relative to the control treatments.

Figure 5-8 shows that although treatment of hepatocytes with Group I and Group II has some effect on metabolite distributions, those samples cluster close to the high

Group	siRNAs
I	*Sh3kbp1, BMP2, Gabrr1, IL6st, Rab3c*
II	*RIK111, Col25a1, Nt5c3, K17RIK, Rnf148*
III	*Eva, LRR, Hgfl, LIM, RIK*

Table 5.2: Groups of siRNAs used in combinatorial screening experiments.

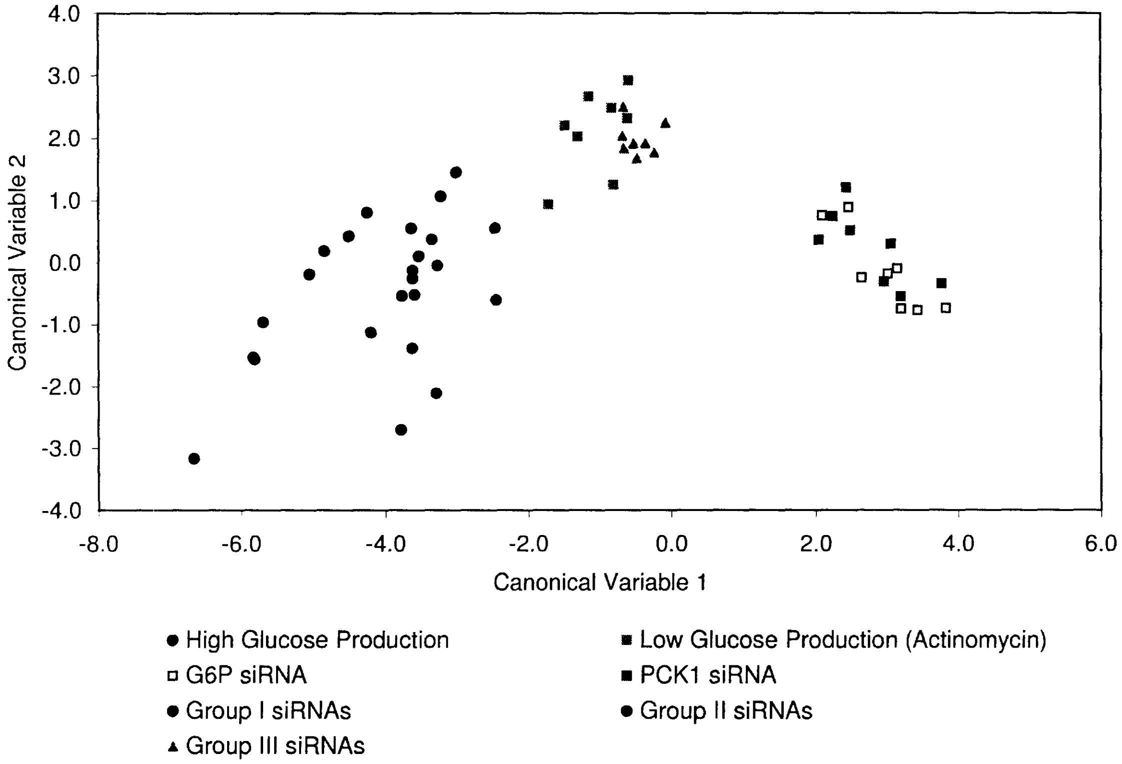


Figure 5-8: Fisher discriminant analysis of hepatic samples treated with either Group I, Group II, or Group III siRNAs from Table 5.2.

producing control samples. In contrast, the Group III samples clustered with the low glucose producing samples treated with actinomycin. Based on these results we proceeded by focusing on the Group III siRNAs (*Eva*, *LRR*, *LIM*, *Hgfl*, *RIK*), measuring the gene expression levels by RT–PCR and conducting additional silencing experiments.

To investigate the effects of the Group III siRNAs, we divided the group into two subgroups, Group A and Group B. Group A was composed of siRNAs for *Hgfl*, *LIM*, and *RIK*. Group B contained siRNAs for *Eva* and *LRR*. Figure 5-9 presents the FDA projection of Group A gene silencing, including silencing results of the individual genes. Figure 5-10, Figure 5-11, and Figure 5-12 present the corresponding RT–PCR results for gene expression during the siRNA treatments. Likewise, Figure 5-13 presents the FDA projection of Group B gene silencing, while Figure 5-14 and Figure 5-15 shows the results of the corresponding RT–PCR measurements.

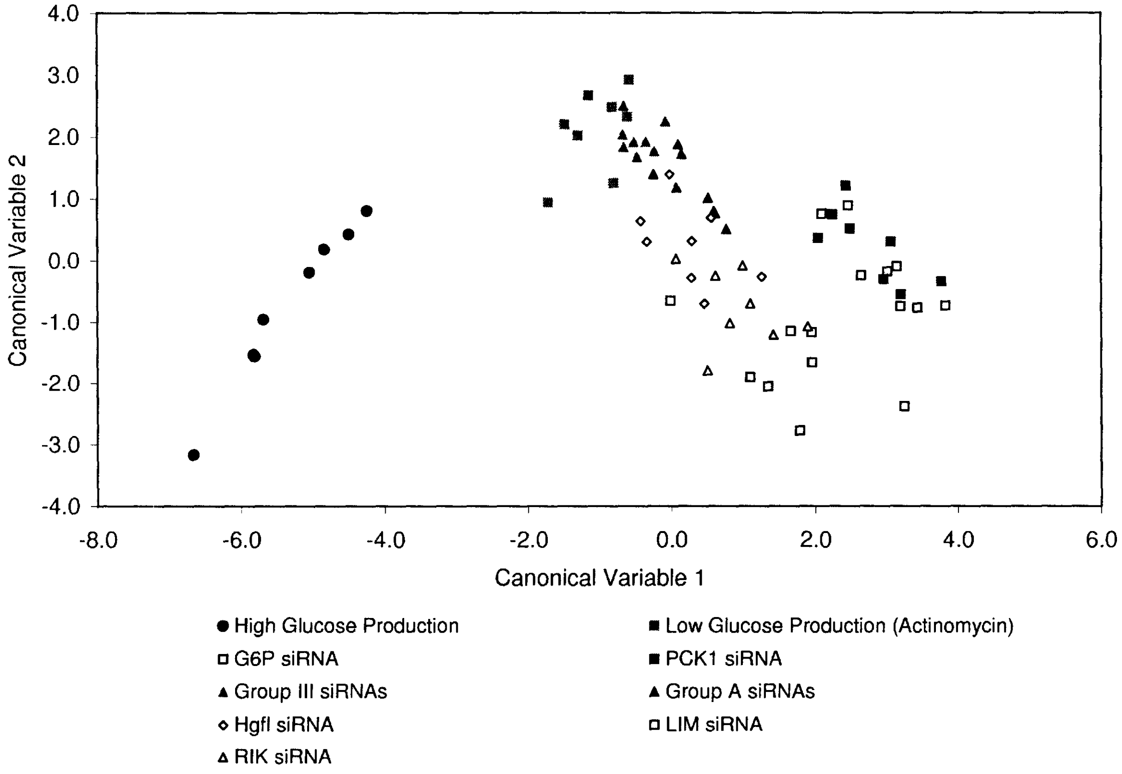


Figure 5-9: Fisher discriminant analysis of Group A siRNAs.

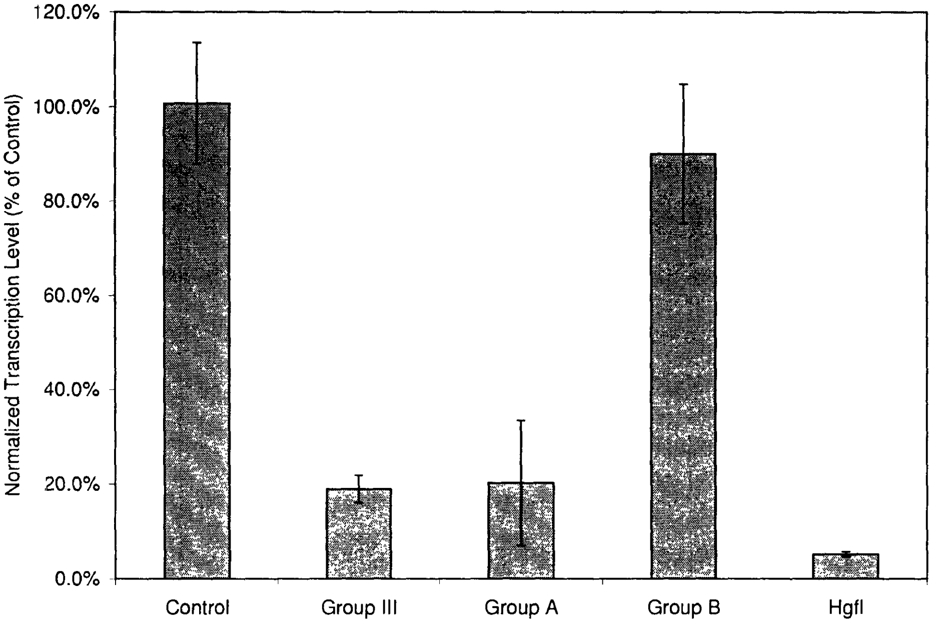


Figure 5-10: RT–PCR results for *Hgfl* expression in different hepatic treatments.

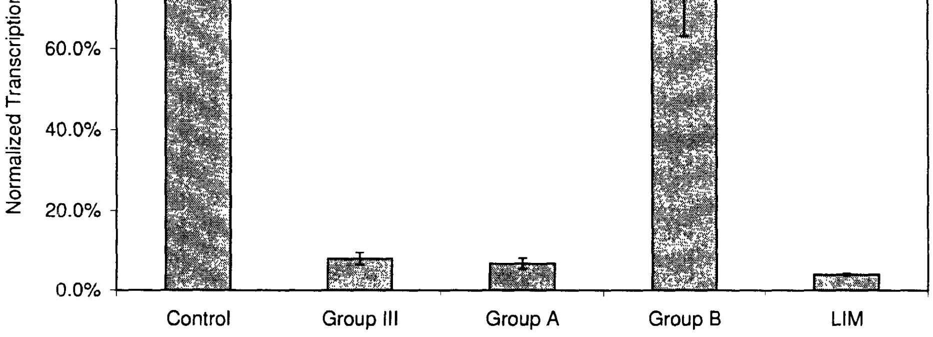


Figure 5-11: RT–PCR results for *LIM* expression in different hepatic treatments.

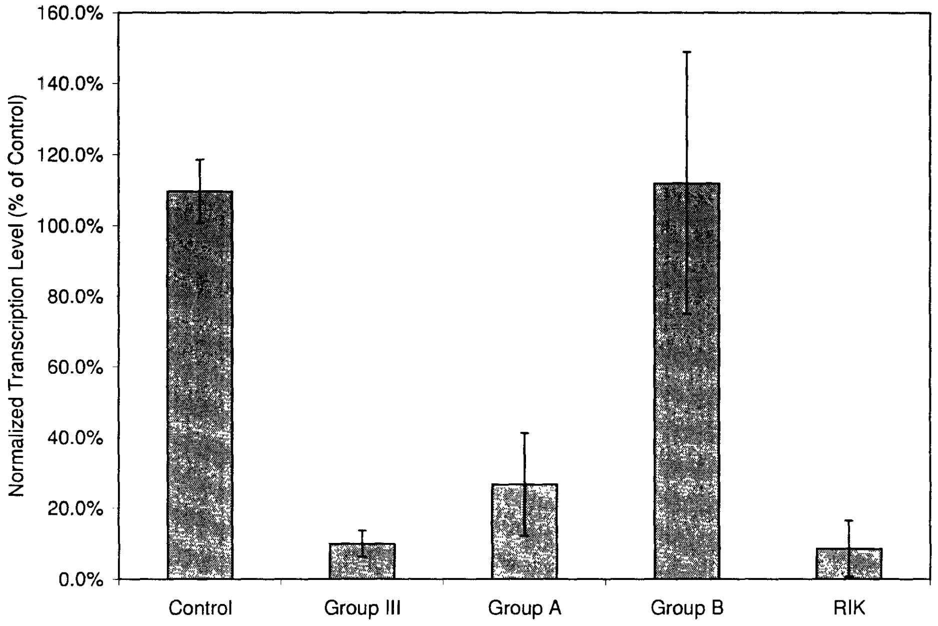


Figure 5-12: RT–PCR results for *RIK* expression in different hepatic treatments.

Figures 5-10, 5-11, 5-12, demonstrate effective, specific silencing was achieved in each of the siRNA treatments. Because Group A clusters close to Group III in Figure 5-9, and the individual genes of Group A cluster in a region near the samples treated with *G6P* and *PCK1* siRNAs, it appears that *Hgfl*, *LIM*, and *RIK* each contribute to the sample classification and individually affect hepatic metabolism during these treatments.

Figures 5-14 and 5-15 show that specific gene silencing was also attained for *Eva* and *LRR*. In this case, however, only *Eva* appears to have a direct effect on hepatic metabolism, because samples treated with *Eva* siRNA cluster near those treated with actinomycin, Group III siRNAs, and Group B siRNAs. In contrast, samples treated with *LRR* siRNA cluster with the high glucose producing control samples, thus indicating that hepatic metabolism was not significantly altered by silencing this gene. Because the *Eva* samples do not cluster directly with the Group B samples, it may be that in the presence of *LRR and Eva* silencing there is an interaction that leads to a further change in metabolism. Also, it should be noted that in Figure 5-

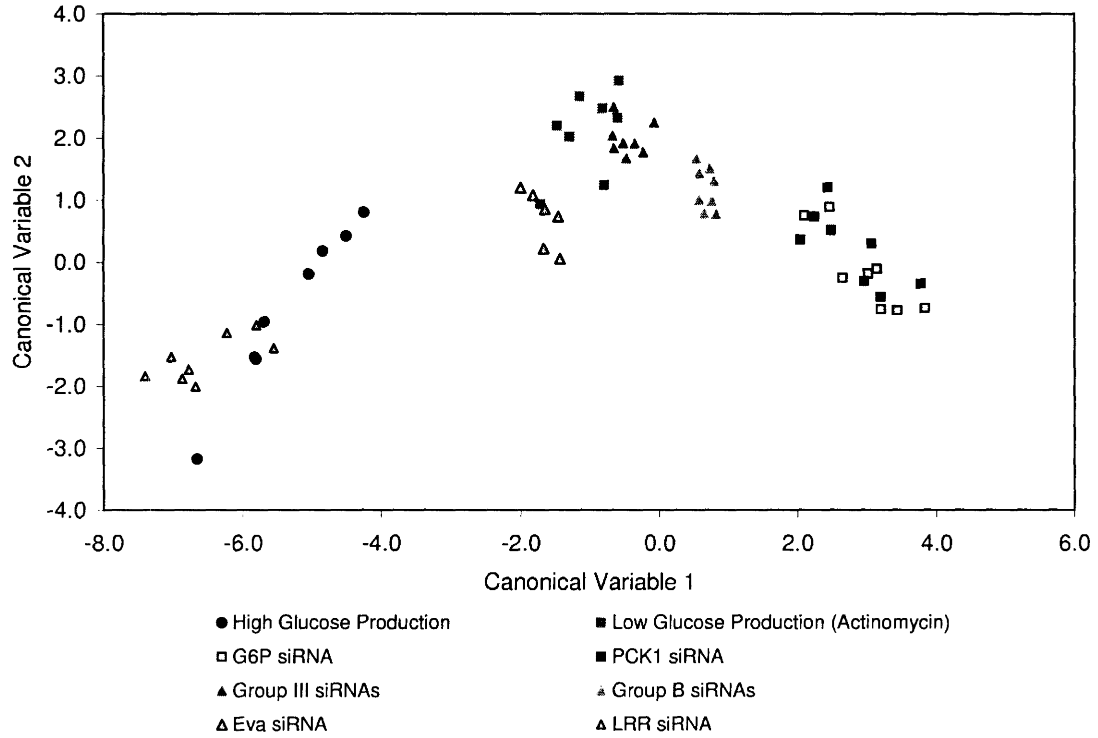


Figure 5-13: Fisher discriminant analysis of Group B siRNAs.

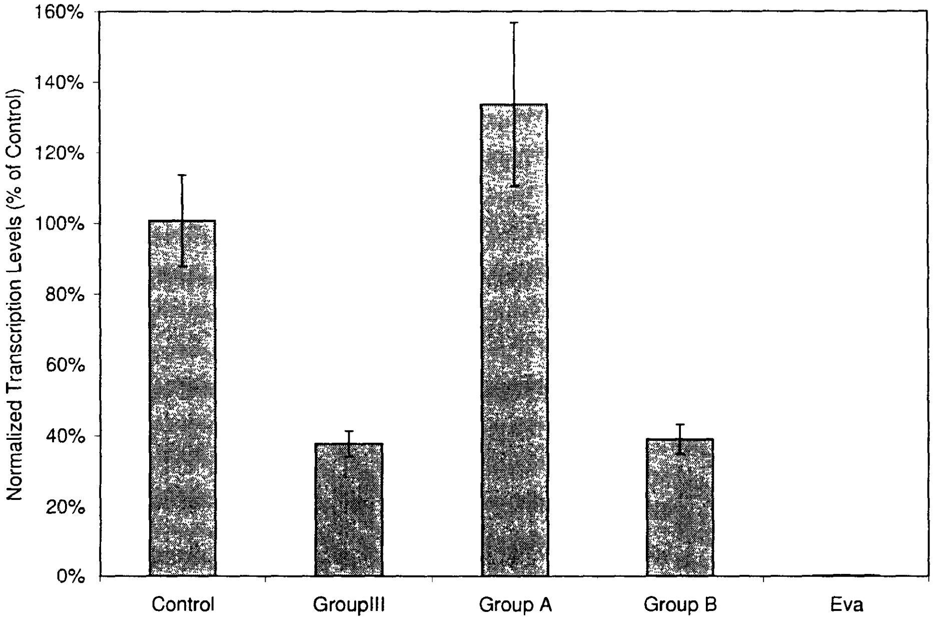


Figure 5-14: RT–PCR results for *Eva* expression in different hepatic treatments.

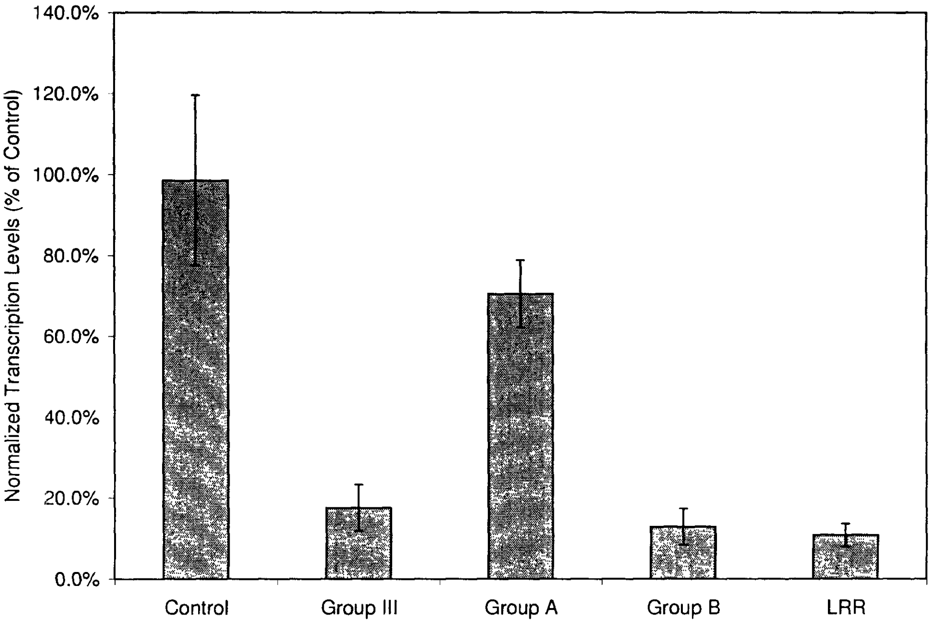


Figure 5-15: RT–PCR results for *LRR* expression in different hepatic treatments.

15 there may be some non–specific effects on *LRR* silencing from Group A siRNAs. Although this difference was not significant ($P > 0.05$), the mean value is decreased.

To determine whether the metabolite projections correlated with glucose production, cells were treated with each group of siRNAs, and each individual siRNA, and glucose production into the medium was measured. From this data, the reduction in hepatic glucose output was calculated and plotted in Figure 5-16. Figure 5-16 shows that the combined effect of all siRNAs in Group III results in approximately a 10% reduction in hepatic glucose output, comparable to the effect of silencing *G6P* and *PCK1*. Likewise, each subgroup and individual gene has a lessor effect and these are approximately additive. The effect of *LRR* silencing did not significantly reduce hepatic glucose output, in agreement with the clustering of those samples with the control samples.

In addition to the FDA projections, the data set was also analyzed using PCA, a non–supervised method (see Section 2.3.2 on page 45 for more information). Figure 5-17 shows the projection obtained from principle components 2 and 3. This projection

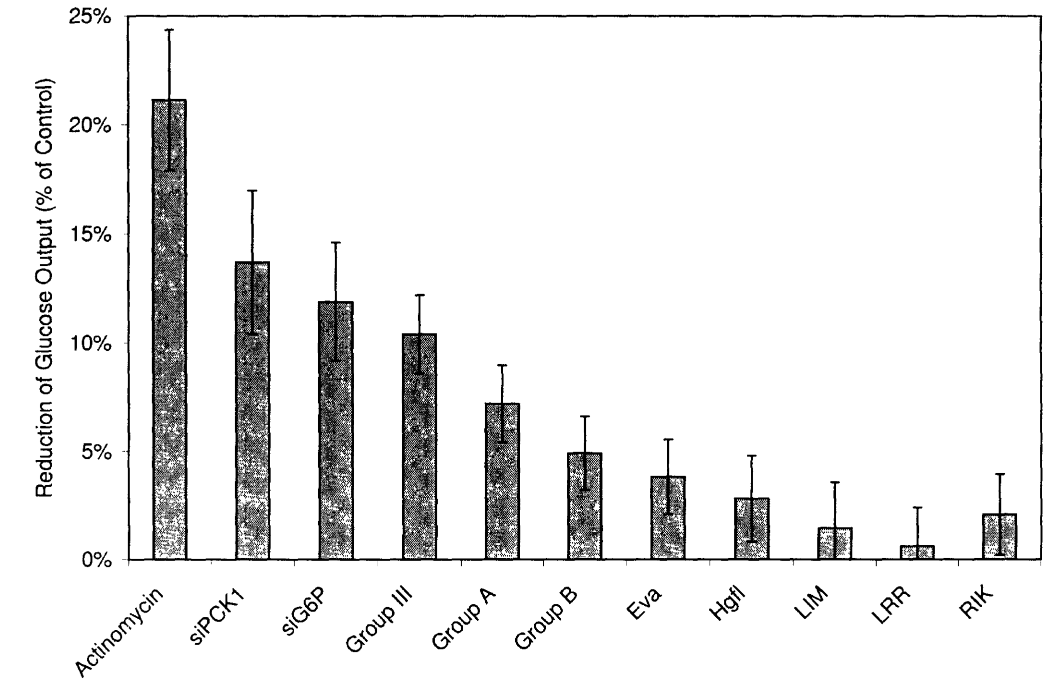


Figure 5-16: Reduction in hepatic glucose output from actinomycin and siRNA treatment.

clearly separates samples producing high amounts of glucose and those producing low amounts of glucose and the overall separation between the treatments is consistent with the FDA projections.

Based on these results, of the 15 genes in Table 5.2, only *Eva* (GenBank Accession # NM_007962, previously BC015076), *Hgfl* (GenBank Accession # NM_178149, previously AK005141), *LIM* (GenBank Accession # NM_024263, previously AK007076), and *RIK* (GenBank Accession # AK017674) significantly effect on hepatic metabolite levels, and of these, only *Eva*, *Hgfl*, and *RIK* cause a significant individual reduction in glucose output. When silenced in combination, the reduction in glucose production was approximately additive from these three genes. Conversely, *LRR* (GeneBank Accession # NM_026253) did not appear to have a substantial effect on metabolism, although it was differentially expressed in the liver samples from C57/BL/6J mice (see Chapter 4) and also did not have a significant effect on glucose production.

Eva (Epithelial V–like Antigen) is a putative transmembrane type 1 glycoprotein with an immunoglobulin V–type domain and although it has only been studied in the

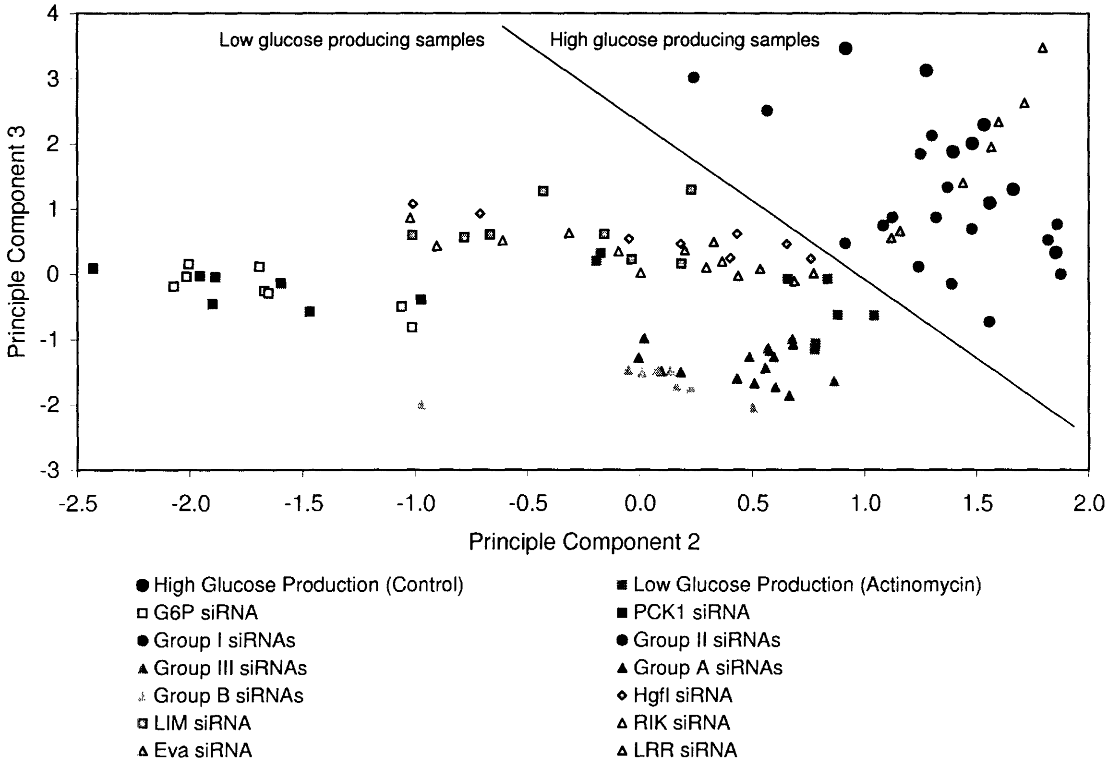


Figure 5-17: PCA projection of hepatic metabolite distributions for different experimental treatments.

context of thymus development, its expression in liver tissue has been demonstrated previously [91]. Because increased expression of *Eva* was observed in DIO–C57/BL/6J mice and then renormalized in DIO–C57/BL/6J mice fasted for 48 hours, and that siRNA silencing of *Eva* lead to a significant change in hepatic metabolite levels, suggests that further studies in hepatocytes are warranted to understand its role in hepatic physiology.

Hgfl (Hepatocyte Growth Factor–Like; also called macrophage stimulating 1) encodes a highly conserved factor found in many species including *Mus musculus*, *Homo sapiens*, *Rattus norvegicus*, *Gallus gallus*, and *Canis familiaris*. *Hgfl* has been widely studied in a variety of contexts, probably because of its sequence and structural similarity to Hepatocyte Growth Factor (HGF)[1] and high–levels of expression. Expression of *Hgfl* is controlled by Hepatocyte Nuclear Factor–4 (HNF–4) and coactivated by CREB–binding protein (CBP) [169, 265]. The appearance of HNF–4[2], as a regulator or *Hgfl* is interesting because several other genes in Table 4.2 are either directly influenced by HNF–4 (such as *Ttr* [43]) or belong to a gene class for which HNF–4 regulates other members of the class (such as *Cyp2c37* [12, 194] and *Serpina5* [205]), suggesting that HNF–4 may have played an important role in the differential expression observed in these genes. The Hgfl protein contains a conserved Kringle Domain, which are generally believed to play a role in protein binding, and a nonfunctional proteolytic serine protease–like domain. It binds the murine receptor stk, which belongs to the tyrosine kinase family or receptors (similar to the insulin receptor) [142], and induces complex pleiotropic effects, some of which are implicated in inflammation. Indeed, *Hgfl* knock–out mice have even been constructed and studied [16]. These mice develop normally and have no obvious unchallenged phenotype other than the formation of large lipid vacuoles in the cytoplasm of hepatocytes. The lipid vacuoles have no effect on hepatic production of serum albumin, aminotransferases, bilirubin, gamma–glutamyltranspeptidase, or alkaline phosphatase. Personal communication

[1] Hepatocyte Growth Factor has been shown to be essential for embryogenesis [246], as well as useful in prolonging β–cell life and cell mass [45, 80]. HGF has also been implicated in protecting against diabetic neuropathy in rats [127] and protecting against diabetic nephropathy [46] in mice.

[2] HNF–4 mutations cause Maturity Onset Diabetes of the Young 1 (MODY1), a rare syndrome that resembles Type II diabetes but is caused by single gene mutations [52].

with the lead author who constructed the *Hgfl* knock–out mice indicated that glucose homeostasis and hepatic glucose output had not been studied in these mice [17].

LIM (Limitrin), like *Eva*, is a putative type I transmembrane protein with two immunoglobulin V–type domains. Members of this protein superfamily often mediate either homophilic or heterophilic cell adhesion interactions, but also serve as receptors that facilitate interactions between the cell and extracellular matrix. Thus far *Lim* has only been studied in the context of neurobiology, where it is suggested to be involved with brain maintenance and maturation [272]. LIM's sequence is identical to asp3, an adipocyte–specific protein that is up–regulated during the differentiation of the 3T3–L1 cell line.

It is interesting that *Eva* and *LIM*, two genes from the same class of molecules, were identified in these studies. Although *Eva* was induced by high–fat feeding and normalized by 48 hours of fasting and weight reduction, *LIM* was increased by high–fat feeding and even more so by 48 hours of fasting and weight reduction. Thus despite similar regulation during high–fat feeding, their response to fasting and weight reduction differs. Although members of the transmembrane immunoglobulin superfamily are not commonly associated with metabolic changes, they are associated with diabetic complications and inflammation processes, to which *Eva* and *LIM* may potentially contribute.

RIK (or *5730458M16Rik*) is an unclassified gene identified in an EST sequencing project [30, 223]. The top two genes with which it has a high degree of similarity, as determined by BLASTn are *Mettl3* (expectation score $2x10^{-86}$ using default BLASTn parameters) and *Flot1* (expectation score $2x10^{-83}$ using default BLASTn parameters). *Mettl3* is a N6–adenosine–methyltransferase, which is involved in the methylation of eukaryotic mRNA on adenosine residues. Although it is believed that adenosine methylation performs a regulatory role in mRNA maturation and translation, the mechanism and its consequences have not yet been determined [21]. In contrast, *Flot1* is caveolae–associated integral membrane protein involved in vesicle trafficking and signal transduction [18]. Neither gene has been directly implicated in Type II diabetes or metabolism, and the connection to *RIK* is tenuous.

It should be noted that because the screening strategy only measured effects on intracellular hepatic metabolites and did not explicitly measure the level of gene silencing in every case. Therefore, it is possible that the degree of silencing for some genes in Group I and Group II, that are important, was not high enough to result in a measurable effect. Likewise, the results are specific to our model system, which displays some physiological characteristics that hepatocytes possess *in vivo*, but may be missing other characteristics. Thus some genes may have their *in vivo* effects hidden in our cell model, but may still be important to controlling hepatic glucose output.

Despite the experimental caveats, our approach to studying the genetics of hepatic glucose output, beginning in Chapter 4 with transcription monitoring using genome scale DNA microarrays, proceeding with a bioinformatics analysis and focusing on 15 genes in silencing experiments, has allowed us to identify four genes that influence hepatic metabolism out of an initial set of over 17,000. This strategy and approach is a good model for future investigations as it enables efficient and meaningful gene identification based upon no *a priori* assumptions about the system or data.

5.5 Methods

Isolation and culture of C57/BL/6J mouse primary hepatocytes

Hepatocytes from C57/BL/6J mice were isolated as described previously [219, 190] and seeded onto Type I collagen–coated T25 tissue culture flasks (BD Biosciences, Bedford, MA) at 4.0 million cells/well in 4.0 ml of Hepatocyte Medium Base supplemented with 1 nM insulin, 100 nM dexamethasone, and 20 mM glucose. Hepatocyte Medium Base is composed of DMEM powder (Sigma) was supplemented with 3.7 g/L $NaHCO_3$, 30 mg/L proline, 100 mg/L ornithine, 610 mg/L niacinimide, 0.544 mg/L $ZnCl_2$, 0.75 mg/L $ZnSO_4$ 7 H_2O, 0.2 mg/L $CuSO_4$ 5 H_2O, 0.025 mg/L $MnSO_4$, 146 mg/L glutamine, 2 g/L bovine serum albumin, 100,000 units/L penicillin, and

100,000 mg/L streptomycin. The medium was sterilized by filtration through a 0.22 μm filter and stored at -20 °C. The cells were allowed to attach for one hour at 37 °C in a humidified atmosphere containing 5% CO_2. After the attachment period, cells were washed with PBS and incubated in 2.6 mL of medium (Hepatocyte Medium Base supplemented with 1 nM insulin and 100 nM dexamethasone) formulated with the apppropriate siRNAs.

Transfection of siRNA

siRNA transfection was carried out as reported previously [202] (See Section 3.3.3 on page 86 for further details). Briefly, the corresponding siRNA (Ambion, siRNAs are listed in Table 5.2) was formulated with Lipofectamine2000 (Invitrogen) in Hepatocyte Medium Base supplemented with 1 nM insulin and 100 nM dexamethasone as described above for the appropriate experiment. Transfection exposure times were at least four hours in every treatment. Metabolites and RNA were isolated from the cells 24 hours after transfection.

RT–PCR of selected genes

A two-step RT-PCR protocol was performed to confirm the mRNA levels of *Eva*, *Hgfl*, *LIM*, *LRR*, and *RIK*. cDNA synthesis was performed as detailed previously [28] (See Section 4.4). PCR was conducted in 94-well plates using the iQ SYBR Green Supermix Kit (Bio–Rad), according to the manufacturer's instructions on an iCycler RT-PCR machine (Bio-Rad). Briefly, 1 μL of the final, diluted cDNA template was mixed with 19 μL of RNase free water, 25 μL of Bio-Rad RT-PCR Supermix (Bio-Rad), 2 μL of sense and antisense primers, and 1 μL of 12.5 mM dNTPs. The final primer concentration was 0.25 μM. The PCR cycle used a single three minute hot-start at 95 °C, followed by 50 cycles of 30 seconds at 95 °C, one minute at 60 °C, and two minutes at 72 °C during which time the reaction fluorescence was measured. RNA from treatment sample was measured in at least quadruplicate.

Eva, *Hgfl*, and *LIM* were measured using primers designed for the Quantitect

RT–PCR kit(Qiagen) according to the above protocol. The specificity of these reactions were checked by gel electrophoresis and gene specific standard curves were generated that resulted in an R^2 value of greater than 0.97 in every set of reactions. For the other genes monitored, the sense and antisense primer sequences were: *G6P* 5'- GTGATTGCTGACCTGAGGAACG - 3', and 5' - TGCCACCCAGAGGAGATTGATG - 3'; for *PCK1* 5'- CAGAGAGACACAGTGCCCATCC - 3', and 5' - AAGTCCTCTTCCGACATCCAGC - 3'; for *LRR* 5'- ATCACATTTGATGGGAGAAAACGCC - 3', and 5' - GCAAGATACACTTGGGGAAGGTGGT - 3'; and for *RIK* 5' - TGTGGTTGCTGGGACTTGAACTTCA - 3', and 5' - TGGTGAGATGGCTCAGTGGGTAAGA - 3', respectively. It's noteworthy that Quantitect primers for *LRR* and *RIK* were also ordered, however the primers obtained did not result in specific, reproducible sequences. New primer sets were designed, from which primers were selected that gave specific fragments of the correct length when viewed upon a 4% agarose gel (data not shown). As an internal control β-Actin mRNA levels were also measured. The sense and antisense sequences were 5' - AATAAGTGGTTACAGGAAGTC - 3' and 5' - ATGAAGTATTAAGGCGGAAG - 3', respectively.

Gene specific standards were developed for each gene by PCR from a cDNA library, gel purifying the resulting band, and then diluting it to concentrations from 10^{-4} μg/ μL to 10^{-11} μg/ μL. The R^2 value of the standard curve, relating the threshold cycle to the amount of standard template, was always greater than 0.97. The mRNA levels of β-actin measured were not significantly ($p > 0.05$) different between the treatments for any of the groups.

Metabolite isolation and profiling

Metabolite isolation and profiling was carried out using methods described previously [72]. Briefly, cultured hepatocytes were lysed using 0.7 mL of methanol (Sigma) per T25 flask and allowed to incubate for 15 minutes. During the incubation 4 μg of ribitol were added to each flask to serve as an internal control. The complete sample was then transferred to a polypropylene 15 mL tube (Falcon) and 0.7 mL of sterile water (Ambion) and 0.38 mL of chloroform (Sigma) were added. The samples were mixed vigorously and then centrifuged for three minutes at 3200 × g. Following centrifugation, 1.3 mL of the aqueous phase was transferred to a 1.5 mL microcentrifuge tube (Eppendorf). Samples were subsequently dried overnight in a vacufuge (Eppendorf). The dried samples were resuspended in 50 μL of methoxyamine hydrochloride (20 mg/ mL pyridine) and the liquid was transferred to a glass vial and incubated at 30 °C for 90 minutes. Following the incubation the samples were derivitized using 80 μL of MSTFA + 1% TMCS (Pierce) and incubated for 30 minutes at 37 °C. The sample was finally transferred to vials compatible with the mass spectrometer autosampler and loaded onto the instrument for injection.

The resulting spectrum from each sample was then analyzed for 3–phosphoglycerate, α–ketoglutarate, citrate, fructose–6–phosphate, fumarate, glyceraldehyde–3–phosphate, glucose–6–phosphate, malate, phosphoenolpyruvate, pyruvate, ribose–5–phosphate, ribitol, and succinate.

Glucose measurements

The amount of hepatic glucose output was measured by sampling of the culture medium 24 hours following transfection (or medium change). The measurements were made using a YSI Glucose/ Lactate analyzer.

Data analysis

Metabolite concentration data were normalized to the ribitol internal control. The data were then assembled into a single data matrix, where each row represented a sample and each column a metabolite. The matrix was autoscaled by subtracting the mean of each column vector from each sample and dividing the difference by the standard deviation of the column vector. The data was then analyzed using BioSystAnse [19] for the FDA projections and by using `Matlab` for the PCA projections (see Section 2.3.2 on page 45 for more information regarding these techniques).

Chapter 6

Summary and Significance of Work

6.1 Summary of Thesis Results

The spread of diabetes is an important problem throughout the world and the subject of intense research. Because blood glucose control is mediated through a number of complex systems, whose molecular basis is not completely understood, further insight can be gained by identifying genes that contribute to its regulation. Ultimately glycemic regulation is a balance between glucose up-take (primarily by the muscle and adipose tissue) and glucose production (primarily by the liver). This thesis focused on identifying genes that influence hepatic metabolism and glucose output.

Traditionally researchers have used gene mapping and linkage studies to determine genes involved in a particular phenotype. This approach has been successful for some single gene disorders, however, it is more difficult to employ in polygenic traits, such as hepatic glucose output. For this reason we used genome scale DNA microarrays, containing 17,000 gene probes, to monitor hepatic gene transcription in control mice, mice with diet induced obesity and insulin resistance, and mice with diet induced obesity that had been calorically restricted, returning their weight to control levels. This approach is fundamentally different than mapping studies that search for *mutations* in genes associated with the phenotype. Instead, we searched for genes that are differentially expressed under the experimental conditions and therefore may directly affect changes in phenotype, such as the development of insulin resistance. It is important to understand that these genes may or may not have mutations associated with the disease, however, because they are differentially expressed, they are likely to be related to the molecular mechanisms that define the physiological differences resulting from the experimental treatments.

Our microarray studies identified 41 genes that were differentially expressed and could rigorously classify the samples. These genes were found using t–tests, Wilks–λ based ranking, Fisher Discriminant Analysis (FDA) and Principle Components Analysis (PCA). Because of the successful treatment classification based on expression levels of the 41 identified genes, they represent good candidates that contribute to the observed changes in hepatic physiology.

In our work we were primarily interested in genes that were involved with hepatic metabolism, particularly hepatic glucose output (HGO). Because the liver performs many diverse physiological functions, such as glucose production, lipid production and metabolism, serum protein production, and xenobiotic detoxification, more screening was necessary to determine which genes affected hepatic metabolism. Traditionally studying loss of function phenotypes has been one of the most important ways of determining a gene's role *in vivo*, thus we developed RNA interference (RNAi) based gene silencing techniques to further screen the identified genes to determine their potential metabolic roles.

RNAi is a method that can be used to post–transcriptionally silence genes; that is, it directs the specific degradation of RNAs resulting gene transcription. By degrading the mRNA of candidate genes, the RNAi pathway prevents translation into protein products or any direct activities that the RNA molecules may possess. To efficiently screen the 15 over–expressed genes from our mouse studies, we developed combinatorial gene silencing protocols that utilized short–interfering RNAs (siRNAs). Using the combinatorial approach it was possible to determine the effects of 15 genes, mapping down to a single gene effect, within seven experiments.

Metabolite profiling was employed to quantitatively describe the metabolic state of hepatocytes during the silencing experiments. Metabolites were chosen from central carbon metabolism (glycolytic and gluconeogenic pathways), the TCA cycle, and pentose phosphate pathway to represent the physiological state. Control treatments that induced different levels of HGO were analyzed with Fisher Discriminant Analysis (FDA) to generate lower dimensional metrics that represented the information from the metabolite profiles. The two metrics developed (the canonical variables resulting from the FDA analysis) could accurately classify the control treatments and were used to project experimental treatments in the reduced FDA space. Based upon the metabolite dependent classification of the experimental siRNA treatments, we focused on one group of five genes that clustered within the lowest glucose producing control samples. The genes contained within this group were *Eva, Hgfl, LIM, LRR*, and *RIK*.

Potent gene silencing of all five genes simultaneously resulted in an 11% reduction in HGO, which was similar to the results of silencing *PCK1* or *G6P*, two key gluconeogenic genes. When the five genes were silenced in smaller groups, or individually, their classification changed within the FDA space and their effects on HGO decreased. Of the five genes, four (*Eva, Hgfl, LIM, RIK*) had a significant effect on hepatic metabolism. Of these four genes, only *Eva, Hgfl* and *RIK* reduced HGO significantly, each by about 5%.

Although each of these genes individually contribute a small amount to HGO, they would have been very difficult, if not impossible, to identify using traditional mapping techniques. It is not currently known whether these genes have mutations that are associated with disease, however, it is now known that in primary hepatocytes, they individually and collectively can influence HGO. For this reason, *Eva, Hgfl,* and *RIK* provide important new research targets and further insight into a very complex phenotype.

6.2 Significance of Results

While classical gene identification techniques work for well defined, single gene phenotypes, more complex, multigenic phenotypes can benefit from other approaches to gene identification. Complex diseases, such as type II diabetes, require additional information to rapidly identify genes involved in the defining molecular pathways. This work integrated experimental techniques including DNA microarrays, RT–PCR, RNAi based gene silencing, and metabolite profiling, along with multivariate analysis techniques, to study glucose production in primary hepatocytes. In doing so we were able to efficiently identify genes that affected hepatic metabolism and HGO.

While the methods used here are not unique, our application provided a new process for identifying genes involved in complex phenotypes. The high through–put nature of DNA microarrays provides enormous amounts of data, driving the identification of candidate genes through statistical and multivariate analysis. In this work we were interested in hepatic genes that mediate the metabolic changes that occur when C57/BL/6J mice are placed on a high–fat diet, or when obese mice are calorically restricted. By analyzing the DNA microarray data, we discovered 41 genes that could discriminate the dietary treatments. Each of these genes represents a rigorous candidate for future investigations. Such applications represent a growing shift from the single hypothesis testing that biologists traditionally rely upon, to an expanding landscape of analyses, which not only investigate large numbers of individual components, but also the systemic features of their interactions.

Instead of conducting linkage studies to associate a chromosomal region with a phenotype, DNA microarrys enable studies work directly at the molecular level. Thus instead of genotyping populations to determine regions that associate with disease, and then dissecting the region for the genetic components that affect a given phenotype, genes can now be readily linked directly to a phenotype of interest. This opportunity also presents some serious challenges associated with understanding how large numbers of variables interact, interpreting true signals within a "noisy" environment, and often performing experiments in information limited systems. These circumstances have been encountered in engineering before and are therefore amenable to some of the analysis techniques that have been previously developed.

Certainly there are caveats to pursuing such lines of research and the resulting data may be confusing to interpret. However, high–through–put studies markedly overcome the time constraints and tedious work involved in identifying QTLs, can be performed with limited staff, and usually leave no shortage of interesting genes for further investigation. Furthermore these types of investigations provide an important conceptual shift in studying biology; one that moves away from the link between genetic alleles and phenotype to one that focuses on the *mechanism* of the phenotype. That is, our approach has not analyzed how the distribution of alleles influence a specific phenotype and instead focuses on how differences in gene transcription define the phenotype. In this way we can more rapidly develop models that explain our observations, look for relevant mutations if need be, and use the results to help guide the development of improved therapies.

While stringent analysis of microarray data can be used to find genes whose transcription levels are associated with a phenotype, directly testing the genes for specific physiological effects still proves challenging. To more rapidly sort through the gene candidates, equally high–through–put methods of gene characterization and further relevance screening must be developed.

Unlike prokaryotes and yeast, genetic manipulation of mammalian cells is very tedious, difficult to perform, and time consuming. With the discovery of RNA interference (RNAi), this situation changed. For the first time it has become possible to efficiently study the effects of specific loss of function phenotypes on genes of interest. Although still in its infancy, the cost, through–put, and required optimization of RNAi based gene silencing promises to decrease in the future, thereby providing a complementary gene "characterization" technology to DNA microarrays as a gene discovery technology.

The RNAi protocols we developed were used to study how silencing of genes identified in our mouse studies influenced glucose production from primary hepatocytes. Because our assay for HGO was time consuming, measuring glucose production over a 24–hour period, and provided little mechanistic insight on the intracellular environment, we used metabolite profiling as the basis for quantitatively assessing hepatic metabolism. The metabolite profiles from control treatments that induced different HGO rates were analyzed using Fisher Discriminant Analysis (FDA) and could be used to classify the treatments. The classification metrics derived from the FDA analysis allowed us to further delineate the changes in hepatic metabolism caused by gene silencing.

In order to increase the efficiency of our silencing experiments, we developed combinatorial screening using siRNAs. The combinatorial strategy allowed us to find four genes out of an initial set of 15, that had a significant effect on hepatic metabolism in seven experiments. In addition to decreasing the number of experiments required, it also showed that multiple genes could be simultaneously silenced, revealing (non)additive effects of silencing gene combinations. This provides a method for testing individual genes and determining if they have a complementary or antagonistic impact a phenotype. In our study four of the genes (*Eva, Hgfl, LIM, RIK*) had an influence on metabolism in primary hepatocytes, but only three (*Eva, Hgfl, RIK*) decreased hepatic glucose output, and those decreases were approximately additive.

Eva, *Hgfl*, *LIM*, and *RIK* may have been difficult to discover using other experimental methods, given that they are not well characterized and intuitively are distantly linked to HGO. While the putative roles of these genes are not closely related to metabolism, our studies support the assertion that they may be involved and could potentially contribute to the variance in HGO observed in Type II diabetics. Although this assertion needs to be further to be validated and studied in greater detail, the new direction resulting from this research is an important contribution: it adds a new piece to the puzzle of HGO regulation. Determining the *in vivo* effects of these genes on hepatic glucose production in a mouse model would provide an important validation, or caveat, of this work. Because biological science is very serendipitous, with a large element of unpredictability, such discoveries deserve to be explored based upon their merit.

While we focused on silencing genes that were over–expressed during the dietary treatments in our mouse experiments, over half of the genes identified from the microarrays were repressed. Many of these genes encoded signalling proteins that may also be involved in the development of insulin resistance. Thus future work should look at over–expression of the repressed genes during high–glucose producing conditions. Determining the effects of these repressed genes on metabolism and HGO is a complementary approach that we did not explore. Our approach also could be extended to study glucose up–take by muscle and adipose tissues, insulin secretion by β–cells or other interesting phenotypes.

The ability to integrate new technologies with advanced analysis methods and readily employ those to medical research has the potential to make significant future contributions. The strategy used here to identify candidate genes is a good case study upon which other investigations may be based. Our approach certainly lends itself to the investigation of other relevant phenotypes and should find growing applications in the future.

Appendix A

Partial Least Squares Simulations

Partial least squares (PLS) is a linear regression method that creates correlations between multivariate sets of data. For this reason it can be a valuable tool for empirical modeling of biological systems, which often have multiple relevant input and output variables of interest.

In this thesis, PLS was used to analyze and derive potential relations between a dependent data set, $\mathbf{Y}$, and the underlying gene transcription data, $\mathbf{X}$. In this way, we have attempted to identify the genes that regulate or influence some biologically relevant features of the cellular system.

Generally speaking, transcriptional data sets have vastly more genes, g, than samples, s, and are referred to as "underspecified" (that is, $s << g$). The converse situation may arise if a predefined gene set were used such that the number of samples were equal to, or greater than the number of genes in the model; such a data set, with $s > g$, is referred to as "overspecified."

To test the ability of PLS to create models that cannot be produced by random occurrence, simulations were conducted with the aim of determining how the information structure of the input data affected the resulting model. The input data, $\mathbf{X}$, to the PLS algorithm has three primary components that define its structure and content: the number of samples, s, the number of "relevant" genes, g, and the number of "irrelevant" genes, ng (the "noisy genes")[1]. Within this framework, several

[1] By relevant, we refer to genes that have some affect or contribution to the output data contained

immediate questions were investigated. First, in DNA microarray experiements it is most common for the number of genes in our data set to grossly outnumber the number of samples, that is, the case of the underspecified data matrix. How is PLS's ability to create models that predict the Y-Block affected by the number of relevant and irrelevant genes? Second, it is important to know how sensitive PLS is to the number of samples contained. Is there an optimal ratio of genes to samples? Does the distribution of relevant and irrelevant genes affect the required number of samples?

A.1 PLS Gene Simulations

The first set of simulations were designed to test the effect of the number of irrelevant genes on the PLS model prediction. This is an important consideration for studying quantitative, polygenic traits, particularly if DNA microarrays are going to be used as the basis for gene identification. In this case, the number of genes monitored in an experiment is almost certainly much greater than the number of genes that actual regulate a given trait, and therefore understanding how the added genes may impact correlations derived from PLS is an important consideration.

Gene simulations investigated under what conditions the PLS model predictions, based on either actual or random data, converged as a function of the number of irrelevant genes. In describing these simulations, "actual data" refers to data that was used in a linear model to construct the dependent data matrix, **Y**. Thus in every simulation, **Y** is some linear function of the number of relevant genes, whose values comprise the actual data, as defined by

$$\mathbf{Y} = \mathbf{X}\ \mathbf{b} \tag{A.1}$$

where **X** is the randomly generated X-Block of actual data, and **b** is the randomly generated correlation vector. To the X-Block, irrelevant genes are added that are simply random data placed within the data matrix, and *do not contribute* to the

in the Y-Block. By irrelevant, we refer to genes that do not substantially contribute to the data contained in the Y-Block.

values of the dependent data matrix, $\mathbf{Y}$. As $\mathbf{X}$ becomes larger by the incorporation of more irrelevant genes, the elements of $\mathbf{b}$ that correspond to the additional irrelevant genes are given values of 0, so that in any given simulation the number of relevant genes remains constant. Once $\mathbf{X}$, $\mathbf{b}$, and $\mathbf{Y}$ are constructed, PLS is used to relate $\mathbf{X}$ to $\mathbf{Y}$, and to relate a randomly generated X-Block, to $\mathbf{Y}$. The result of the correlation between $\mathbf{X}$ and $\mathbf{Y}$ is called $\mathbf{Y}_p$, while the result of the randomly generated X-Block and $\mathbf{Y}$ is called $\mathbf{Y}_r$. Each test is repeated 100 times for each combination of s, g, and ng.

To evaluate how PLS performed in the simulation, the Euclidean distance between $\mathbf{Y}$ and $\mathbf{Y}_p$, and between $\mathbf{Y}$ and $\mathbf{Y}_r$ was calculated. Once the mean and standard deviation of the 100 test results were compiled for each metric, the test was restarted by adding another irrelevant gene. When the difference between the last ten simulation results of the actual data was not significantly different from the last ten simulation results of the random data (as measured by a two-way t-test, $P = 0.05$), the simulation ended and the number of irrelevant genes was recorded. Figure A-1 schematically shows how the first set of simulations was conducted.

The simulations began with a two gene, two sample, one Y-variable model, and proceeded by first increasing the number of irrelevant genes (ng), for different numbers of actual genes (g) used to make the Y-Block, and finally repeating the entire simulation by including more Y-variables, all of which were functions of the X-Block. Although the difference in the distance in Figure A-2is the largest when the X-Block is full rank, the variance in the model fits derived from random data is large and therefore the differences between the models is not statistically significant. Also the mean agreement between the prediction and the actual $\mathbf{Y}$ continues to improve as more irrelevant genes are added. This implies that the algorithm is overfitting, especially since the variance in the model is decreasing. Table A.1 summarizes the data from this set of simulations.

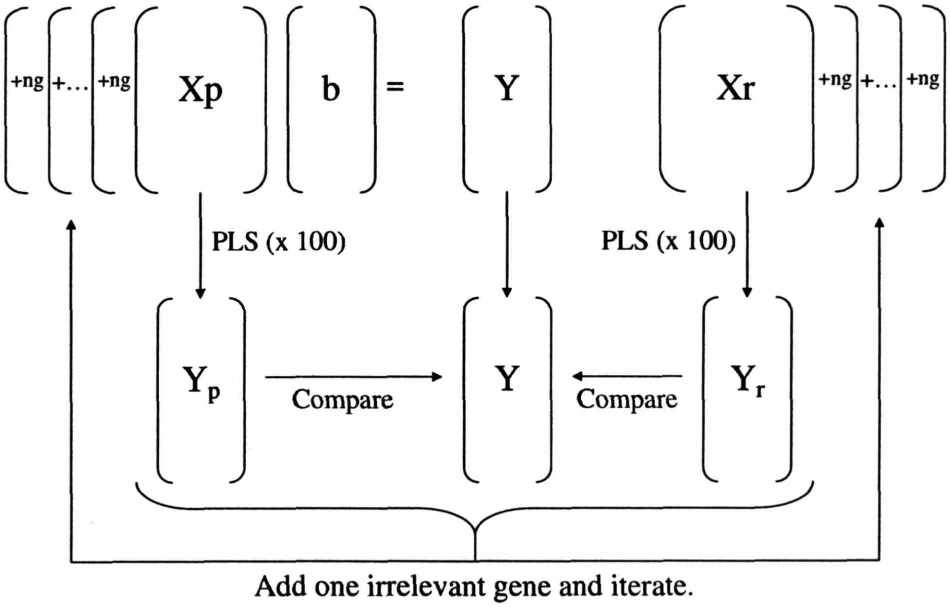


Figure A-1: Schematic of PLS simulation algorithm for testing the effect of additional irrelevant genes on model prediction.

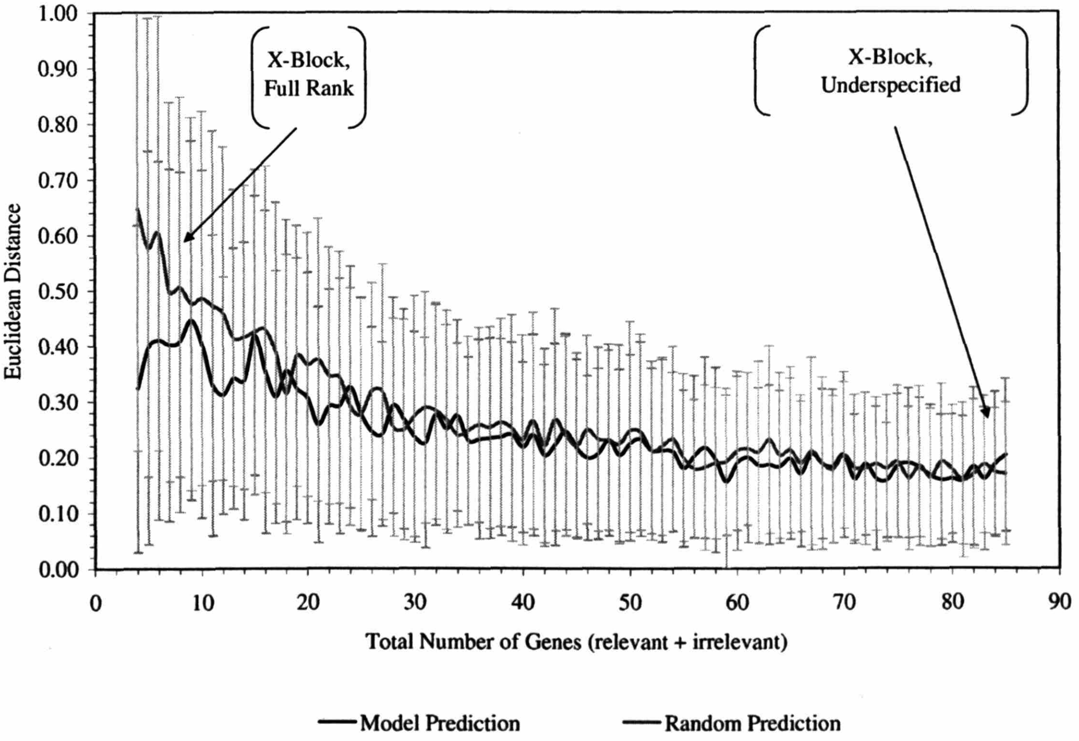


Figure A-2: PLS simulation results from adding additional irrelevant genes to the model.

Number of Initial Samples	Number of Acutal Genes	Number of Y–Variable	Number of Total Genes at Convergence	Number of Samples at Divergence	Initial Difference between Model and Random Prediction
2	2	1	83	20	0.324
2	10	1	83	20	0.093
2	20	1	110	19	0.073
2	30	1	104	19	-0.045
2	2	2	52	17	0.444
2	10	2	44	18	0.030
2	20	2	64	18	0.030
2	30	2	51	19	0.005
2	2	3	37	16	0.380
2	10	3	58	19	0.021
2	20	3	52	19	0.045
2	30	3	52	19	0.018
2	2	4	48	17	0.363
2	10	4	46	19	0.052
2	20	4	48	19	0.085
2	30	4	51	21	0.044

Table A.1: Summary of PLS Simulation Data.

In Table A.1 there are several things to notice. First the number of actual genes pertinent to the model does not markedly influence the number of total genes it takes for the models to converge. This means that very complex relationships, as found in polygenic quantitative traits, will be more difficult to substantiate because such relationships can easily arise simply from noise in the data set. Next, increasing the number of Y-variables in the model decreases the number of irrelevant genes necessary for convergence. This affect only differentiates the single Y-variable model from the others, and is the result of the Y-Block transitioning from being overspecified, to full rank, and finally to being underspecified.

Another interesting aspect of Table A.1 is the fact that as the number of actual genes in the model increases, the initial difference between the actual model prediction and the random model prediction decreases. This trend is independent of the number of Y-variables included in the model, which further substantiates the previous point that as the genetic model becomes more complex by including more genes, PLS will have an increasingly difficult time identifying its exact salient features as opposed to noise in the data.

Finally, the amount of variance captured from the X-block and Y-block in the first latent variable, did not vary substantially between the models. It's noteworthy that as the number of samples increases, the variance captured by the model spreads through the subsequent latent variables. Because different simulations have different numbers of latent variables included, only the amount of variance captured by the first variable was tracked. This variable captures the largest amount of variance from the Y-block of any of the latent variables. Because the algorithm is designed to both capture variance, and find correlation, it is not surprising that the random data model obtains a comparable level of variance in the first latent variable for both blocks.

Based on these initial simulations some conclusions can be made about employing PLS with different data sets:

- As the number of real genes pertinent to the Y-block data increases, the more difficult it is for PLS to identify a statistically significant model.
- As the number of irrelevant genes in the X-block increases, the more difficult it is for PLS to identify a statistically significant model.

A.2 PLS Sample Simulations

In addition to investigating the effect of the gene number on the model predictions, it is also important to understand the effect of the number of samples. For example, for a given number of genes in the model, is there a minimum number of samples necessary to find a statistically significant model? How sensitive is the model prediction to the number of samples? If a large number of genes are included, is it possible to do enough experiments to make PLS a worthwhile technique, or is it only relevant to look for correlations between subsets of genes? These questions are important because they have an impact on the practical applications of PLS and what kind of data is necessary for its implementation.

To test the effect of the number of samples on the model prediction, simulations were conducted in two ways, using the same aforementioned criteria. The first set of simulations simply continued the gene expansion simulations conducted in Section A.1 by adding more samples to the matrices and investigating how long it took the random and actual model to diverge. More samples were included to see if the model would perform the same "overfitting" that appeared to happen as more genes were added.

As shown in Figure A-3, as additional samples are added to the matrices, the random data model performs less and less well relative to the actual data model. Relative to the random data model, the model based on the actual data quickly yields better predictions of the Y-Block then what could be obtained by chance, making the model results statistically more significant.

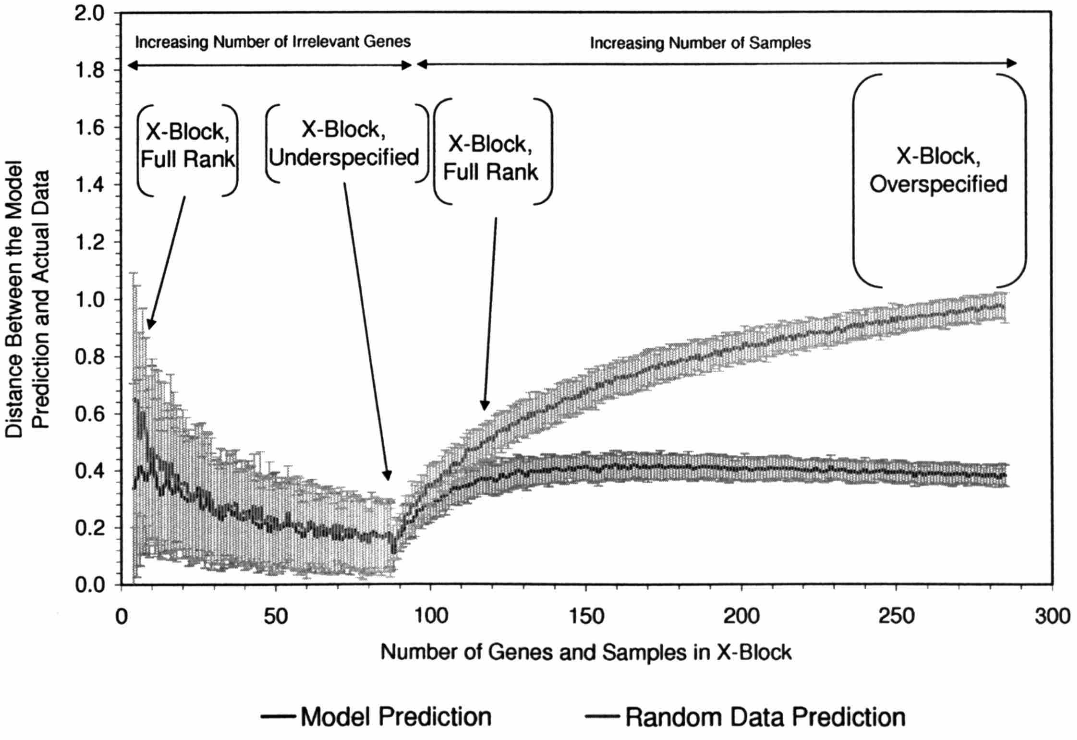


Figure A-3: PLS simulation results from adding additional samples to the data matrix.

These differences are reflected in the amount of variance captured by the first latent variable in the model. Figure A-4 shows the amount of variance captured from the X-Block and Y-Block from both models as a function of the number of samples.

In Figure A-4 we see that the Y-block variance captured by the random model drops off monotonically as more samples are added to the experiment. Interestingly, although the means for the X-block latent variables also decrease as more samples are added, the variance in the actual data model latent variable is much smaller than the random data model latent variable. Despite this increased variance, it appears that the decreasing accuracy of the random data model is primarily due to the algorithm's inability to capture the Y-Block variance with a random X-block as the number of samples increases.

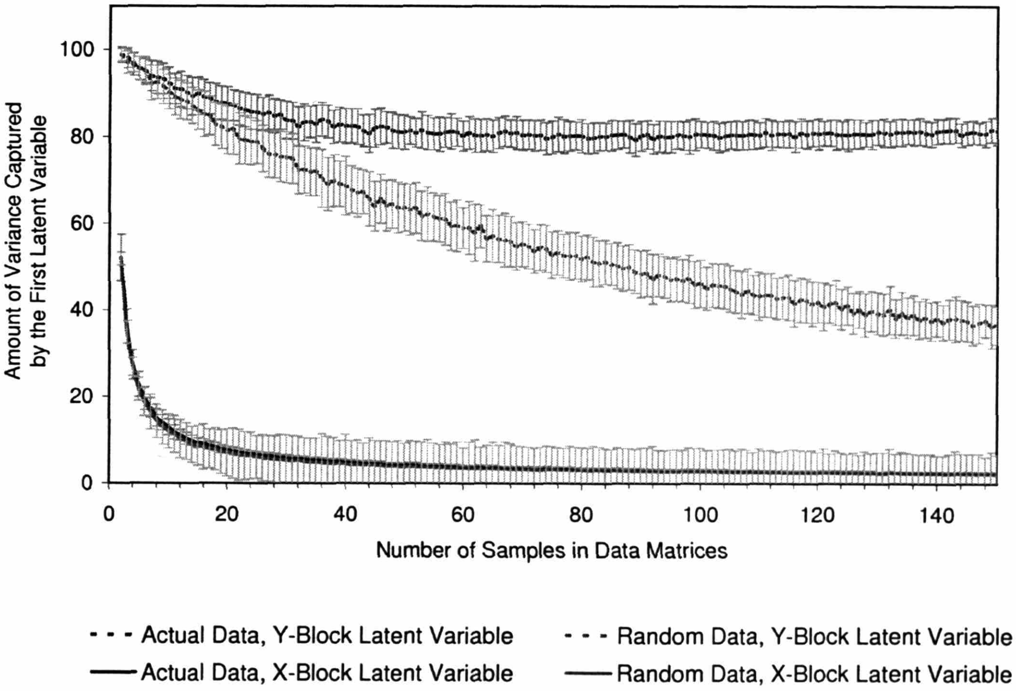


Figure A-4: Variance captured by the first latent variable of the Y-Block and X-Block in PLS model simulations.

The other method of simulation used to test the effect of increased numbers of samples, is to first increase the number of samples, with only relevant genes included in the model, and then see how many irrelevant genes are required for the model predictions to converge. This is similar to the simulation carried out above in Section A.1, however here the samples are first increased, followed by the number of irrelevant genes included in the data matrix. In this case, the maximum number of samples used in the simulation was 80, and the maximum number of irrelevant genes used was 800. Figure A-5 shows the results of the model simulations.

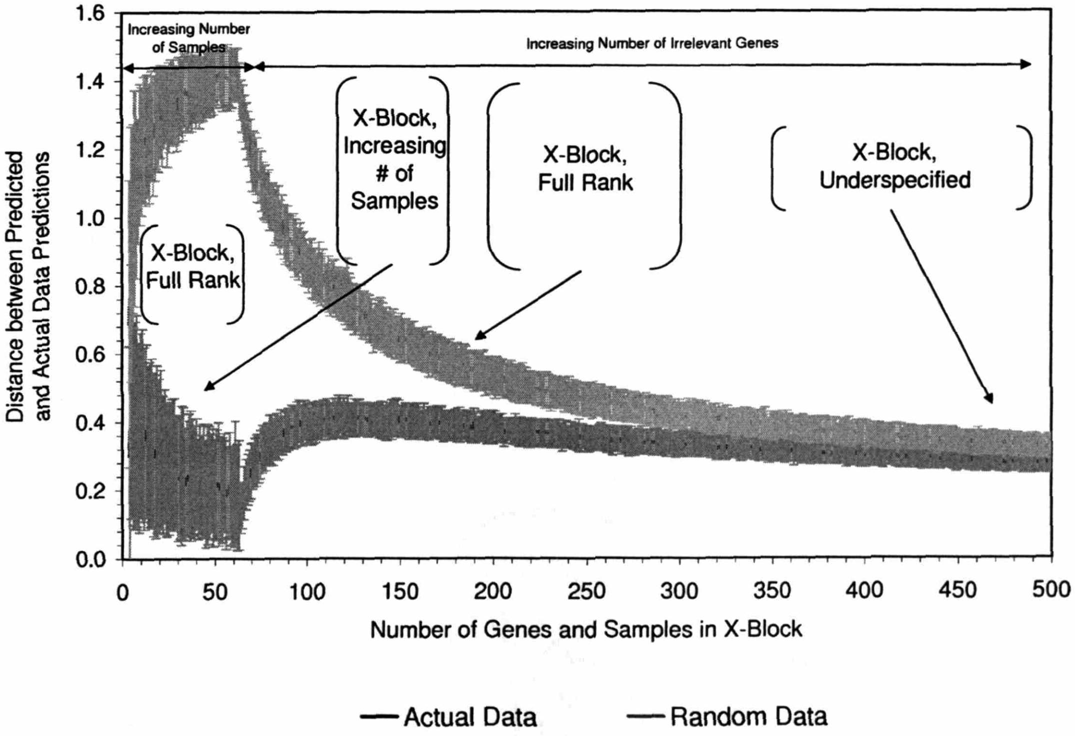


Figure A-5: PLS model simulation wherein the number of samples was initially increased, followed by the number of irrelevant genes.

This simulation took four days of computer time to complete, and after all 800 irrelevant genes had been included, the two models still had not completely converged (although the standard deviations of the models did overlap). It is interesting in Figure A-5 that the number of samples causes for a very large initial divergence between the actual and random data, and that the actual data prediction appears to get continually better as more samples are included. However, once the number of irrelevant genes is increased, the two models quickly begin to converge. Because the model initially does a better job of prediction as the number of samples is increasing, the amount of variance captured from the Y–Block is increasing. Likewise, as the number of irrelevant genes increases, the variance captured from the Y–Block decreases. Both of these effects are reflected in the amount of variance captured by the first latent variable during the simulations (data not shown).

Together, these simulations support the notion that in order to use PLS to identify statistically meaningful relationships in the data, having an adequate number of samples is crucial. While processing more samples on full genome microarrays is time consuming, costly, and can be tedious for large numbers of arrays, the relationships obtained should hold information that is both relevant and significant to the investigation.

Biology is currently attempting to address, and understand, the problems associated with the systemic features of phenotypic behavior. The potential of large scale biological experiments has yet to be completely realized partially because of the difficulty in obtaining useful information from these projects. This difficulty is enhanced by the absence of any clear method for linking, or integrating, different data types. Partial least squares, PLS, has been proposed as one method for identifying trends between different data types. A large number of simulations were undertaken to test the reliability and application of PLS to using an independent data set to predict a dependent data set. In our laboratory the independent data set usually takes on the identity of gene transcription data, while the predicted data set is usually chemical reaction flux data, or physiological data, however the algorithm itself can use any type of data matrices.

The results of the simulations have shown that for experiments typical in most DNA microarray studies, the resulting correlations are not significant unless a large number of samples are used. In one set of simulations, 20 samples were required to identify statistically significant models with 100 total genes, regardless of the number of actual genes contributing to the Y–Block. In practice, if the number of genes included in a set of experiments can be limited in a meaningful way, so as to decrease the relative number of irrelevant genes in the data set, then both statistically significant and unique relations can be identified which underlie the regulatory genetic structure.

It should be emphasized that prudent use of this and other methods is necessary, in addition to good experimental design. For many microarray experiments utilizing cDNA technology, the researcher generally loses genes as the number of samples grows (because as a requirement for inclusion in most models, the gene must be reliably detected in every sample), which complicates these problems. If genes relevant to the variables contained in the Y–Block are lost due to poor observability, then the resulting model will have a poor fit, and although the number of samples may rival the number of genes in the set, the resulting model will still not be significant due to its poor predictive capability. Conversely, if a large number of samples are not obtained, then the likelihood of having a statistically relevant model is low. Based on these problems, the best the researcher can do is identify relevant gene sets to the Y–Block variables (either from previous knowledge and intuition or from other methods such as discriminant analysis or statistics), process a lot of samples that span the space of values defined by the Y–Block variables, and conduct as many replicates as is feasible. If these three criterions can be met, then PLS can provide researchers the ability to link diverse sets of data in the study of systemic biological properties.

Bibliography

[1] Rexford S. Ahima, Daniel Prabakaran, Christos Mantzoros, Daqing Qu, Bradford Lowell, Eleftheria Maratos-Flier, and Jeffrey S. Flier. Role of leptin in the neuroendocrine response to fasting. *Nature*, 383:250–252, 1996.

[2] T.J. Aitman, A.M Glazier, C.A. Wallace, L.D. Cooper, P.J. Norsworthy, F.N. Wahid, K.M. Al-Majali, P.M. Trembling, C.J. Mann, C.C. Shoulders, D. Graf, E. St. Lezin, T.W. Kurtz, V. Kren, M. Pravenac, A. Ibrahimi, N.A. Abumrad, L.W. Stanton, and J. Scott. Identification of CD36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genetics*, 21:76–83, 1999.

[3] O. Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences - USA*, 97(18):10101–10106, 2000.

[4] P. Angel, M. Imagawa, R. Chiu, B. Stein, R.J. Imbra, H.J. Rahmsdorf, C. Jonat, P. Herrlich, and M. Karin. Phorbol ester-inducible genes contain a common *cis* element recognized by a TPA-modulated *trans*-acting factor. *Cell*, 49:729–739, 1987.

[5] M Appel, E Couderc, and J Feger. Comparative studies of protein biosynthesis: the main experimental parameters in pulse and pulse-chase experiments must be standardized. *Biol Cell*, 74(2):235–238, 1992. Letter.

[6] A. Arkin and J. Ross. Statistical construction of chemical–reaction mechanisms from measured time–series. *Journal of Physical Chemistry*, 99:970–979, 1995.

[7] V.Y Arshavsky and E.N. Pugh Jr. Lifetime regulation of G protein-effector complex: Immerging importance of RGS proteins. *Neuron*, 20:11–14, 1998.

[8] Ernest Asante-Appiah and Brian P. Kennedy. Protein tyrosine phosphatases: The quest for negative regulators of insulin action. *American Journal of Physiology, Endocrinology and Metabolism*, 284:E663–E670, 2003.

[9] Kaveh Ashrafi, Francesca Y. Chang, Jennifer L. Watts, Andrew G. Fraser, Ravi S. Kamath, Julie Ahringer, and Gary Ruvkun. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature*, 421:268–272, 2003.

[10] A.H. Barnett, C. Eff, R.D. Leslie, and D.A. Pyke. Diabetes in identical twins: A study of 200 pairs. *Diabetologia*, 20:87–93, 1981.

[11] A.S. Bassett, B.C. McGillivray, B.D. Jones, and J.T. Pantzar. Partial trisomy chromosome 5 cosegregating with schizophrenia. *Lancet*, 1(8589):799–801, 1988.

[12] M.J. Beaudet, M. Desrochers, A.A. Lachaud, and A. Anderson. The CYP2B2 phenobarbital response unit contains binding sites for hepatocyte nuclear factor 4, PBX–PREP1, the thyroid hormone receptor beta and the liver X receptor. *Biochemical Journal*, 388:407–418, 2005.

[13] F. Belfiore, S. Ianneloo, A.M. Rabuazzo, and R. Campione. Metabolic effects of short–term fasting in obese hyperglycaemic humans and mice. *International Journal of Obesity*, 11:631–640, 1987.

[14] Robert M. Berne and Matthew N. Levy. *Physiology.* Mosby, St. Louis, MO, fourth edition, 1998.

[15] Robert M. Berne and Matthew N. Levy. *Physiology, 4th Edition.* Mosby, St. Louis, MO, 1998.

[16] J.A. Bezerra, T.L. Carrick, J.L. Degen, D. Witte, and S.J.F. Degen. Biological effects of targeted inactivation of hepatocyte growth factor–like protein in mice. *Journal of Clinical Investigation*, 101(5):1175–1183, 1998.

[17] Jorge A. Bezerra. Personal communication. Email to Dr. Bezerra confirmed that the Hgfl knockout mice had not been studied in the context of glucose homeostasis nor hepatic glucose output, August 2005.

[18] P.E. Bickel, P.E. Scherer, J.E. Schnitzer, P. Oh, M.P. Lisanti, and H.F. Lodish. Flotillin and epidermal surface antigen define a new family of caveolae–associated integral membrane proteins. *Journal of Biological Chemistry*, 272(21):13793–13802, 1997.

[19] Bioinformatics and Metabolic Engineering Laboratory at MIT. http://web.mit.edu/cheme/gnswebpage/links.shtml. URL, 2005.

[20] M.A. Black and R.W. Doerge. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, 18(12):1609–1616, 2002.

[21] J.A. Bokar, M.E. Shambaugh, D. Polayes, A.G. Matera, and F.M. Rottman. Purification and cDNA cloning of the adomet–binding subunit of the human mRNA (N6–adenosine)–methyltransferase. *RNA*, 3:1233–1247, 1997.

[22] A. Borkhardt and O. Heidenreich. RNA interference as a potential tool in the treatment of leukaemia. *Expert Opinion on Biological Therapy*, 2004.

[23] F. Bosch, A. Pujol, and A. Valera. Transgenic mice in the analysis of metabolic regulation. *Annual Review of Nutrition*, 18:207–232, 1998.

[24] M Boshart, F Weber, G Jahn, K Dorsch-Hasler, B Fleckenstein, and W Schaffner. A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. *Cell*, 41(2):521–530, Jun 1985.

[25] Michael Boutros, Amy A Kiger, Susan Armknecht, Kim Kerr, Marc Hild, Britta Koch, Stefan A Haas, Heidelberg Fly Array Consortium, Renato Paro, and Norbert Perrimon. Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science*, 303(5659):832–835, Feb 2004.

[26] Gudrun A. Brockmann and Marianna R. Bevova. Using mouse models to dissect the genetics of obesity. *TRENDS in Genetics*, 18:367–376, 2002.

[27] N.E. Broude, T. Sano, C.L. Smith, and C.R. Cantor. Enhanced DNA sequencing by hybridization. *Proceedings of the National Academies of Science — U.S.A.*, 91:3072–3076, 1994.

[28] John W Bullen Jr., Mary Ziotopoulou, Linda Ungsunan, Jatin Misra, Ilias Alevizos, Efi Kokkotou, Eleftheria Maratos-Flier, Gregory Stephanopoulos, and Christos S. Mantzoros. Short–term resistance to diet–induced obesity in a/ j mice is not associated with regulation of hypothalamic neuropeptides. *American Journal of Physiology: Endocrinology and Metabolism*, 287(4):E662–E670, 2004.

[29] Corey M Carlson and David A Largaespada. Insertional mutagenesis in mice: new perspectives and tools. *Nat Rev Genet*, 6(7):568–580, Jul 2005.

[30] P. Carninci, Y. Shibata, N. Hayatsu, Y. Sugahara, K. Shibata, M. Itoh, H. Konno, Y. Okazaki, M. Muramatsu, and Y. Hayashizaki. Normalization and subtraction of cap–trapper–selected cDNAs to prepare full–length cDNA libraries for rapid discovery of new genes. *Genome Research*, 10(10):1617–1630, 2000.

[31] R.K. Chan and C.A. Otte. Isolation and genetic analysis of S*accharomyces cerevisiae* mutants supersensitive to G1 arrest by a factor and α-factor pheromones. *Molecular Cellular Biology*, 2:11–20, 1982.

[32] Erwin Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6:201–209, 1950.

[33] Tapan K. Chatterjee, Alex K. Eapen, and Rory A. Fisher. A truncated form of RGS3 negatively regulates G protein-coupled receptor stimulation of adenylyl cyclase and phosphoinositide phospholipase C. *The Journal of Biological Chemistry*, 272(24):15481–15487, 1997.

[34] Cecil Chen, Krzysztof J. Grzegorzewski, Steve Barash, Qinghai Zhao, Helmut Schneider, Qi Wang, Mallika Singh, Laurie Pukac, Adam C. Bell, Roxanne Duan, Tim Coleman, Alokesh Duttaroy, Susan Cheng, Jon Hirsch, Linyi Zhang, Yanick Lazard, Carrie Fischer, Melisa Carey Barber, Zhi-Dong Ma, Ya-Qin Zhang, Peter Reavey, Lilin Zhong, Baiqin Teng, Indra Sanyal, Steve M. Ruben, Olivier Blondel, and Charles E. Birse. An integrated functional genomics screening program reveals a role for BMP-9 in glucose homeostasis. *Nature Biotechnology*, 21:294–301, March 2003.

[35] G. Chen, T.G. Gharib, C.C. Huang, J.M.G. Taylor, D.E. Misek, S.L.R. Kardia, T.J. Giordano, M.D. Iannettoni, M.B. Orringer, S.M. Hanash, and D.G. Beer. Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular & Cellular Proteomics*, 1:304–313, 2002.

[36] Alan Cheng, Nadia Dube, Feng Gu, and Michel L. Tremblay. Coordinated action of protein tyrosine phosphatases in insulin signal transduction. *European Journal of Biochemistry*, 269:1050–1059, 2002.

[37] Janice Yang Chou, Adriana Zingone, and Chi-Jiunn Pan. Adenovirus-mediated gene therapy in a mouse model of glycogen storage disease type 1a. *Eur J Pediatr*, 161 Suppl 1:56–61, Oct 2002.

[38] B. Christ, E. Yazici, and A. Nath. Phosphatidylinositol 3–kinase and protein kinase c contribute to the inhibition by interleukin 6 of phosphoenolpyruvate carboxykinase gene expression in cultured hepatocytes. *Hepatology*, 31:461–468, 2000.

[39] Francis S. Collins, Michael Morgan, and Aristides Patrinos. The human genome project: Lessons from large-scale biology. *Science*, 11:286–290, 2003.

[40] F.S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. New goals for the u.s. human genome project: 1998–2003. *Science*, 282:682–689, 1998.

[41] S Connelly. Adenoviral vectors for liver-directed gene therapy. *Curr Opin Mol Ther*, 1(5):565–572, Oct 1999.

[42] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[43] R.H. Costa, D.R. Grayson, and J.E. Darnell. Multiple hepatocyte–enriched nuclear factors function in the regulation of transthyretin and alpha 1–antitrypsin genes. *Molecular Cell Biology*, 9(4):1415–1425, 1989.

[44] J.A. Coyne and E. Beecham. Heritability of two morphological characters within and between natural populations of drosophlia melanogaster. *Genetics*, 117:727–737, 1987.

[45] C. Dai, Y. Li, J. Yang, and Y. Liu. Hepatocyte growth factor preserves beta cell mass and mitigates hyperglycemia in streptozotocin–induced diabetic mice. *Journal of Biological Chemistry*, 278(29):27080–27087, 2003.

[46] C. Dai, J. Yang, S. Bastacky, J. Xia, Y. Li, and Y. Liu. Intravenous administration of hepatocyte growth factor gene ameliorates diabetic nephropathy in mice. *Journal of the American Society of Nephrology*, 15(10):2637–2647, 2004.

[47] A. Darvasi and M.A. Soller. A simple method to calculate resolving power and confidence interval of qtl map location. *Behavioral Genetics*, 27:125–132, 1997.

[48] S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003.

[49] S Davis and J C Watson. In vitro activation of the interferon-induced, double-stranded RNA-dependent protein kinase PKR by RNA from the 3' untranslated regions of human alpha-tropomyosin. *Proc Natl Acad Sci U S A*, 93(1):508–513, Jan 1996.

[50] A.G. de Brevern, S. Hazout, and A. Malpertuy. Influence of microarrays experiments missing values on the stability of gene groups by hierarchial clustering. *BMC Bioinformatics*, 5(1):160, 2004.

[51] G. de Jong. Quantitative genetics of reaction norms. *Journal of Evolutionary Biology*, 3:447–468, 1990.

[52] Laura Dean and Johanna McEntyre. *The Genetic Landscape of Diabetes.* National Center for Biotechnology Information, Bethesda, MD, 2004.

[53] F. Diehl, S. Grahlmann, M. Beier, and J.D. Hoheisel. Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Research*, 29(7):E38, 2001.

[54] William R. Dillon and Matthew Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

[55] Olivier Donze and Didier Picard. RNA interference in mammalian cells using siRNAs synthesized with T7 RNA polymerase. *Nucleic Acids Res*, 30(10):e46, May 2002.

[56] E.M. East. Studies on size inheritance in nicotiana. *Genetics*, 1:164–176, 1916.

[57] K. Ebihara, Y. Ogawa, H. Masuzaki, M. Shintani, F. Miyanaga, M. Aizawa-Abe, T. Hayashi, K. Hosoda, G. Inoue, Y. Yoshimasa, O. Gavrilova, M.L. Reitman, and K. Nakao. Transgenic overexpression of leptin rescues insulin resistance and diabetes in a mouse model of lipoatrophic diabetes. *Diabetes*, 50:1440–1448, 2001.

[58] S M Echwald, T I Sorensen, T Andersen, A Tybjaerg-Hansen, J O Clausen, and O Pedersen. Mutational analysis of the proopiomelanocortin gene in Caucasians with early onset obesity. *Int J Obes Relat Metab Disord*, 23(3):293–298, Mar 1999.

[59] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, August 2002.

[60] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome–wide expression patterns. *Proceedings of the National Academies of Science — U.S.A.*, 95(25):14863–14868, 1998.

[61] K. El-Haschimi, D.D. Pierroz, S.M. Hileman, C. Bjorbaek, and J.S. Flier. Two defects contribute to hypothalamic leptin resistance in mice with diet-induced obesity. *Journal of Clinical Investigation*, 105:1827–1832, 2000.

[62] S M Elbashir, J Martinez, A Patkaniowska, W Lendeckel, and T Tuschl. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO Journal*, 20(23):6877–6888, Dec 2001.

[63] S.M. Elbashir, J. Harboth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. Duplexes of 21–nucleotide RNAs for mediate RNA interference in cultured mammalian cells. *Nature*, 411:494–498, 2001.

[64] M. Elchebly, P. Payette, E. Michaliszyn, W. Cromlish, S. Collins, A.L. Loy, D. Normandin, A. Cheng, J. Himms-Hang, C.C. Chan, C. Ramachandran, M.J. Gresser, M.L. Tremblay, and B.P. Kennedy. Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1b gene. *Science*, 283:1544–1548, 1999.

[65] M. Escalante-Pulido, A. Escalante-Herrera, M.E. Milke-Najar, and M. Alpizar-Salazar. Effects of weight loss on insulin secretion and *in vivo* insulin sensitivity in obese diabetic and non–diabetic subjects. *Diabetes Nutrition and Metabolism*, 16:277–283, 2003.

[66] H.J. Evans, K.E. Buckton, and A.T. Sumner. Cytological mapping of human chromosomes: Results obtained with quinacrine fluorescence and the acetic-saline-giemsa techniques. *Chromosoma*, 35(3):310–325, 1971.

[67] Boston University Transgenic/ Knock Out Core Facility. http://www.bumc.bu.edu/dept/content.aspx?departmentid=439&pageid=9220. URL, 2005.

[68] A. Fagot-Campagna and K. Narayan. Type 2 diabetes in children. *British Medical Journal*, 322:377–387, 2001.

[69] A. Fagot-Campagna, D.J. Pettit, M.M. Engelgau, N.R. Burrows, L.S. Geiss, R. Valdez, G.L. Beckles, J. Saaddine, E.W. Gregg, D.F. Williamson, and K.M. Narayan. Type 2 diabetes among north american children and adolescents: An epidemiologic review and a public health perspective. *Journal of Pediatrics*, 136:664–672, 2000.

[70] R. Kretschmer-Kazemi Far, W. Nedbal, and G. Sczakiel. Concepts to automate the theoretical design of effective antisense oligonucleotides. *Bioinformatics*, 17(11):1058–1061, 2001.

[71] A. Favello, L. Hillier, and R.K. Wilson. Genomic DNA sequencing methods. *Methods in Cell Biology*, 48:551–569, 1995.

[72] O. Fiehn, J. Kopka, R.N. Trethewey, and L. Willmitzer. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromotography and quadrupole mass spectrometry. *Analytical Chemistry*, 72(15):3573–3580, 2000.

[73] A Fire, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–811, Feb 1998.

[74] Jeffrey S. Flier and Christos Mantzoros. *Syndromes of Insulin Resistance and Mutant Insulin.* Diabetes Mellitus, Carbohydrate Metabolism, and Lipid Disorders. Harcourt Canada, Canada, 2000.

[75] Jonathan Flint, William Valdar, Sagiv Shifman, and Richard Mott. Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Reviews Genetics*, 6:271–286, 2005.

[76] R.C. Frederich, A. Hamann, S. Anderson, B. Lollmann, and J.S. Flier. Leptin levels reflect body lipid content in mice: Evidence for diet-induced resistance to leptin action. *Nature Medicine*, 1:1311–1314, 1995.

[77] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. In *Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan, 2000.

[78] D.J. Galas and S.J. McCormack. An historical perspective on genomic technologies. *Current Issues in Molecular Biology*, 4:123–127, 2003.

[79] S. Ganguly and A.I. Skoultchi. Absolute rates of globin gene transcription and mRNA formation during differentiation of culture mouse erythroleukemia cells. *Journal of Biological Chemistry*, 260:12167–12173, 1985.

[80] A. Garcia-Ocana, K.K. Takane, M.A. Syed, W.M. Philbrick, R.C. Vasavada, and A.F. Stewart. Hepatocyte growth factor overexpression in the islet of transgenic mice increases beta cell proliferation, enhances islet mass, and induces mild hypoglycemia. *Journal of Biological Chemistry*, 275(2):1226–1232, 2000.

[81] A. Gasch and M. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering. *Genome Biology*, 3(11):0059.1, 2002.

[82] H. Ge, A.J.M. Walhout, and M. Vidal. Integrating 'omic' information: A bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10):551–560, 2003.

[83] P. Geladi and B.R. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[84] A L Goldberg, T N Akopian, A F Kisselev, D H Lee, and M Rohrwild. New insights into the mechanisms and importance of the proteasome in intracellular protein degradation. *Biol Chem*, 378(3-4):131–140, Mar 1997.

[85] G. Golub and V. Pereya. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19:R1–R26, 2003.

[86] Michael I. Goran, Geoff D.C. Ball, and Martha L. Cruz. Obesity and risk of type 2 diabetes and cardiovascular disease in children and adolescents. *Journal of Clinical Endocrinology and Metabolism*, 88(4):1417–1427, 2003.

[87] Ivan Gout, Gayle Middleton, Jimi Adu, Natalia N. Ninkina, Ludmila B. Drobot, Valery Filonenko, Gennady Matsuka, Alun M. Davies, Michael Waterfield, and Vladimir L. Buchman. Negative regulation of PI 3-kinase by RUK, a novel adaptor protein. *EMBO Journal*, 19(15):4015–4025, 2000.

[88] B. Gowda, M. Sar, X. Mu, J. Cidlowski, and T. Welbourne. Coordinate modulation of glucocorticoid receptor and glutaminase gene expression in LLC-PC$_1$-F$^+$ cells. *American Journal of Physiology*, 270:C825–C831, 1996.

[89] Anthony J.F. Griffiths, Jeffry H. Miller, David T. Suzuki, Richard C. Lewontin, and William M. Gelbart. *An Introduction to Genetic Analysis*, chapter 27, page 818. W.H. Freeman and Company, New York, NY, sixth edition, 1996. This is a full INBOOK entry.

[90] Helge Grosshans and Frank J Slack. Micro-RNAs: small is plentiful. *J Cell Biol*, 156(1):17–21, Jan 2002.

[91] M. Guttinger, F. Sutti, M. Panigada, S. Porcellini, B. Merati, M. Mariani, T. Teesalu, G.G. Consalez, and F. Grassi. Epithelial v–like antigen EVA, a novel member of the immunoglobulin superfamily, expressed in embryonic epithelia with a potential role as homotypic adhesion molecule in thymus histogenesis. *Journal of Cell Biology*, 141(4):1061–1071, 1998.

[92] S.P. Gygi, Y. Rochon, B.R. Franza, and R. Aebersold. Correlation between protein and mRNA in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730, 1999.

[93] S.M. Haffner, R. D'Agostino, M.F. Saad, M. Rewers, L. Mykkanen, J. Selby, G. Howard, P.J. Savage, R.F. Hamman, L.E. Wagenknecht, and R.N. Bergman. Increased insulin resistance and insulin secretion in nondiabetic african-americans and hispanics compared to non-hispanic whates: the insulin resistance atherosclerosis study. *Diabetes*, 45:742–748, 1996.

[94] S.M. Haffner, H. Miettinen, and M.P. Stern. Are risk factors for conversion to niddm similar in high and low risk populations? *Diabetologia*, 40:62–66, 1997.

[95] Steven M Haffner. Epidemiology of type 2 diabetes: Risk factors. *Diabetes Care*, 21:C3–C6, 1998.

[96] Makiko Hamada, Toshiaki Ohtsuka, Reimi Kawaida, Makoto Koizumi, Koji Morita, Hidehiko Furukawa, Takeshi Imanishi, Makoto Miyagishi, and Kazunari Taira. Effects on RNA interference in gene expression (RNAi) in cultured mammalian cells of mismatches and the introduction of chemical modifications at the 3'-ends of siRNAs. *Antisense Nucleic Acid Drug Dev*, 12(5):301–309, Oct 2002.

[97] S.M. Hammond, E. Bernstein, D. Beach, and G.J. Hannon. An RNA–directed nuclease mediates post–transcriptional gene silencing in *Drosophila* cells. *Nature*, 404:293–296, 2000.

[98] G.J. Hannon. RNA Interference. *Nature*, 418:244–251, 2002.

[99] M.I. Harris. Noninsulin-dependent diabetes mellitus in black and white americans. *Diabetes Metabolism Revues*, 6:71–90, 1990.

[100] R B Harris, J Zhou, M Shi, S Redmann, R L Mynatt, and D H Ryan. Overexpression of agouti protein and stress responsiveness in mice. *Physiol Behav*, 73(4):599–608, Jul 2001.

[101] Daniel L. Hartl and Vitezslav Orel. What did Gregor Mendel think he discovered. *Genetics*, 131(2):245–253, 1992.

[102] V Hatzimanikatis and K H Lee. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metab Eng*, 1(4):275–281, Oct 1999.

[103] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cDNA microarray analysis. *Biotechniques*, 29(3):548–563, 2000.

[104] Leoine K. Heilbronn and Eric Ravussin. Calorie restriction and aging: Review of the literature and implications for studies in humans. *American Journal of Clinical Nutrition*, 78:361–369, 2003.

[105] Peter C. Heinrich, Iris Behrmann, Gerhard Muller-Newen, Fred Schaper, and Lutz Graeve. Interleukin–6–type cytokine signalling through the gp130/jak/STAT pathway. *Biochemistry Journal*, 334:297–314, 1998.

[106] Michael T Hemann, Jordan S Fridman, Jack T Zilfou, Eva Hernando, Patrick J Paddison, Carlos Cordon-Cardo, Gregory J Hannon, and Scott W Lowe. An epi-allelic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes in vivo. *Nat Genet*, 33(3):396–400, Mar 2003.

[107] Joene Hendry. FDA approves exenatide. *Diabetes, Obesity, and Cardiovascular Disease*, pages 1–6, July 2005.

[108] Christophe Himber, Patrice Dunoyer, Guillaume Moissiard, Christophe Ritzenthaler, and Olivier Voinnet. Transitivity-dependent and -independent cell-to-cell movement of RNA silencing. *EMBO J*, 22(17):4523–4533, Sep 2003.

[109] Joel N. Hirschhorn and Mark J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95–108, 2005.

[110] Alexander Hoffmann, Andre Levchenko, Martin L Scott, and David Baltimore. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science*, 298(5596):1241–1245, Nov 2002.

[111] P. Hogan, T. Dall, and P. Nikolov. Economic costs of diabetes in the us in 2002. *Diabetes Care*, 26:917–932, 2003.

[112] A.L. Holleran, D.A. Briscoe, G. Fiskum, and J.K. Kelleher. Glutamine metabolism in AS-30D hepatoma cells. evidence for its conversion into lipids via reductive carboxylation. *Molecular Cell Biochemistry*, 152:95–101, 1995.

[113] N.S. Holter, M. Mitra, A. Maritan, M. Cleplak, J.R. Banavar, and N.V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academies of Science — U.S.A.*, 97:8409–8414, 2000.

[114] Leroy Hood, M.W. Hunkapiller, and L.M. Smith. Automated DNA sequencing and analysis of the human genome. *Genomics*, 1(3):201–212, 1987.

[115] Daehee Hwang, William A. Schmitt, Gregory Stephanopoulos, and George Stephanopoulos. Determination of minimum sample size and discriminatory expression patterns. *Bioinformatics*, 18:1184–1193, 2002.

[116] H. Inoue, W. Ogawa, M. Ozaki, S. Haga, M. Matsumoto, K. Furukawa, N. Hashimoto, Y. Kido, T. Mori, H. Sakaue, K. Teshigawara, S. Jin, H. Iguchi, R. Hiramatsu, D. LeRoith, K. Takeda, S. Akira, and M. Kasuga. Role of STAT-3 in regulation of hepatic gluconeogenic genes and carbohydrate metabolism *in vivo*. *Nature Medicine*, 10(2):168–174, 2004.

[117] Aimee L Jackson, Steven R Bartz, Janell Schelter, Sumire V Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol*, 21(6):635–637, Jun 2003.

[118] Jean-Marc Jacque, Karine Triques, and Mario Stevenson. Modulation of HIV-1 replication by RNA interference. *Nature*, 418(6896):435–438, Jul 2002.

[119] S.A. Jelinsky and L.D. Samson. Global response of S*accharomyces cerevisiae* to an alkylating agent. *Proceedings of the National Academy of Sciences - USA*, 96:1486–1491, 1999.

[120] S.W. Jeong and S.R. Ikeda. Endogenous regulator of G-protein signalling proteins modify N-type calcium channel modulation in rat sympathetic neurons. *Journal of Neuroscience*, 20:4489–4496, 2000.

[121] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[122] Anthony W. Ferrante Jr., Marie Thearle, Ted Liao, and Rudolph L. Leibel. Effects of leptin deficiency and short-term repletion on hepatic gene expression in genetically obese mice. *Diabetes*, 50(10):2268–2278, 2001.

[123] W.A. Schmitt Jr. *Extracting Transcriptional Regulatory Information from DNA Microarray Expression Data.* PhD thesis, Massachusetts Institute of Technology, July 2000.

[124] R.T. Kamimura. *Application of Multivariate Statistics to Fermentation Database Mining.* PhD thesis, Massachusetts Institute of Technology, June 1997.

[125] M.D. Kane, T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, and S.J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, 28(22):4552–4557, 2000.

[126] K.C. Kao, Y.L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury, and J.C. Liao. Transcriptome–based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proceedings of the National Academies of Science — U.S.A.*, 101(2):641–646, 2004.

[127] N. Kato, K. Nemoto, K. Nakanishi, R. Morishita, Y. Kaneda, M. Uenoyama, T. Ikeda, and K. Fujikawa. Nonviral gene transfer of human hepatocyte growth factor improves streptozotocin–induced diabetic neuropathy in rats. *Diabetes*, 54(3):846–854, 2005.

[128] K.B. Keller and L. Lemberg. Obesity and the metabolic syndrome. *American Journal of Critical Care*, 12(2):167–170, 2003.

[129] O. Kempthorne. *An Introduction to Genetic Statistics.* Wiley, New York, 1957.

[130] M.K. Kerr and G.A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2:183–201, 2001.

[131] M.K. Kerr and G.A. Churchill. A statistical design and the analysis of gene expression microarrays. *Genetics Research*, 77:123–128, 2001.

[132] Sander Kerstein, Beatrice Desvergne, and Walter Wahli. Roles of ppars in health and disease. *Nature*, 405:421–424, 2000.

[133] William M. Kettyle and Ronald A. Arky. *Endocrine Pathophysiology.* Lippincott — Raven, 227 East Washington Square, Philadelphia, PA, 1998.

[134] K.R. Khrapko, Yu P. Lysov, A.A. Khorlin, I.B. Ivanov, G.M. Yershov, S.K. Vasilenko, V.L. Florentiev, and A.D. Mirzabekov. A method for DNA sequencing by hybridization with oligonucleotide matrix. *DNA Sequencing*, 1:375–388, 1991.

[135] K Kobayashi, K Oka, T Forte, B Ishida, B Teng, K Ishimura-Oka, M Nakamuta, and L Chan. Reversal of hypercholesterolemia in low density lipoprotein receptor knockout mice by adenovirus-mediated gene transfer of the very low density lipoprotein receptor. *J Biol Chem*, 271(12):6852–6860, Mar 1996.

[136] W.J. Koch, B.E. Hawes, L.F. Allen, and R.J. Lefkowitz. Direct evidence that G_i-coupled receptor stimulation of mitogen-activated protein kinase is mediated

by $G_{\beta\gamma}$ activation of p21^{ras}. *Proceedings of the National Academy of Sciences: U.S.A.*, 91:12706–12710, 1994.

[137] Ron Korstanje and Beverly Paigen. From qtl to gene: the harvest begins. *Nature Genetics*, 31:235–236, 2002.

[138] Anna M Krichevsky and Kenneth S Kosik. RNAi functions in cultured mammalian neurons. *Proc Natl Acad Sci U S A*, 99(18):11926–11929, Sep 2002.

[139] H Krude, H Biebermann, W Luck, R Horn, G Brabant, and A Gruters. Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature Genetics*, 19(2):155–157, Jun 1998.

[140] Klas Kullander and Rudiger Klein. Mechanisms and functions of EPH and EPHRIN signalling. *Nature Reviews Molecular Cell Biology*, 3:475–486, 2002.

[141] W.P. Kuo, T.K. Jenssen, A.J. Butte, L. Ohno-Machado, and I.S. Kohane. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–412, 2002.

[142] N. Kurihara, A. Iwama, J. Tatsumi, K. Ikeda, and T. Suda. Macrophage–stimulating protein activates stk receptor tyrosine kinase on osteoclasts and facilitates bone resorption by osteoclast–like cells. *Blood*, 87:3704–3710, 1996.

[143] Finny G. Kuruvilla, Peter J. Park, and Stuart L. Schreiber. Vector algebra in the analysis of genome-wide expression data. *Genome Biology*, 3(3):0011.1–0011.11, 2002.

[144] Eric S. Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199, 1989.

[145] J.J. Lebrun, S. Ali, L. Sofer, A. Ullrich, and P.A. Kelly. Prolactin–induced proliferation of Nb2 cells involves tyrosine phosphorylation of the prolactin receptor and its associated tyrosine kinase JAK2. *Journal of Biological Chemistry*, 269(19):14021–14026, 1994.

[146] F. Li and G.D. Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17(11):1067–1076, 2001.

[147] J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, and V.P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academies of Science — U.S.A.*, 100(26):15522–15527, 2003.

[148] S. Lin, T.C. Thomas, L.H. Storlien, and X.F. Huang. Development of high fat diet–induced obesity and leptin resistance in C57Bl/6J mice. *International Journal of Obesity and Related Metabolic Disorders*, 24:639–646, 2000.

[149] R. Lipshutz and S.P. Fodor. Advanced DNA sequencing technologies. *Current Biology*, 4:376–380, 1994.

[150] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittman, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

[151] John A. Luckey, Howard Drossman, Anthony J. Kostichka, David A. Mead, Jonathan D'Cunha, Tracy B. Norris, and Lloyd M. Smith. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research*, 18(15):4417–4421, 1990.

[152] D.S. Ludwig, N.A. Tritos, J.W. Mastaitis, R. Kulkarni, E. Kokkotou, J. Elmquist, B. Lowell, J.S. Flier, and E. Maratos-Flier. Melanin–concentrating hormone overexpression in transgenic mice leads to obesity and insulin resistance. *Journal of Clinical Investigation*, 107(3):379–386, 2001.

[153] O.A. MacDougald, C.S. Hwang, H. Fan, and M.D. Lane. Regulated expression of the obese gene product (leptin) in white adipose tissue and 3T3–L1 adipocytes. *Proceedings of the National Academy of Science — U.S.A.*, 92:9034–9037, 1995.

[154] I. Maeda, Y. Kohara, M. Yamamoto, and A. Sugimoto. Large scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Current Biology*, 11:171–176, 2001.

[155] Takahiko Matsuda and Constance L Cepko. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc Natl Acad Sci U S A*, 101(1):16–22, Jan 2004.

[156] A.M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academies of Science — U.S.A.*, 74:560–564, 1977.

[157] Michele M Maxwell, Piera Pasinelli, Aleksey G Kazantsev, and Robert H Jr Brown. RNA interference-mediated silencing of mutant superoxide dismutase rescues cyclosporin A-induced death in cultured neuroblastoma cells. *Proc Natl Acad Sci U S A*, 101(9):3178–3183, Mar 2004.

[158] Anton P McCaffrey, Leonard Meuse, Thu-Thao T Pham, Douglas S Conklin, Gregory J Hannon, and Mark A Kay. RNA interference in adult mice. *Nature*, 418(6893):38–39, Jul 2002.

[159] J. Denis McGarry. What if Minkowski had been aguesic. *Science*, 258:766 – 770, October 1992.

[160] M.T. McManus and P.A. Sharp. Gene silencing in mammals by small interfering RNAs. *Nature Reviews Genetics*, 3:737–747, 2002.

[161] D. Meldrum. Automation for genomics, part two: Sequencers, microarrays and future trends. *Genome Research*, 10:1288–1303, 2000.

[162] O.J. Miller, D.A. Miller, and D. Warburton. Application of new staining techniques to the study of human chromosomes. *Progress in Medical Genetics*, 9:1–47, 1973.

[163] W.M. Miller, C.R. Wilke, and H.W. Blanch. Transient responses of hybridoma cells to nutrient additions in continuous culture. 1. glucose pulse and step changes. *Biotechnology & Bioengineering*, 33:477–486, 1989.

[164] J. Misra, W. Schmitt, D. Hwang, L.L. Hsiao, S. Gullans, and G. Stephanopoulos. Interactive exploration of microarray expression patterns in a reduced dimensional space. *Genome Research*, 12:1112–1120, 2002.

[165] Asif Mohmmed, Palakodeti V N Dasaradhi, Raj K Bhatnagar, Virander S Chauhan, and Pawan Malhotra. In vivo gene silencing in Plasmodium berghei–a mouse malaria model. *Biochem Biophys Res Commun*, 309(3):506–511, Sep 2003.

[166] B C Moon and J M Friedman. The molecular basis of the obese mutation in ob2J mice. *Genomics*, 42(1):152–156, May 1997.

[167] R.K. Mortimer and D.C. Hawthorne. Genetic mapping in yeast. *Methods in Cell Biology*, 11:221–233, 1975.

[168] U Muller. Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis. *Mech Dev*, 82(1-2):3–21, Apr 1999. Historical Article.

[169] R.S. Muraoka, S.E. Waltz, and S.J. Friezner Degen. Expression of hepatocyte growth factor–like protein is repressed by retinoic acid and enhanced by cyclic adenosine 3',5'–monophosphate response element binding protein (creb)–binding protein (cbp). *Endocrinology*, 140(1):187–196, 1999.

[170] Jason W Myers, Joshua T Jones, Tobias Meyer, and James E Jr Ferrell. Recombinant Dicer efficiently converts large dsRNAs into siRNAs suitable for gene silencing. *Nat Biotechnol*, 21(3):324–328, Mar 2003. Evaluation Studies.

[171] R. Nadon and J. Shoemaker. Statistical issues with microarrays: Processing and analysis. *TRENDS in Genetics*, 18(5):265–271, 2002.

[172] Yusaku Nakabeppu and Daniel Nathans. A naturally occurring truncated for of FosB that inhibits Fos/Jun transcriptional activity. *Cell*, 64:751–759, 1991.

[173] NCBI, Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/. URL, 2005.

[174] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry.* Freeman, 4 edition, April 2004.

[175] Richard R. Neubig and David P. Siderovski. Regulators of G-protein signalling as new central nervous system drug targets. *Nature Reviews Drug Discovery*, 1:187–197, 2002.

[176] R.H. Nicholson and A.W. Nicholson. Molecular characterization of a mouse cDNA encoding dicer, a ribonuclease III ortholog involved in RNA interference. *Mammalian Genome*, 13:67–73, 2002.

[177] Charles E. Novitski. Revision of fisher's analysis of mendel's garden pea experiments. *Genetics*, 166(3):1139–1140, 2004.

[178] R.J. Oakey, M.L. Watson, and M.F. Seldin. Construction of a physical map on mouse and human chromosome 1: Comparison of 13 Mb of mouse and 11 Mb of human DNA. *Human Molecular Genetics*, 1(8):613–620, 1992.

[179] Richard M. O'Brien and Daryl K. Granner. Regulation of gene expression by insulin. *Biochemistry Journal*, 278:609–619, 1991.

[180] Richard M. O'Brien and Daryl K. Granner. Regulation of gene expression by insulin. *Physiological Reviews*, 76(4):1109–1161, 1996.

[181] R.M. O'Doherty, D.L. Lehman, S. Telemaque-Potts, and C.B. Newgard. Metabolic impact of glucokinase overexpression in liver. *Diabetes*, 48:2022–2027, 1999.

[182] Duncan T. Odom, Nora Zizlsperger, D. Benjamin Gordon, George W. Bell, Nicola J. Rinaldi, Heather L. Murray, Tom L. Volkert, Jorg Schreiber, P. Alexander Rolfe, David K. Gifford, Ernest Fraenkel, Graeme I. Bell, and

Richard A. Young. Control of pancreas and liver gene expression by HNF transcription factors. *Nature*, 303:1378–1381, 2004.

[183] U.S. Department of Agriculture. http://www.itis.usda.gov/servlet/singlerpt /singlerpt. URL, 2005.

[184] U.S. Department of Health and Human Services and U.S. Department of Energy. Understanding our genetic inheritance — the u.s. human genome project. *DOE/ER-0452P; NIH Publication No. 90-1590*, 1990. http://www.ornl.gov/sci /techresources/Human_Genome/project/5yrplan/summary.shtml.

[185] Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, 26:S5–S20, 2003.

[186] DOE-NIH Joint Subcommittee on the Human Genome. NIH, DOE guidelines encourage sharing of data, resources. *Human Genome News*, 4(5), January 1993. http://www.ornl.gov/sci/techresources/Human_Genome/publicat /hgn/v4n5/04share.shtml.

[187] Patrick J Paddison, Amy A Caudy, and Gregory J Hannon. Stable suppression of gene expression by RNAi in mammalian cells. *Proc Natl Acad Sci U S A*, 99(3):1443–1448, Feb 2002.

[188] Manika Pal-Bhadra, Utpal Bhadra, and James A Birchler. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. *Mol Cell*, 9(2):315–327, Feb 2002.

[189] P.E. Parekh, A.E. Petro, J.M. Tiller, M.N. Feinglos, and R.S. Surwit. Reversal of diet-induced obesity and diabetes in C57BL/6J mice. *Metabolism*, 47(9):1089–1096, 1998.

[190] H. Pertoft and B. Smedsrod. *Separation and Characterization of Liver Cells*, volume 4 of *Cell Separation: Methods and Selected Applications*. Academic Press, 1987.

[191] G Pesole, F Mignone, C Gissi, G Grillo, F Licciulli, and S Liuni. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, 276(1-2):73–81, Oct 2001.

[192] P.A. Pevzner, Yu P. Lysov, K.R. Khrapko, A.V. Belyavsky, V.L. Florentiev, and A.D. Mirzabekov. Improved chips for sequencing by hybridization. *Journal of Biomolecule Structure Dynamics*, 9:399–410, 1991.

[193] M. Pigliucci, J. Whitton, and C.D. Schlichting. Reaction norms of *Arabidopsis*. I. Plasticity of characters and correlations across water, nutrient and light gradients. *Journal of Evolutionary Biology*, 8(4):421–438, 1995.

[194] M. Pitarque, C. Rodriguez-Antona, M. Oscarson, and M. Ingelman-Sundberg. Transcriptional regulation of the human CYP2A6 gene. *Journal of Pharmacological Experimentation*, 313(2):814–822, 2005.

[195] P.L. Podolin, P. Denny, N. Armitage, C.J. Lord, and N.J. Hill. Localization of two insulin-dependent diabetes (idd) genes to the idd10 region on mouse chromosome 3. *Mammalian Genome*, 9:283–286, 1998.

[196] J.C. Portais, P. Voisin, M. Merle, and P. Canioni. Glucose and glutamine metabolism in C6 glioma cells studied by carbon 13 NMR. *Biochimie*, 78:155–164, 1996.

[197] Roy Porter. *The Greatest Benefit to Mankind: A Medical History of Humanity*. W.W. Nortion & Company, New York, NY, 1997.

[198] Patrick Provost, David Dishart, Johanne Doucet, David Frendewey, Bengt Samuelsson, and Olof Radmark. Ribonuclease activity and RNA binding of recombinant human Dicer. *EMBO J*, 21(21):5864–5874, Nov 2002.

[199] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[200] John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

[201] R. Michael Raab, John Bullen, Joanne Kelleher, Christos Mantzoros, and Gregory Stephanopoulos. Regulation of mouse hepatic genes in response to diet induced obesity, insulin resistance and fasting. *Nutrition & Metabolism*, 2:15, 2005.

[202] R. Michael Raab and Gregory Stephanopoulos. Dynamics of gene silencing by RNA interference. *Biotechnology & Bioengineering*, 88(1):121–132, 2004.

[203] Arvind Raghavan, Rachel L Ogilvie, Cavan Reilly, Michelle L Abelson, Shalini Raghavan, Jayprakash Vasdewani, Mitchell Krathwohl, and Paul R Bohjanen. Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res*, 30(24):5529–5538, Dec 2002.

[204] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S. Marshall, and A. Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22:326–330, 2004.

[205] P. Rollini and R.E. Fournier. The HNF–4/ HNF–1alpha transactivation cascade regulates gene activity and chromatin structure of the human serine protease inhibitor gene cluster at 14q32.1. *Proceedings of the National Academies of Science — U.S.A.*, 96(18):10308-10313, 1999.

[206] C. Des Rosiers, L. Di Donato, B. Comte, A. Laplante, C. Marcoux, F. David, C.A. Fernandez, and H. Brunengraber. Isotopomer analysis of citric acid cycle and gluconeogenesis in rat liver. reversibility of isocitrate dehydrogenase and involvement of ATP–citrate lyase in gluconeogenesis. *Journal of Biological Chemistry*, 270:10027-10036, 1995.

[207] R. Rosipal, L.J. Trejo, and B. Matthews. Kernel PLS-SVC for linear and nonlinear classification. In *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, D.C., 2003.

[208] Hong Ruan, Nir Hacohen, Todd R. Golub, Luk Van Parijs, and Havey F. Lodish. Tumor necrosis factor-α suppresses adipocyte-specific genes and activates expression of preadipocyte genes in 3T3-L1 adipocytes. *Diabetes*, 51:1319–1336, 2002.

[209] Douglas A Rubinson, Christopher P Dillon, Adam V Kwiatkowski, Claudia Sievers, Lili Yang, Johnny Kopinja, Dina L Rooney, Melanie M Ihrig, Michael T McManus, Frank B Gertler, Martin L Scott, and Luk Van Parijs. A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nat Genet*, 33(3):401–406, Mar 2003.

[210] C. Sabatti, L. Rohlin, M.K. Oh, and J.C. Liao. Co–expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, 30(13):2886–2893, 2002.

[211] A Sacchetti, T El Sewedy, A F Nasr, and S Alberti. Efficient GFP mutations profoundly affect mRNA transcription and translation rates. *FEBS Lett*, 492(1-2):151–155, Mar 2001.

[212] A.A. Sandberg, R.M. Gemmill, B.K. Hecht, and F. Hecht. The philadelphia chromosome: A model of cancer and molecular cytogenetics. *Cancer Genetics and Cytogenetics*, 21(2):129–146, 1986.

[213] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academies of Science — U.S.A.*, 74(12):5463–5467, 1977.

[214] Peter C Scacheri, Orit Rozenblatt-Rosen, Natasha J Caplen, Tyra G Wolfsberg, Lowell Umayam, Jeffrey C Lee, Christina M Hughes, Kalai Selvi Shanmugam, Arindam Bhattacharjee, Matthew Meyerson, and Francis S Collins. Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. *Proc Natl Acad Sci U S A*, 101(7):1892–1897, Feb 2004.

[215] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[216] Michaela Scherr, Karin Battmer, Arnold Ganser, and Matthias Eder. Modulation of gene expression by lentiviral-mediated delivery of small interfering RNA. *Cell Cycle*, 2(3):251–257, May 2003.

[217] W.A. Schmitt, R.M. Raab, and G. Stephanopoulos. Elucidation of gene interaction networks through time–lagged correlation analysis of transcriptional data. *Genome Research*, 14:1654–1663, 2004.

[218] R.A. Segal and M.E. Greenberg. Intracellular signalling pathways activated by neurotrophic factors. *Annual Reveiws in Neuroscience*, 19:463–489, 1996.

[219] P.O. Seglen. Preparation of isolated rat liver cells. *Methods in Cell Biology*, 13:29–83, 1976.

[220] D. Shalon, S.J. Smith, and P.O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6:639–645, 1996.

[221] H. Shi, N. Chamond, C. Tschudi, and E. Ullu. Selection and characterization of RNA interference–deficient trypanosomes impaired in target mRNA degradation. *Eukaryotic Cell*, 3(6):1445–1453, 2004.

[222] Y. Shi. Mammalian RNAi for the masses. *Trends in Genetics*, 19:9–12, 2003.

[223] K. Shibata and *et. al.* RIKEN integrated sequence analaysis (RISA) system—384–format sequencing pipeline with 384 multicapillary sequencer. *Genome Research*, 10(11):1757–1771, 2000.

[224] D.P. Siderovski, B. Strockbine, and C.I. Behe. Whither goest the RGS proteins? *Critical Reviews in Biochemistry and Molecular Biology*, 34:215–251, 1999.

[225] Srikumar Sinnarajah, Carmen W. Dessauer, Deepa Srikumar, Jun Chen, John Yuen, Solomon Yilma, John C. Dennis, Edward E. Morrison, Vitaly Vodyanoy, and John H. Kehrl. RGS2 regulates signal transduction in olfactory neurons by attenuating activation of adenylyl cyclase III. *Nature*, 409:1051–1055, 2001.

[226] Paul Smolen, Douglas A Baxter, and John H Byrne. Reduced models of the circadian oscillators in Neurospora crassa and Drosophila melanogaster illustrate mechanistic similarities. *OMICS*, 7(4):337–354, Winter 2003.

[227] R. Somogyi and S. Fuhrman. Distributivity, a general information theoretic network measurement, or why the whole is more than the sum of its parts. In *The International Workshop on Information Processing in Cells and Tissues*, Sheffield, UK, 1997.

[228] G.D. Sorenson, O.S. Pettengill, T. Brinck-Johnson, C.C. Cate, and L.H. Maurer. Hormone production by cultures of small–cell carcinoma of the lung. *Cancer*, 47(6):1289–1296, 1981.

[229] Edwin M. Southern. *DNA Arrays, Methods and Protocols*, volume 170 of *Methods in Molecular Biology*, chapter DNA Microarrays: History and Overview, pages 1–15. Humana Press, Totowa, NJ, March 2001.

[230] Gregory Stephanopoulos, Daehee Hwang, William A. Schmitt, Jatin Misra, and George Stephanopoulos. Mapping physiological states from microarray expression measurements. *Bioinformatics*, 18(8):1054–1063, 2002.

[231] M.P. Stern, M. Rosenthal, S.M. Haffner, H.P. Hazuda, and L.J. Franco. Sex difference in the effects of sociocultural status on diabetes and cardiovascular risk factors in mexican americans: the san antonio heart study. *American Journal of Epidemiology*, 120:834–851, 1984.

[232] C. Strohm, M. Barancik, M.L. von Bruehl, M. Strniskova, C. Ullmann, R. Zimmermann, and W. Schaper. Transcription inhibitor actionmycin–d abolishes

the cardioprotective effect of ischemic reconditioning. *Cardiovascular Research*, 55:602–618, 2002.

[233] T. Strother, W. Cai, X. Zhao, R.J. Hamers, and L.M. Smith. Synthesis and characterization of DNA-modified silicon (111) surfaces. *Journal of the American Chemical Society*, 122:1205–1209, 2000.

[234] R.S. Surwit, C.M. Kuhn, C. Cochrane, J.A. McCubbin, and M.N. Feinglos. Diet-induced type II diabetes in C57BL/6J mice. *Diabetes*, 37(9):1163–1167, 1988.

[235] K Takaya, Y Ogawa, N Isse, T Okazaki, N Satoh, H Masuzaki, K Mori, N Tamura, K Hosoda, and K Nakao. Molecular cloning of rat leptin receptor isoform complementary DNAs–identification of a missense mutation in Zucker fatty (fa/fa) rats. *Biochem Biophys Res Commun*, 225(1):75–83, Aug 1996.

[236] P. Tamayo, D. Slonim, J. Mesirov, J. Zgu, Q. Kitareewan, S. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self–organizing maps: Methods and applications to homeopoietic differentiation. *Proceedings of the National Academies of Science — U.S.A.*, 96:2907–2912, 1999.

[237] Y.X. Tao and D.L. Segaloff. Functional characterization of melanocortin–4 receptor mutations associated with childhood obesity. *Endocrinology*, 144:4544–4551, 2003.

[238] S. Taylor, S. Smith, B. Windle, and A. Guiseppi-Elie. Impact of surface chemistry and blocking strategies on DNA microarrays. *Nucleic Acids Research*, 31(16):E87, 2003.

[239] Scott A Tenenbaum, Craig C Carson, Ulus Atasoy, and Jack D Keene. Genome-wide regulatory analysis using en masse nuclear run-ons and ribonomic profiling with autoimmune sera. *Gene*, 317(1-2):79–87, Oct 2003.

[240] J. Theilhaber, T. Connolly, S. Roman-Roman, S. Bushnell, A. Jackson, K. Call, T. Garcia, and R. Baron. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Research*, 12(1):165–76, 2002.

[241] Jeffrey G. Thomas, James M. Olson, Stephen J. Tapscott, and Lue Ping Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11:1227–1236, 2001.

[242] M. Tijsterman and R.H.A. Plasterk. Dicers at RISC: The mechanism of RNAi. *Cell*, 117:1–4, 2004.

[243] Kazuyuki Tobe, R.Y.O. Suzuki, Masashi Aoyama, and Yasuo Terauchi. Increased expression of SREBP-1 gene in IRS-2(-/-) mice liver. *Diabetes*, 50(Supplement 2):A324–A324, 2001.

[244] J Torchia, C Glass, and M G Rosenfeld. Co-activators and co-repressors in the integration of transcriptional responses. *Curr Opin Cell Biol*, 10(3):373–383, Jun 1998.

[245] G.C. Tseng, M.K. Oh, L. Rohlin, J.C. Liao, and W.H. Wong. Issues in cDNA microarray analysis: Quality filtering, channel normailzation, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29(12):2549–2557, 2001.

[246] Y.O. Uehara, C. Minowa, K. Mori, J. Shiota, T. Kuno, and N. Kitamura. Placental defect and embryonic lethality in mice lacking hepatocyte growth factor/ scatter factor. *Nature*, 373:702–705, 1995.

[247] Kumiko Ui-Tei, Yuki Naito, Fumitaka Takahashi, Takeshi Haraguchi, Hiroko Ohki-Hamazaki, Aya Juni, Ryu Ueda, and Kaoru Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res*, 32(3):936–948, 2004. Guideline.

[248] I. Uphues, Y. Chern, and J. Eckel. Insulin-dependent translocation of the small GTP-binding protein RAB3C in cardiac muscle: Studies on insulin-resistant zucker rats. *FEBS Letters*, 377(2):109–112, 1995.

[249] H Vaucheret, C Beclin, and M Fagard. Post-transcriptional gene silencing in plants. *J Cell Sci*, 114(Pt 17):3083–3091, Sep 2001.

[250] Frédérique Verdier, Taras Valovka, Alexander Zhyvoloup, Ludmila B. Drobot, Vladimir Buchman, Mike Waterfiedl, and Ivan Gout. Ruk is ubiquitinated but not degraded by the proteasome. *European Journal of Biochemistry*, 269:3402–3408, 2002.

[251] S A Verploegen, G Plaetinck, R Devos, J Van der Heyden, and Y Guisez. A human leptin mutant induces weight gain in normal mice. *FEBS Lett*, 405(2):237–240, Mar 1997.

[252] Tom Volkert. Unpublished correspondance. Personal communication, 2001.

[253] L. De Vries, B. Zheng, T. Fischer, E. Elenko, and M.G. Farquhar. The regulator of G protein signalling family. *Annual Reviews in Pharmacology and Toxicology*, 40:235–271, 2000.

[254] V. Wallenius, K. Wallenius, B. Ahren, M. Rudling, H. Carlsten, S.L. Dickson, C. Ohlsson, and J.O. Jansson. Interleukin–6–deficient mice develop mature–onset obesity. *Nature Medicine*, 8(1):75–79, 2002.

[255] F Wang, R M Raab, M W Washabaugh, and B C Buckland. Gene therapy and metabolic engineering. *Metab Eng*, 2(2):126–139, Apr 2000.

[256] Y. Wang, T. Rea, J. Bian, S. Gray, and Y. Sun. Identification of the genes responsive to etoposide-induced apoptosis: Application of DNA chip technology. *FEBS Letters*, 445:269–273, 1999.

[257] A. Watson, A. Mazumder, M. Stewart, and S. Balasubramanian. Technology for microarray analysis of gene expression. *Current Opinion in Biotechnology*, 9:609–614, 1998.

[258] Affymetrix Web-site. http://www.affymetrix.com/products/arrays/specific /hgu133.affx. URL, 2005.

[259] Human Genome Project Web-site. http://www.ornl.gov/sci/techresources/ human_genome/faq/seqfacts.shtml#whatis. URL, 2005.

[260] Human Genome Project Web-site. http://www.ornl.gov/sci/techresources/ human_genome/project/about.shtml. URL, 2005.

[261] Human Genome Project Web-site. http://www.ornl.gov/sci/techresources/ human_genome/publicat/hgn/v7n3/04progre.shtml#sequencing. URL, 2005.

[262] Human Genome Project Web-site. http://www.ornl.gov/sci/techresources/ human_genome/research/instrumentation.shtml. URL, 2005.

[263] S.M. Weissman. Molecular genetic techniques for mapping the human genome. *Molecular Biology and Medicine*, 4(3):133–143, 1987.

[264] B.L. Welch. The generalization of student's problem when several populations are involved. *Biometrika*, 34:28–35, 1947.

[265] C.C. Wetzel, S.J.F. Degen, and S.E. Waltz. *cis*–acting elements in the hepatocyte growth factor–like protein gene regulate kidney and liver–specific expression in mice. *DNA and Cell Biology*, 22(5):293–301, 2003.

[266] D.L. Wheeler, D.M. Church, S. Federhen, A.E. Lash, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.A. Tatusova, and L. Wagner. Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, 31(1):28–33, 2003.

[267] Morris F. White and Martin G. Myers Jr. *Diabetes Mellitus, Carbohydrate Metabolism, and Lipid Disorders.* Harcourt Canada, Canada, 2000.

[268] C J Wilusz, M Wormington, and S W Peltz. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol*, 2(4):237–246, Apr 2001.

[269] M.S. Wong, R.M. Raab, I. Rigoutsos, G.N. Stephanopoulos, and J.K. Kelleher. Metabolic and transcriptional patterns accompanying glutamine depletion and repletion in mouse hepatoma cells: A model for physiological regulatory networks. *Physiological Genomics*, 16:247–255, 2004.

[270] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[271] Y.H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3:579–588, 2002.

[272] T. Yonezawa, A. Ohtsuka, T. Yoshitaka, S. Hirano, H. Nomoto, K. Yamamoto, and Y. Ninomiya. Limitrin, a novel immunoglobulin superfamily protein localized to glia limitans formed by astrocyte endfeet. *Glia*, 44:190–204, 2003.

[273] So youn Kim, Ha il Kim, Sang-Kyu Park, Seung-Soon Im, Tianzhu Li, Hyae Gyeong Cheon, and Yong ho Ahn. Liver glucokinase can be activated by peroxisome proliferator-activated receptor-gamma. *Diabetes*, 53(Supplement 1):S66–S70, 2004.

[274] M. Zechmeister-Machhart, P. Hufnagl, P. Uhrin, I. Korschineck, B.R. Binder, and M. Geiger. Molecular cloning and sequence analysis of the mouse protein C inhibitor gene. *Gene*, 186:61–66, 1997.

[275] Yan Zeng and Bryan R Cullen. Sequence requirements for micro RNA processing and function in human cells. *RNA*, 9(1):112–123, Jan 2003.

[276] S. Zhang and H.K. Gershenfeld. Genetic contributions to body weight in mice: Relationship of exploratory behavior to weight. *Obesity Research*, 11(7):828–838, 2003.

[277] Paul Zimmet, K.G.M.M. Alberti, and Jonathan Shaw. Global and societal implications of the diabetes epidemic. *Nature*, 414:782–787, 2001.

[278] M. Ziotopoulou, C.S. Mantzoros, S.M. Hileman, and J.S. Flier. Differential expression of hypothalamic neuropeptides in the early phase of diet–induced obesity in mice. *American Journal of Physiology: Endocrinology and Metabolism*, 279(4):E838–E845, 2000.