SYSTEMS APPROACH TO ANALYZING THE TGFβ/SMAD3 GENE REGULATORY PATHWAY IN A549 CELLS

by

Daniel Edward Handley

B.A., Biophysics, The Johns Hopkins University, 1990

M.S., Logic and Computation, Carnegie Mellon University, 2002

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Daniel Edward Handley

It was defended on

March 24, 2008

and approved by

Susanne Gollin, Ph.D., Professor, Department of Human Genetics,
Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, Ph.D., Associate Professor, Department of Human Genetics,
Graduate School of Public Health, University of Pittsburgh

Panayiotis Benos Ph.D., Assistant Professor, Department of Computational Biology,
School of Medicine, University of Pittsburgh

Dissertation Advisor: Naftali Kaminski, M.D., Associate Professor of Medicine,
Pathology, Human Genetics and Computational Biology, Division of Pulmonary,
Allergy, and Critical Care Medicine, University of Pittsburgh Medical Center

SYSTEMS APPROACH TO ANALYZING THE TGFβ/SMAD3 GENE REGULATORY PATHWAY IN A549 CELLS

Daniel Edward Handley, Ph.D.

University of Pittsburgh, 2008

Public Health Significance: Idiopathic pulmonary fibrosis (IPF) is a devastating lung disease affecting over 100,000 people every year in the U.S. There is no prevention, cure, or effective treatment for the disease, and the life expectancy after diagnosis is about 3 years. The disease is characterized by progressive and irreversible deposition of fibrotic proteins in the lung. The etiology of the disease is poorly understood, but there is abundant evidence the pro-fibrotic cytokine transforming growth factor beta (TGFβ) plays a major role in the disease process. TGFβ acts principally through the DNA-binding transcription factor SMAD3. The research presented here may lead directly to new pharmacological interventions for IPF, thereby substantially decreasing morbidity and mortality rates for the disease.

To gain new insights into how the TGFβ/SMAD3 transcriptional regulatory pathway might promote pulmonary fibrosis, I combined high-throughput molecular biology measurements with systems biology computational tools to study transcriptional regulation of the TGFβ/SMAD3 pathway in A549 alveolar epithelial cells. The first tier of measurement consisted of chromatin immunoprecipitation combined with whole-genome promoter microarrays (ChIP-on-chip). This technique globally identifies the promoter regions of genes bound by the SMAD3 transcription factor. A second tier of systems-wide information consisted of whole-genome gene expression microarrays, which measures levels of mRNA activated by the TGFβ/SMAD3

pathway. These two tiers of transcription information were integrated and analyzed using systems biology computational tools. The analysis yielded three novel findings. The first is that the TGFβ/SMAD3 pathway transcriptionally regulates transgelin, a protein that signifies the TGFβ-induced transition of epithelial cells into collagen-secreting myofibroblasts. The second is that the TGFβ/SMAD3 pathway also regulates the transcription factor FOXA2, which plays a major role in lung development and surfactant production. The third is possible TGFβ/SMAD3 transcriptional regulation of PINX1, which is a potent suppressor of telomere reverse transcriptase (hTERT). All three of these proteins are mechanistically linked to genes or processes that are already suspected of being involved in the pathophysiology of IPF. Thus, a systems-level approach to studying transcriptional regulatory networks is a valuable tool for discovering new biological pathways or new connections between known biological pathways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would also like to extend my sincerest thanks to my advisor, Naftali Kaminski, M.D., as well as my committee members: Eleanor Feingold, Ph.D., Susanne Gollin, Ph.D., and Takis Benos, Ph.D. Any substantial body of work is likely to be the result of numerous formal and informal collaborations over time, and this work is no exception. Those individuals who have helped me through a combination of illuminating discussions and technical assistance include: Ivan Rosas, M.D., Tom Richards, Ph.D., Stefan Ryter, Ph.D., Yingze Zhang, Ph.D., Paul Reynolds, Ph.D., David Peters, Ph.D., Kevin Gibson, M.D., Oliver Eickelberg, M.D., Tommy Kaplan, Ph.D., Nir Friedman, Ph.D., Guoying Yu, Ph.D., Ansuman Chattopadhyay, Ph.D., Carol Feghali-Bostwick, Ph.D., David Corcoran, Jessica Dominick, Elisa Heinrich O'Hare, Lara Chensny, Kazu Kazuhisa, M.D., Megan LeJeune, Louis Vuga, M.D., Kusum Pandit. I am very appreciative of each of these individual's help and friendship during this arduous process.

1 INTRODUCTION

1.1 PUBLIC HEALTH SIGNIFICANCE

Idiopathic pulmonary fibrosis (IPF) is a chronic and fatal interstitial lung disease affecting over 100,000 people every year in the United States alone [1, 2]. It is characterized by the progressive and irreversible replacement of healthy, flexible lung tissue with stiff, fibrous proteins. It is a severely debilitating disease for which there is currently no prevention, cure, or effective therapy [2-8]. The life expectancy of IPF patients from time of diagnosis is three to four years, and about 50% of patients will die within 2 to 3 years after diagnosis [2, 9-11]. The etiology of IPF is poorly understood, but it appears some complex combination of genetic and environmental factors play a role. The central issue in understanding the etiology and pathophysiology of IPF is elucidating specific genomic and molecular mechanism(s) responsible for the abnormal cellular behavior that results in pulmonary fibrosis. Further, the research presented here may lead directly to new pharmacological interventions for IPF, thereby substantially decreasing morbidity and mortality rates for the disease.

New technologies—particularly sophisticated computational tools and massively parallel (high-throughput) biological measurements—allow for simultaneous global discovery of interconnections and functional pathways between thousands of genes and the proteins they encode. This holistic, top-level approach to biological investigation is known as systems biology.

It promises to provide biological insights not readily discernable using traditional single gene, single protein experimentation.

In this study, a systems-level approach was applied to a cell culture model system to better understand the etiology and molecular pathophysiology of IPF. One conclusion from this exercise is that a systems-level approach is a valuable tool for gaining new insights into the molecular basis for complex diseases, and therefore is well-suited for studying IPF and other complex diseases of great public health significance.

1.2 OVERVIEW OF THE APPROACH USED IN THIS STUDY

Multiple environmental, aging, and genetic factors likely play causal roles in the etiology of IPF. However, in the pathophysiology of IPF, one of the most robust findings is the involvement of transforming growth factor beta (TGFβ). TGFβ is a major anti-inflammatory cytokine that is well-established as inducing epithelial cells to convert to fibroblasts, and then inducing fibroblasts to secrete fibrotic proteins [12-16]. This is the case both in the context of normal wound repair and in abnormal fibrotic organ diseases. TGFβ regulates transcription of particular genes principally through the DNA-binding transcription factor SMAD3. Although much is known about the TGFβ/SMAD3 signal transduction proteins on the cell surface and in the cytoplasm, little is still known about what and how genes are transcriptionally regulated in the nucleus [17-24]. Therefore, it is clear TGFβ plays a major role in IPF, and the transcription factor SMAD3 is a major mediator of TGFβ signaling to the nucleus. But exactly what genes SMAD3 activates and how that pathway might be responsible for abnormally depositing fibrotic proteins in the lung is still unclear.

By employing a systems biology approach to studying transcriptional regulation in pulmonary epithelial cells at the global level, the hope is to gain new insights into what the TGFβ/SMAD3 pathway might be doing to promote fibrotic deposition. The first such systems-level approach used in this work is a high-throughput measurement technique that combines chromatin immunoprecipitation with promoter microarrays (ChIP-on-chip). The principle behind ChIP is that a specific transcription factor (in this case SMAD3) can be isolated from the nucleus along with a portion of each of the promoter regions in the DNA to which it is bound. By amplifying these short stretches of promoter-region DNA, labeling them with a fluorescent dye, and then hybridizing them to a promoter microarray, we can tell what gene promoter regions had been originally bound by SMAD3. By identifying what gene promoter regions had been bound by SMAD3, we then know which gene coding regions are potentially regulated by the SMAD3 transcription factor. Therefore, ChIP-on-chip gives us a method for identifying from among the entire cell genome which subset of genes might be transcriptionally regulated specifically through the TGFβ/SMAD3 pathway.

A second systems-level approach used in the present work is the gene expression microarray. Gene expression microarrays are firmly established in all areas of life science research as a method for globally identifying up- and down-regulated mRNA levels from cells [25-30]. In this case, gene expression microarrays provides a second tier of information to complement that from ChIP-on-chip. While ChIP-on-chip provides information on which gene promoters are bound by SMAD3, gene expression microarrays tell us what actually happens to transcript levels of each gene—whether gene transcription is activated, repressed, or whether there is no actual change in mRNA transcription level at all.

Finally, these high-throughput technologies would be of little use unless we can evaluate the information in a proper biological context. The third element of the systems-level approach I applied in this work is therefore the use of a number of sophisticated computational methods, including gene functional network discovery tools [31-33]. We first use computational biology tools to process the raw information from promoter and gene expression microarrays. The array images must be converted to computational data structures containing the biologically-relevant measurement data. Since all measurements in any scientific context necessarily contain both systematic and random error, the microarray data must then be evaluated statistically to identify significantly bound, up-regulated, and down-regulated genes. Since potentially hundreds or thousands of genes are likely to be identified, this information must be further evaluated with the assistance of computational tools. First, genes can be grouped in terms of their similar behaviors (up- or down-regulation) and/or their known functions by statistical algorithms that cluster them according to their degree of behavior or functional similarity [34]. Using massive databases that contain curated information about already-known transcriptional networks, signaling pathways, and molecular functions of genes and gene products, our identified gene lists can be further mapped according to their participation in these known pathways [32, 33]. This allows us to (1) verify that our data agree with what functions they are known to participate in—this serves to give us confidence in the quality of our experimental techniques, and (2) discover connections of our identified gene groups with previously unknown functions or pathways. Thus, a systems-level approach to studying transcriptional regulatory networks promises to discover new biological pathways or connections between biological pathways not previously known.

1.3 OVERVIEW OF FINDINGS

Using the approaches described above yielded three specific findings. One is the identification of TGFβ/SMAD3 regulation of a gene called transgelin, whose protein participates in the transformation of pulmonary epithelial cells into fibrotic-protein-secreting myofibroblasts [35]. The second is the identification of a connection between the TGFβ/SMAD3 pathway and the forkhead winged helix transcription factor FOXA2. FOXA2 is an important transcriptional regulator necessary for both proper lung development during embryogenesis and for proper lung function in maturity [36-40]. A key function of FOXA2 is the transcriptional regulation of surfactant proteins, which play a major role in lung health and whose dysregulation can result in pulmonary fibrosis [41-45]. Finally, this study identified a novel connection between the TGFβ/SMAD3 pathway and the *PinX1* gene, whose protein product is a potent inhibitor of telomerase (*hTERT*) [46-49]. PinX1-mediated inhibition of *hTERT* can induce apoptosis in cells as well as limit their ability to appropriately proliferate in response to tissue damage [46, 48, 49]. Further, loss-of-function mutations in *hTERT* have been reported in familial forms of pulmonary fibrosis, which strongly implicates the involvement of *hTERT* in IPF [50, 51]. All three of these findings, the result of a systems-level experimental approach, link specific targets of the TGFβ/SMAD3 transcriptional regulatory pathway with the pathogenesis of IPF.

Additionally, at the global level, this study demonstrated the functional pathways involved in the TGFβ–induced transition of A549 alveolar epithelial cells toward cells that have the molecular characteristics of myofibroblasts. While the observation that TGFβ induces A549 cells to undergo epithelial-mesenchymal transition was published in 2007 [13], this is the first study to globally show the functional groups and networks of genes involved in the process.

1.4 DOCUMENT ORGANIZATION

The study described here is the result of the integration of diverse technologies and approaches. There is also a large amount of background material to discuss to be able to put the study in its proper biological context. Therefore this document necessarily spans a number of varied topics. The following (second) chapter is an overview of what IPF is, how it is diagnosed and characterized, and some of the suspected environmental or genetic contributions to the disease process. The third chapter discusses in more detail what is known about the cytokine TGFβ itself, the SMAD3 protein, and how the TGFβ/SMAD3 signal transduction cascade operates. The fourth chapter takes a short detour from biology to discuss some of the methodological and (lightly) philosophical issues surrounding the systems biology approach to biological research. The fifth and sixth chapters discuss in relative detail the technologies used—the specific principles and limitations of ChIP-on-chip and gene expression microarrays, respectively. Chapter seven formally describes the methods used in the study. Chapter eight exhibits and discusses the specific findings. And finally, chapter nine discusses the importance and relevance of the findings in the context of IPF, as well as suggesting future avenues for research.

2 IDIOPATHIC PULMONARY FIBROSIS

Idiopathic pulmonary fibrosis (IPF) is a fatal lung disease characterized by excessive fibrotic protein deposition in the lungs. This abnormal fibrotic tissue significantly reduces the normal flexibility of the lung and interferes with exchange of O_2 and CO_2 in the alveolar spaces. IPF is a progressively debilitating disease for which there is currently no prevention, cure, or effective therapy [1, 2]. Moreover, the underlying cause(s) and cellular mechanisms responsible for the disease are poorly understood [2-8].

Figure 2-1. Severe end-stage pulmonary fibrosis in a lung taken from an autopsy performed in the 1980s [52].

IPF is a subset of a much larger group of lung disorders known as diffuse parenchymal lung diseases (DPLDs). These diseases broadly affect the interstitium, or tissues between and surrounding the alveolar spaces in the lung. Among the many different diseases that make up the category of DPLDs, many lung diseases involve the development of pulmonary fibrosis as a secondary outcome of the disease process. Some of these lung diseases stem from chronic occupational dust exposures, such as from asbestos, metallic particles, and silica dust [10, 53, 54]. Pulmonary fibrosis can also be a serious side effect of some pharmaceuticals such as the antiarrhythmic agents amiodarone and propanolol and the antibiotic nitrofurantoin [55, 56]. Similarly, some antineoplastic cytotoxic agents used in chemotherapy as well as oncological radiotherapy may induce pulmonary fibrosis [12, 57, 58]. Sarcoidosis, an inflammatory disease characterized by accumulation of granulomas in tissues, can also result in pulmonary fibrosis [10, 55]. Some autoimmune or connective tissue diseases, such as rheumatoid arthritis and systemic sclerosis may induce pulmonary fibrosis as well [10, 53, 55]. In all of these cases, pulmonary fibrosis is a secondary effect of some other disease process. In IPF, fibrosis is the primary effect of the disease.

IPF is the most severe of all of the DPLDs. The life expectancy of IPF patients from the time of diagnosis is three to four years, and about 50% of patients will die within 2 to 3 years after diagnosis [2, 9-11]. Although a number of drug treatments are commonly prescribed for IPF patients, none have been proven to be particularly effective against the disease [2, 7, 56]. The most effective measure is a single or double lobe lung transplantation for suitable recipients, which obviously has severe practical limitations and is certainly not preferable as a first-line medical treatment [3]. At best, some of the disease symptoms can be managed, such as

delivering oxygen to help alleviate the dyspsnea and chronic hypoxia that are common in late stages of the disease [2].

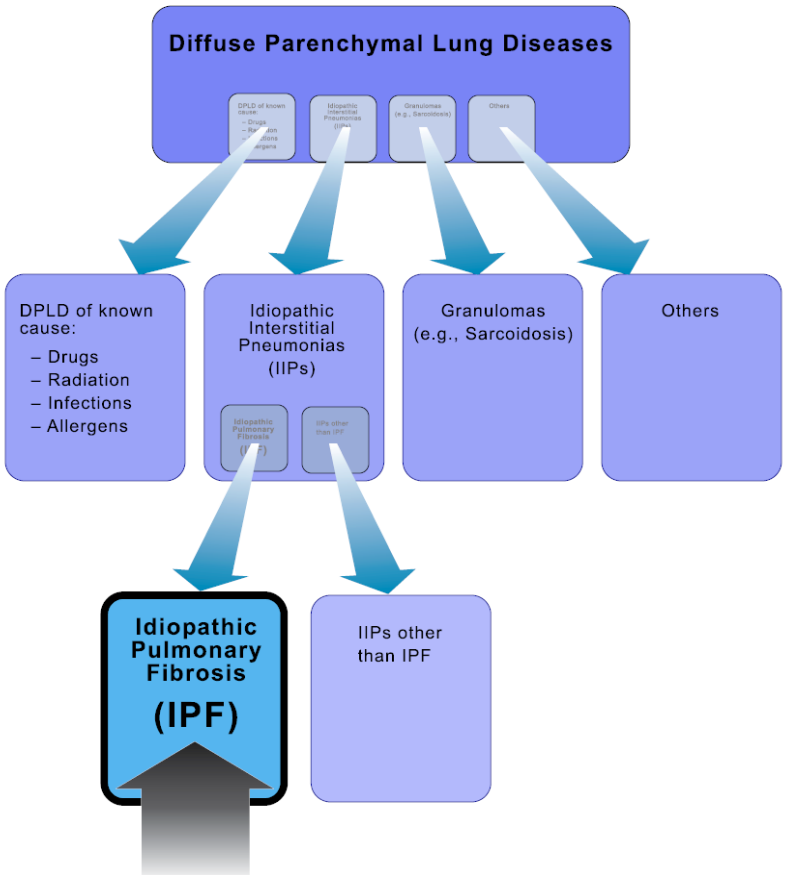


Figure 2-2. Relationship of IPF to other types of Diffuse Parenchymal Lung Diseases.

There are a number of potential risk factors for IPF, however, none of them have been clearly identified as unequivocal causes of IPF. Current or former cigarette smoking has the strongest association with IPF, with an estimated odds ratio of between 1.6 and 2.9 [2, 56, 59, 60]. Exposure to infectious viruses such as Epstein-Barr virus, cytomegalovirus, influenza, and hepatitis C is also suggested as a potential risk factor, however exposure to most of these viruses in the population is very common while only a tiny minority of those exposed will ever develop

IPF [2, 56]. In one study, 87% of IPF patients had gastroesophageal reflux disease (GERD), and it was postulated that chronic aspiration of stomach acid aerosol might play a role in the development of IPF [2, 53, 55, 61]. As in the case of exposure to common viruses, GERD is an extremely widespread diagnosis especially in those 40 years or older; the vast majority of those with GERD will never develop IPF. As of yet there have been no robust genetic associations with IPF (see Genetic Association with IPF section below) as well as no clear race, ethnicity, or geographic distribution [2, 57]. The strongest demographic associations continue to be with age (>40 years) and gender, with an approximately 60%-40% split in favor of males [2, 57]. IPF appears to be a complex disease involving numerous partial ultimate causes that might include pathogenic exposures, polygenic factors, environmental exposures, age, and others not yet identified.

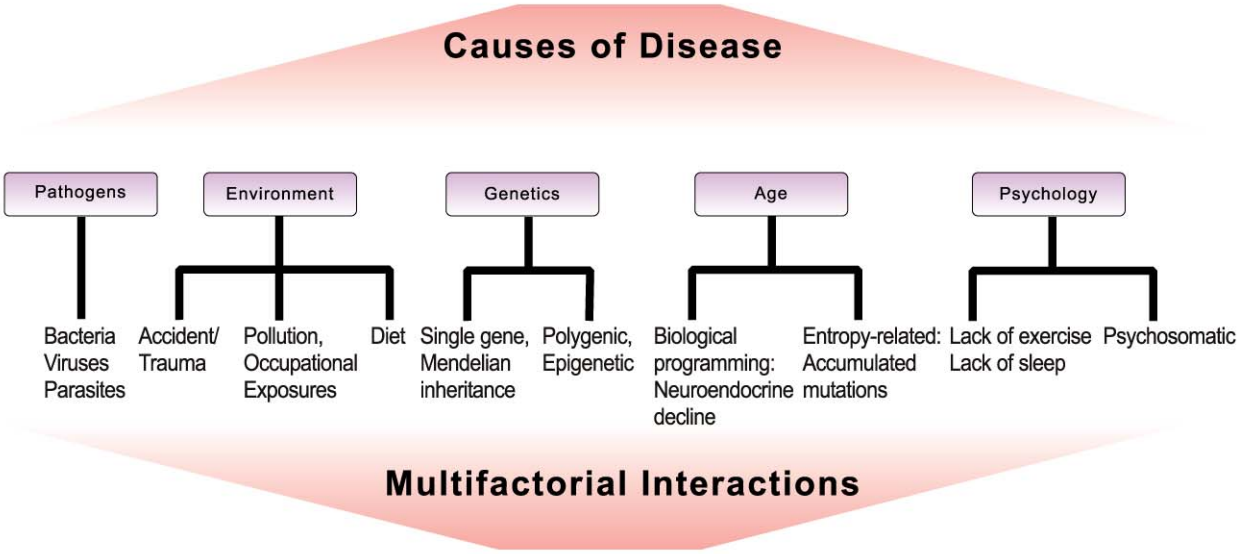


Figure 2-3. Multiple possible causes of disease making for complex interactions.

2.1 SYMPTOMS AND DIAGNOSIS

The most common initial symptom of pulmonary fibrosis is shortness of breath, especially upon exertion. This symptom can be associated with a great many other medical conditions, however, and is often ignored by the patient until it has progressed to a significant degree. Other symptoms include a dry cough or the presentation of "crackles" audible to a physician upon chest auscultation [2]. Some patients develop "clubbing" of the fingertips as a result of chronic oxygen deprivation [2]. Pulmonary fibrosis is often differentially diagnosed through a combination of non-invasive and invasive diagnostic tests. One of the least invasive diagnostic test procedures is spirometry, or pulmonary function testing, where such variables as lung capacity, mechanical compliance, and maximal air flow are measured [10, 62]. Chest x-rays and high-resolution CT scans are commonly used to look for observable, macroscopic changes in the structure of the lung [2, 56, 62]. Bronchoscopy and bronchoalveolar lavage (BAL; washing and sampling of lung spaces with saline) may be used to rule out other diagnoses, such as infections [10]. Finally, a lung biopsy can be used to provide a definitive diagnosis when other potential causes are ruled out [2, 10, 62].

At the histopathology level, idiopathic pulmonary fibrosis is characterized by the appearance of small foci of fibroblasts which deposit fibrotic proteins such as collagens in the interstitium. This fibrotic deposition progresses outwardly from these foci and cause portions of the lung to take on a characteristic "honeycomb" appearance that can be identified through lung X-ray or CT scan [10, 62]. The progressive replacement of healthy, flexible alveolar tissue with stiff fibrotic lesions begins to inhibit proper respiratory movement as well as severely interfere with gas exchange in the lung. Therefore, IPF patients will understandably experience dyspnea as a result of increasingly diminished lung function [56]. In situations in which pulmonary

fibrosis occurs and all known proximal causes (i.e., other DLPDs) are ruled out, the disease is then diagnosed as “idiopathic” pulmonary fibrosis.

The time course of IPF is extremely variable. Some patients have a stable period of disease progression interspersed with periods of acute exacerbations. Others experience a rapid deterioration after diagnosis [4, 63]. A sizable proportion of IPF patients develop pulmonary hypertension and eventually die of cardiovascular events or congestive heart failure; others will die of acute respiratory distress or other IPF-related complications [4, 10, 63, 64].

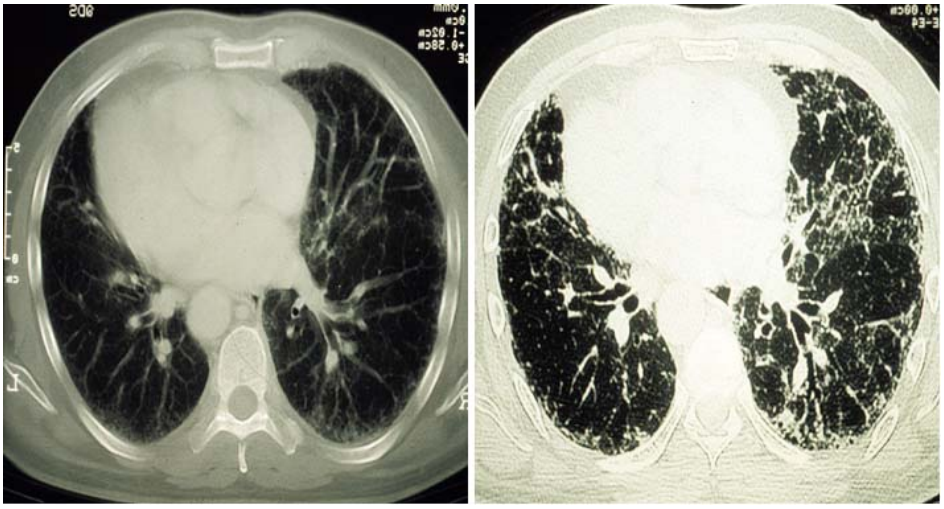


Figure 2-4. Patient CT scans showing progression of IPF after nine months (right)(courtesy Kevin Gibson, M.D.).

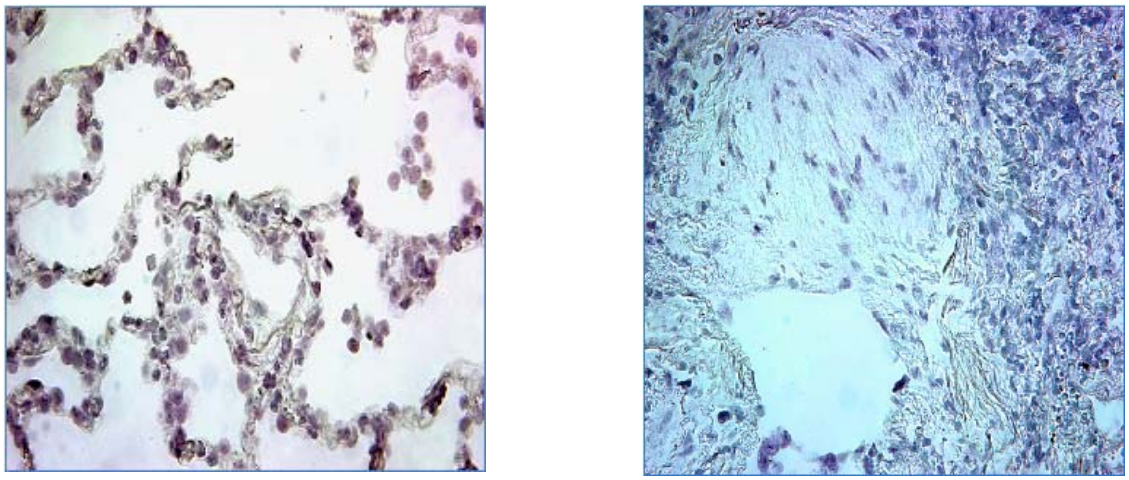


Figure 2-5. Histology, normal lung (left) versus IPF lung (right) showing fibrotic deposits (courtesy Naftali Kaminski, MD).

2.2 GENETIC ASSOCIATIONS WITH IPF

DPLDs in general, and IPF in particular, are relatively rare diseases and therefore difficult to study in the population. The vast majority of cases of IPF occur sporadically, i.e., they rarely cluster in families. IPF appears to be a complex polygenic disease likely involving multiple environmental interactions as well. Further, the relative contribution of many individual genes to the disease may each be small—it is the combinatory effect that may be most important [65]. Elucidating the genetic causes or predispositions for the disease is also complicated by the fact that until recently, there has been no consensus on a precise definition of the IPF phenotype.

Despite these difficulties, a number of studies have shown possible genetic factors involved in the pathogenesis of IPF. First, histologically identical cases of pulmonary fibrosis have been reported in monozygotic twins reared separately [66-68]. Pulmonary fibrosis has also been seen in the same generation of a single family as well as vertical transmission between generations [69-72]

Also, pulmonary fibrosis is associated with a number of well-defined genetic disorders, including Hermansky-Pudlack Syndrome, Gaucher Disease, Nieman-Pick Disease, and Familial Hypocalciuric Hypercalcemia [73-80]. This indicates some mode of inheritance of genes responsible for the development of pulmonary fibrosis.

In 2005 Steele *et. al.*, identified 111 families with 309 people affected and 360 people unaffected by idiopathic interstitial pneumonias (IIPs). They showed a statistically significant association for disease risk among sibling pairs ($p<0.001$). Twenty pedigrees also showed

vertical transmission of the disease, including three families with male-to-male transmission. These results are consistent with an autosomal dominant with incomplete penetrance inheritance model [81].

In addition to familial aggregation, several IIP studies point to specific gene mutations. One such gene, surfactant protein C (Sp-C; *SFTPC*), codes for a hydrophobic membrane protein produced by alveolar type II epithelial cells. It works with other surfactant proteins (A,B,D) to reduce surface tension in the aqueous alveolar space, allowing the alveoli to expand properly for gas exchange. Surfactants are required for proper lung function. In one case study, a c.460 G → A mutation in *SFTPC* and consequent defective Sp-C protein was detected in both a mother and daughter with IIPs [41, 42]. Similarly, Thomas et. al., identified an *SFTPC* mutation (Exon 5 + 128 A → T) in a familial cluster of IPF cases [44].

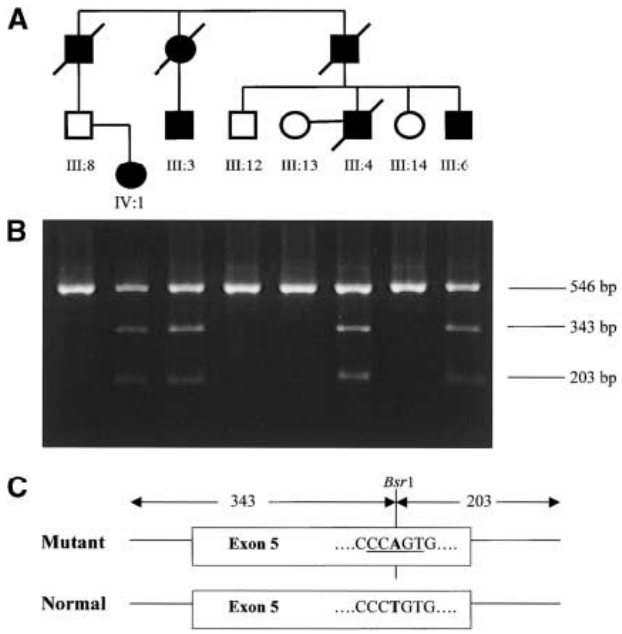


Figure 2-6. Pedigree and associated restriction analysis of *SFTPC* variants.
(A) pedigree showing affected and non-affected individuals. (B) Electrophoretic gel showing Bsr1 restriction digest gene fragment containing Exon 5 of the Sp-C gene. Note non-affecteds uniformly show a single band, whereas affected have three. (C) Diagram

showing restriction site in mutation but not normal Exon 5 SFTPC sequence (reproduced from Thomas, et. al. 2002) [44].

Telomere reverse transcriptase (*hTERT*) is another gene which has an association with IPF. Telomeres contain multiple repeats of the TTAGGG sequence at the ends of chromosomes. Some of these repeat sequences are lost with each cell division. The telomerase enzyme encoded by *hTERT*, along with the RNA component of telomerase (*hTERC*), mediates addition of TTAGGG to replace those lost during cell division. Without the proper telomerase activity, telomeres will shorten to the point that cells will fail to divide [82]. Tsakiri *et al.,* performed linkage analysis of 46 families with two or more individuals affected with IIPs. They identified *hTERT* as a candidate gene. Mutation analysis revealed a missense mutation and a frameshift mutation in *hTERT* that co-segregated with IIP in two families. Sequence comparisons of *hTERT* between 44 sporadic IPF cases and probands from 44 unrelated families revealed five other mutations. One family had a mutation in *TERC* as well. Those individuals with heterozygous mutations in *hTERT* or *hTERC* had shorter telomere lengths compared to controls [50, 51]. Thus, telomerase is another gene with a distinct association with IPF, and may also provide a clue to the overall profoundly increased risk of IPF with age.

Other genetic associations with IPF have been found with ELMO domain-containing protein 2 (*ELMOD2*) [83], variants of the surfactant proteins Sp-A and Sp-B (*SP-A, Sp-B*) [84-86], complement receptor 1 (CR1)[87], and interleukin-1 receptor antagonist (*IL-1ra*) and tumor necrosis factor α (*TNFα*) [88]. Each of these linkages and association with genetic mutations have been found in some studies but not others. In most cases of sporadic IPF, no genetic association has yet been found [43, 65, 88]. This suggests that of many possible genes involved

in the disease, few if any are both necessary and sufficient causes. However, the coincidence of multiple genetic, pathogenic, and environmental causes may constitute sufficiency.

2.3 EXTRACELLULAR MATRIX

In pulmonary fibrosis, the most visible evidence of the disease process takes place in the pulmonary interstitium, or the extracellular material in between and surrounding the alveolar sac. The study of how this abnormal, progressive fibrotic deposition occurs therefore centers largely around how the extracellular matrix, or ECM, is produced, maintained, remodeled, and degraded.

In multicellular animals, extracellular matrix (ECM) refers to material between cells that provides anchorage and mechanical support, as well as serving as mediators of intercellular communication [89, 90]. ECM is primarily made of fibrous structural proteins predominated by collagens, along with various polysaccharides and proteoglycans. Among the 28 different types of characterized collagens are fibrillar (Type I, II, III, V, XI), facit (Type IX, XII, XIV), short chain (Type VIII, X), basement membrane (Type IV) and others (Type VI, VII, XIII) [89, 90]. Other major structural proteins include fibronectin, elastin, and laminins [89, 90]. Among the proteoglycan components of ECM are heparin sulfate, keratin sulfates, and condroitin sulfates [90].

The ECM also includes a myriad of cell signaling molecules such as cytokines, cell adhesion molecules, proteases, growth factors and other molecules that interact with the cell surface receptors on the adjoining cell surfaces [90]. In cartilage, production and maintenance of ECM is performed by chondrocytes [90]. Similarly, in bone, osteoblasts are responsible for

matrix deposition and bone formation [90]. In the majority of tissue types including lung, fibroblasts predominantly produce and maintain ECM structural proteins [14, 91].

Far from being a passive structure, the ECM is a dynamic community of bioactive molecules, its ever-changing architecture the result of continuous protein deposition and active protein remodeling. Chief among these remodeling agents are proteolytic enzymes known as matrix metalloproteinases (MMPs). MMPs are zinc-dependent endopeptidases that catalyze the cleavage and/or degradation of various proteins found in the ECM [92-95]. Since the ECM consists of numerous fibrous structural proteins interwoven with various cytokines, receptors, and growth factors, catalytic cleavage of these bioactive molecules can in some circumstances inactive, or active, them. MMPs are therefore active regulators of both structural architecture and inter-cellular signaling. Among the 23 types of human MMPs characterized, each exhibits specific activity against particular substrates, e.g., collagenases, gelatinases, stromelysins, and membrane-type MMPs (MT-MMPs) [89, 90, 92, 93]. MMPs types also differ in their biological functions: some may cleave cell surface receptors and cell adhesion molecules, some release apoptotic ligands (such as the FAS ligand), and others activate or inactivate chemokines [92, 93]. MMPs therefore play a critical role in mediating cell behavior, such as cell proliferation, cell adhesion and migration, differentiation, angiogenesis, apoptosis and host defense [92, 93]. MMPs are themselves regulated by endogenous tissue inhibitors of metalloproteinases (TIMPs), consisting of a family of four protease inhibitors: TIMP-1, TIMP-2, TIMP-3 and TIMP-4 [92, 93].

Elucidating the complex dynamics of ECM structure and function is therefore extremely important in understanding a great variety of normal biological processes as well as pathological conditions. The interplay of structural and bioactive molecules in the ECM, and particularly

those of MMPs and TIMPs, are of particular interest in understanding the molecular pathophysiology of IPF [89, 96, 97]. In addition to being important in pulmonary fibrosis, these processes are also crucial to wound healing, host defense, embryogenesis and development, and angiogenesis, and tumor metastasis, migration, and metastatic anchorage [14, 92, 93, 95].

2.4 WOUND REPAIR, EPITHELIAL-MESENCHYMAL TRANSITION, AND FIBROTIC DEPOSITION

In mechanical wounding of the epidermis, the first immediate activity is coagulation of blood mediated by a complex cascade of blood clotting factors. The result of this clotting is a temporary mechanical stabilization of the wound by a hemostatic plug of a dense, strong, fibrous protein known as fibrin [90, 91]. Concurrently, an immediate localized inflammatory reaction occurs and produces chemoattractants that recruit monocytes and neutrophils to the area. This acute inflammatory reaction is later followed by the start of a repair phase in which acute inflammation is suppressed and fibrinolytic substances begin partially dissolving clotted fibrin. This allows epithelial cells and fibroblasts to migrate toward the area to begin the process of re-epithelialization [98-100].

In response to pro-repair cytokines and growth signals—most notably TGFβ_1 secreted by activated macrophages—the cytoskeletal structure of fibroblasts begins to change. Fibroblasts begin expressing proteins which give them some of the contractile properties of muscle cells (e.g., alpha smooth muscle actin), and are hence known as *myo*fibroblasts [12, 91, 101]. Similarly in response to TGFβ_1, some epithelial cells proliferate, dissociate from their basement membranes via MMP activity, transform into myofibroblasts, and migrate toward the wound

area. This cell type transformation is known as epithelial-mesenchymal transition, or EMT [12, 14, 98-103]. A third potential source of fibroblasts that may migrate to the injured area via the bloodstream is bone-marrow derived mesenchymal stem cells known as fibrocytes [12, 102, 104].

Being both motile and contractile, the myofibroblasts infiltrate the wounded area and anchor themselves. They then begin "pulling" the edges of the wound together [90, 91]. Myofibroblasts also actively deposit large quantities of collagens and other structural proteins that build up the ECM in and around the wounded area, resulting in a more long-term structural stabilization of the damaged epidermis. MMPs and TIMPs are also secreted which continually remodel and sculpt the fibrous ECM architecture [91, 100]. Normally, after sufficient ECM has been deposited to protect and stabilize a wound, the myofibroblasts are no longer needed and begin to undergo programmed cell death, or apoptosis [91]. Some of the ECM proteins continue to be degraded by MMPs and are resorbed, allowing epithelial cells to migrate into the area, proliferate, and re-epithelialize the previously damaged areas [98-100]. However, when there is excessive deposition of fibrotic material, or the material is not properly degraded and removed, the result is a prominent residual "patch" of fibrous structural proteins we commonly refer to as scar tissue. $TGF\beta_1$ is a major mediator of all of these repair activities, most notably fibroblast and myofibroblast recruitment, EMT, secretion of collagens, and upregulation and secretion of various MMPs and TIMPs [12, 89-91, 105].

In idiopathic pulmonary fibrosis, no such mechanical wounding is evident that would account for either an acute inflammatory phase or a repair phase. The major feature of idiopathic pulmonary fibrosis is uncontrolled, aberrant "wound repair" in the lung where no such repair is apparently necessary. As in the case of mechanical wound healing of epidermis, however,

fibroblasts and myofibroblasts proliferate and migrate to the area. Similarly, epithelial cells undergo EMT, and migrate to the area as myofibroblasts as well. Deposition of collagens and other fibrotic proteins, which in other tissues would serve to mechanically stabilize and repair a wound, proves to be a serious problem in the interstitium around alveolar spaces. Again, $TGF\beta_1$ is a known major mediator of these events and as discussed earlier is present in active form in high concentrations in active IPF foci [13, 91, 106, 107].

For many years the prevailing hypothesis was that IPF must result from some form of chronic inflammation in the lung [8, 10, 99, 100, 108]. However, there is abundant evidence that the etiology of IPF is not quite so simple and that chronic inflammation is unlikely to be its primary cause. First, in histopathological examination of tissue samples taken from IPF patients, there is typically little evidence of inflammation [8, 10, 100, 107]. Similarly, markers of chronic inflammation in IPF patients do not correlate with severity of IPF progression or with deaths. Conversely, in cases of known interstitial lung diseases involving chronic inflammation, the cases do not often progress toward fibrosis [8]. Finally, although anti-inflammatory and immune-suppressing drugs have been traditionally given to IPF patients for many years, these drugs have been shown to have no effect on improving disease outcomes or reducing deaths [1, 8, 108].

More recently, the prevailing ideas about the etiology of IPF surround some as-yet unknown mechanism that induces alveolar epithelial cell damage, followed by pro-fibrotic mechanisms [56]. Based on transgenic animal models and microarray studies, a number of cytokine signaling pathways and cellular functions have been proposed as being involved in the pathogenesis of IPF. Among these are transforming growth factor beta-1 ($TGF\beta_1$), tumor necrosis factor ($TNF\alpha$), angiogenesis and cell recruitment, coagulation cascades, apoptosis, lipid metabolism, and expression of multiple regulatory molecules in the alveolar epithelial cells

[109]. Of these possible key mediators of fibrosis, a significant amount of evidence points toward a major role of the cytokine TGFβ_1. The evidence and potential role of this prominent signaling molecule is discussed in the following chapter.

2.5 EXPERIMENTAL METHODS COMMON IN IPF RESEARCH

There are several general approaches and models common in IPF research. Like most biomedical research, experiments are often conducted using cell culture models, animal models, and human (clinical and post-mortem) studies.

Cell cultures of lung epithelial cells and fibroblasts are often used to study cell signaling and transcriptional regulation of key factors hypothesized to be involved in IPF pathogenesis. While primary human lung fibroblasts are relatively easy to grow in culture, primary human epithelial cells, and in particular Type II alveolar epithelial cells, are very difficult if not impossible to culture and expand in the laboratory. A less desirable but more practical alternative is using transformed or epithelial cancer cell lines, since they are much easier to maintain and grow up in quantities necessary for most molecular biology techniques. Single genes or transcription factors of interest can be knocked out using small interfering RNA (siRNA) or specific molecular pathway inhibitors.

Animal models—most commonly mice, rats, and hamsters—are also often used in IPF research [110]. Transgenic animals, particularly those with homozygous deletions for key genes of interest, are commonly used to study the effect of single gene knockouts [111]. Adenoviral-mediated gene transfer is also used to over-express certain gene products [112].

The most common animal model of IPF in particular is bleomycin-induced pulmonary fibrosis. Bleomycin is a cytotoxic glycopeptide antibiotic most commonly used as an antineoplastic agent in the treatment of lymphomas, squamous cell carcinomas, and testicular cancer. A serious side effect of bleomycin treatment in humans is induction of pulmonary fibrosis. Approximately 10% of treated patients develop side effects associated with pulmonary toxicity, with 1% progressing to full pulmonary fibrosis [113].

In the animal model of bleomycin-induced pulmonary fibrosis, mice or rats are anesthetized and intratracheally administered high doses of bleomycin. In the first several days, an acute pulmonary inflammatory reaction takes place, followed by activation of fibroblasts and subsequent fibrotic deposition. Frank pulmonary fibrosis usually develops within days of dosing. The key disadvantage of the bleomycin rodent model is that the chemically-induced pulmonary fibrosis in rodents does not completely mimic naturally-occurring human idiopathic fibrosis, especially in regards to the initial acute inflammatory phase. More significantly, bleomycin-induced fibrosis in rats and mice will often resolve over a period of time, while IPF in humans progressively worsens [110, 114, 115].

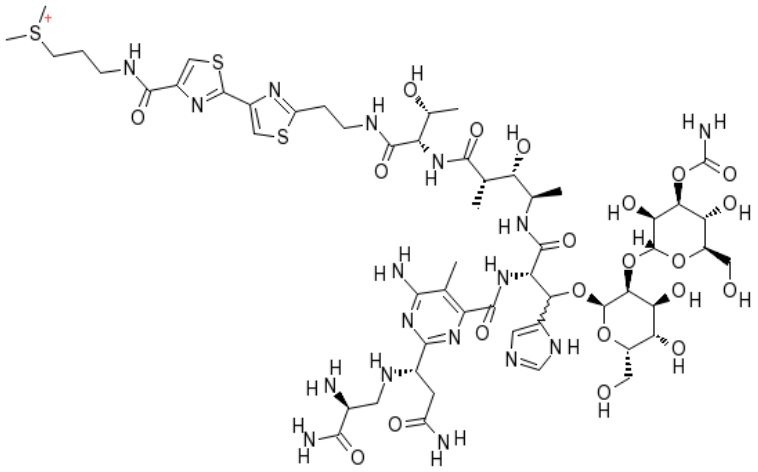


Figure 2-7. Molecular structure of bleomycin.

One method of assessing the severity of a fibrotic response *in vivo* is by measuring collagen content and the relative rates of collagen synthesis and degradation in the lungs. This is accomplished by measuring a major constituent of collagen, hydroxyproline. When collagen is synthesized, the amino acid proline is hydroxylated to form hydroxyproline. This hydroxylated amino acid allows the collagen helical structure to bend tightly, serving as a strong molecular "hinge" or "spring." The only other protein that has significant amounts of hydroxyproline is elastin, which is another highly elastic structural protein found in fibrotic deposits [116]. By administering radio-labeled proline, and then measuring the absolute and relative amounts of proline to hydroxyproline, one can calculate rates of collagen synthesis, degradation, and absolute quantities of the protein in tissue [117].

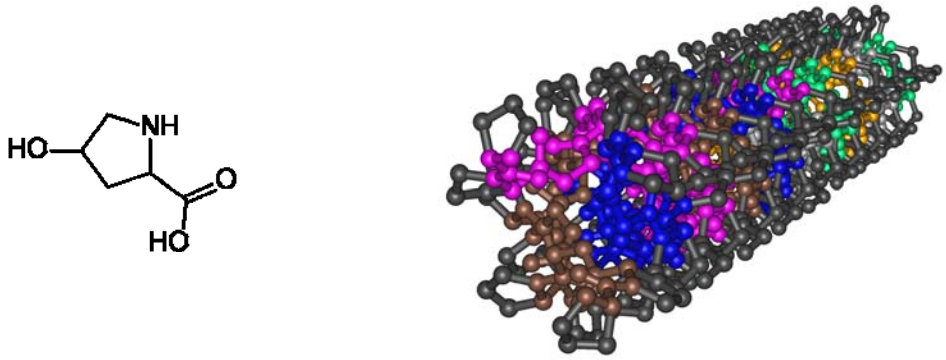


Figure 2-8. Hydroxyproline structure (left) and the collagen triple helix (right).

In humans, clinical studies on IPF include clinically-observable endpoints. Examples of these are imaging (X-rays, CT-scan), immunohistochemical and/or molecular characterization of biopsies, bronchoalevolar lavage [118, 119], and measurements of blood markers. Lung transplants and explants are also sources of biopsies that can be studied through molecular biology techniques or immunohistochemistry. Finally, post-mortem examinations of lung tissue,

especially if done quickly after death before autolysis has set in, provides insight into end-stage IPF disease [63].

3 TRANSFORMING GROWTH FACTOR β1 AND SMAD3

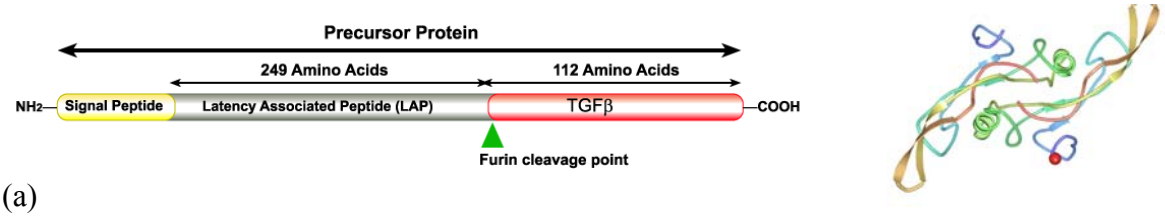
3.1 TRANSFORMING GROWTH FACTOR β1

Transforming Growth Factor β1 ($TGF\beta_1$) is a key anti-inflammatory cytokine involved in numerous cell signaling and cellular processes. These including cell proliferation, differentiation, cell adhesion and migration, extracellular matrix deposition, apoptosis, embryonic development, and immune response [17, 18, 106, 120-123]. Dysregulated or aberrant $TGF\beta_1$ signaling is implicated in numerous pathological conditions including cancer, pulmonary hypertension, and a wide variety of organ-specific fibrotic diseases including idiopathic pulmonary fibrosis (IPF). [123-125].

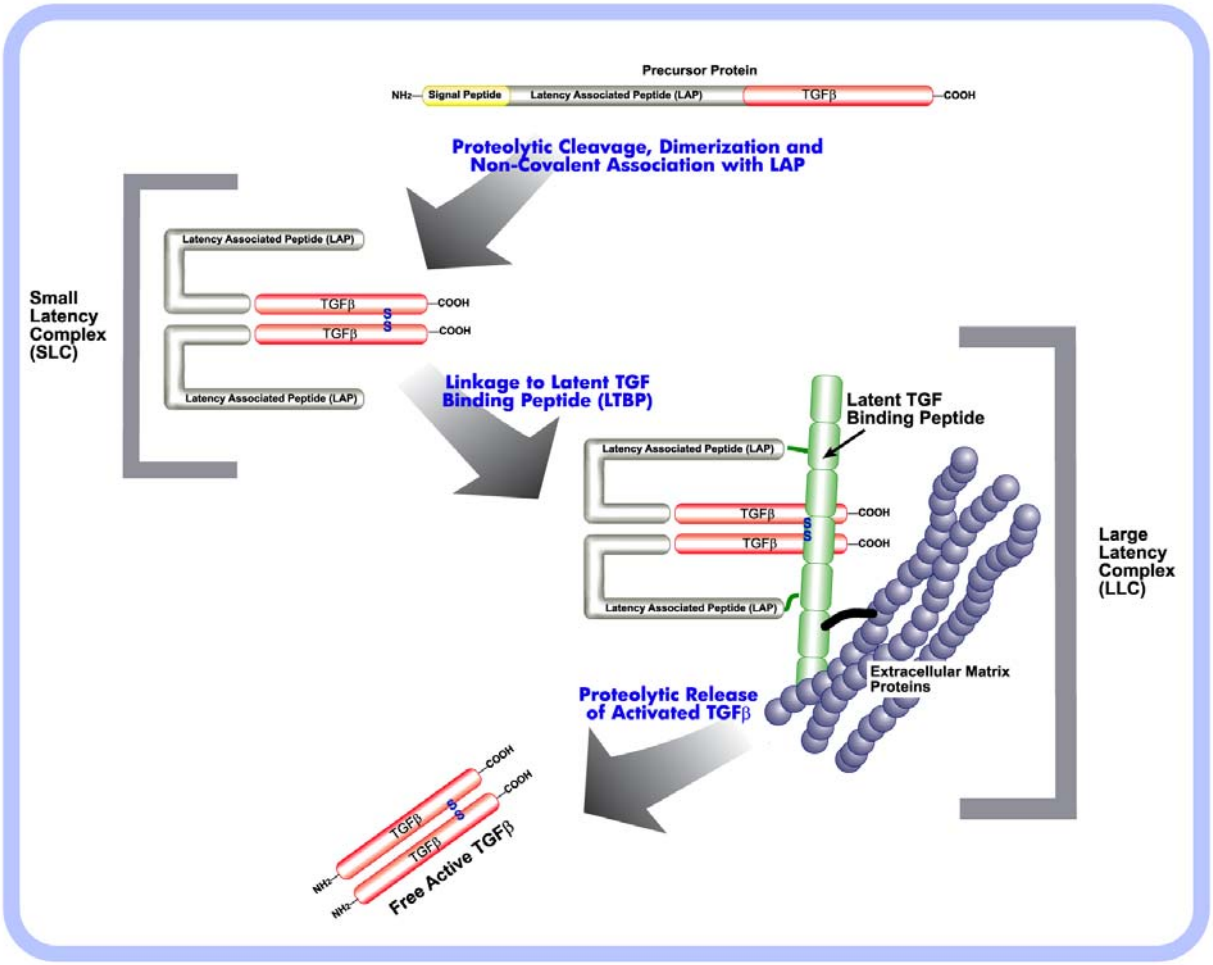
In humans, the TGFβ superfamily of homologous proteins consists of about 42 members that include three isoforms of TGFβ itself (TGFβ1, 2, and 3), nodals, activins, inhibitins, bone morphogenic proteins (BMPs), myostatin, and anti-Müllerian hormone (AMH) [19, 20, 22].The TGFβ family of proteins is also highly conserved across mammalian species [22, 121].

Nearly all cell types both produce and have receptors for TGFβ, although the effects on each type of cell are varied and specific to a particular cell type [17, 18, 20, 126, 127]. The isoform $TGF\beta_1$ is of particular interest in understanding IPF because it is found abundantly in platelets and plays the major role in wound healing as well as both normal and aberrant fibrotic

deposition [122, 128, 129]. More importantly, of the three isoforms, TGFβ$_1$ is found in the alveolar epithelial cells and macrophages of pulmonary fibrosis patients [130].



(a) (b)



(c)

Figure 3-1. TGFβ$_1$ Structure.
(a) Linear structure of TGFβ$_1$ precursor peptide; (b) three-dimensional structure of active TGFβ$_1$ homodimer; (c) Processing of TGFβ$_1$ peptide: TGFβ$_1$ is secreted as a pro-peptide precursor; proteolytically cleaved, dimerized, and non-covalently associated with latency associated peptide (LAP); bound to ECM proteins in inactive form with latent TGF binding peptide (LTBP); and released/activated through proteolytic cleavage [131-134].

$TGF\beta_1$ is transcribed in an inactive precursor polypeptide consisting of 390 amino acids. At the N-terminus lies a signal sequence consisting of 29 amino acids. The active $TGF\beta_1$ peptide lies at the C-terminus of the amino acid chain and is 112 amino acids long. The sequence upstream of the $TGF\beta_1$ sequence is known as latency-associated peptide (LAP). [133-136].

Upon cleavage through proteolytic hydrolysis, the LAP/$TGF\beta_1$ peptide is released and homodimerizes through disulfide linkages at the cysteine residues [17, 18, 133, 135]. The precursor molecule is also non-covalently linked to a latent TGF binding peptide (LTBP). Inactive $TGF\beta_1$-LAP precursor complexes are commonly secreted and stored in high concentrations in the ECM surrounding epithelial cells. Upon perturbation by mechanical injury or epithelial cell migration, proteases such as plasmin, metalloproteinase 9, elastase, cathepsins or the cell surface protein integrin $\alpha v\beta 6$, cleave the $TGF\beta_1$-LAP complex to release activated $TGF\beta_1$ homodimer [121, 133, 135, 137-139]. $TGF\beta_1$ is an immediate-early response autoendocrine cytokine, and as such is ubiquitously sequestered around cells in an inactive state, poised to be released and activated quickly upon demand.

3.2 SMADS

$TGF\beta_1$ exerts its principal effects on epithelial cells and fibroblasts through the $TGF\beta_1$/SMAD3 signal transduction pathway operating between cell surface receptors for $TGF\beta_1$ and the gene regulatory machinery in the nucleus [140, 141]. In humans, there are eight members

of the SMAD family of transcription factors.[1] Of these, five are receptor-regulated SMADs, or R-SMAS: SMAD1, SMAD2, SMAD3, SMAD5 and SMAD9. SMAD4 is referred to as a common-mediator SMAD, or co-SMAD. SMAD6 and SMAD7 are antagonistic or inhibitory SMADs and are therefore referred to as I-SMADs [20].

SMAD proteins have two MAD-homology domains, an N-terminal one called MH-1 and a C-terminal one called MH-2. The MH-1 domain has DNA-binding activity and is necessary for nuclear translocation. The MH-2 domain bind proteins and transactivates other proteins SMADs also have a C-terminal SXS motif containing serines which are targets of activation through phosphorylation [20].

Table 3-1. List of human SMAD proteins

Receptor SMADs	**R-SMADs**	**SMAD1, SMAD2, SMAD3, SMAD5, SMAD9**
Common-mediator SMADs	**co-SMAD**	**SMAD4**
Inhibitory SMADs	**I-SMADs**	**SMAD6, SMAD7**

[1] Human SMAD proteins are homologs of two proteins found in lower species. One is a *Drosophila* protein that represses an embryonic development gene called decapentaplegic. The discoverers called the "mothers against decapentaplegic" or "MAD". The other protein shares homology with the *C. elegans* protein SMA (contracted from "small/male/abnormal"). The name "SMAD" therefore refers to a contraction of the two protein names, "SMA" and "MAD."

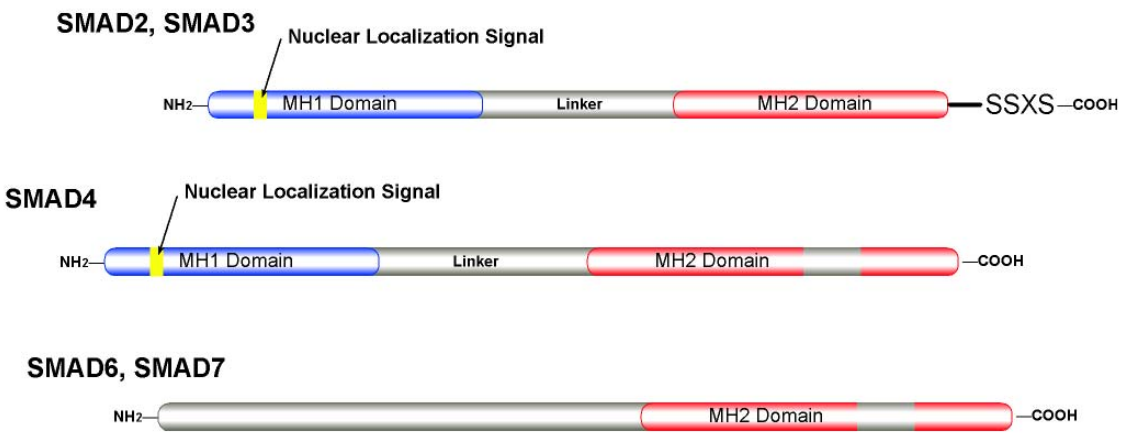


Figure 3-2. Linear structure of SMAD proteins.

Two kinds of TGFβ_1-responsive receptors are found on the cell surface: TGFβ receptor type I (TβRI) and TGFβ receptor type II (TβRII). Each is present as a homodimer on the cell surface. Both receptor types are transmembrane serine/threonine kinases [142]. Both receptor types are also very similar in sequence and structure, but the type I receptor has a conserved glycine/serine-rich sequence (GS-sequence) immediately upstream of its kinase domain. TβRI and TβRII receptors are general names for receptors activated by different members of the TGFβ family of ligands. Each pair of receptor types are also known more specifically as species of activin-receptor-like kinases (ALKs). Type I receptors may consist of one of seven separate ALK species. The type I receptor specific for TGFβ_1 in particular is known as TβRI/ALK5 [20, 143].

The signal transduction cascade begins when a ligand, in this case the active homodimer form of TGFβ_1, binds to and activates the TβRII receptor pair. The activated type II receptor pair recruits the type I receptor pair into the complex through an auxiliary protein known as Smad Anchor for Receptor Activation (SARA). This complex phosphorylates the GS-sequence in the type I receptor pair. The phosphorylated, activated type I receptor pair in turn phosphorylates the

cytoplasmic transcription factors SMAD2 and SMAD3. This heterodimer undergoes a conformational change that releases it from the receptor complex. The SMAD2/3 heterodimer subsequently complexes with a third co-transcription factor, SMAD4. The entire activated heterotrimer complex, consisting of SMAD2/3/4, then translocates into the nucleus to effect transcriptional regulation [17, 18, 20, 140, 144]. Prior to TGFβ_1 activation, the R-SMADs are primarily distributed in the cytoplasm, whereas SMAD4 is distributed in both the cytoplasm and nucleus. The activated SMAD2/3/4 complex translocates to the nucleus through a nuclear localization sequence (NLS) in the MH1 domain of the R-SMADs [143].

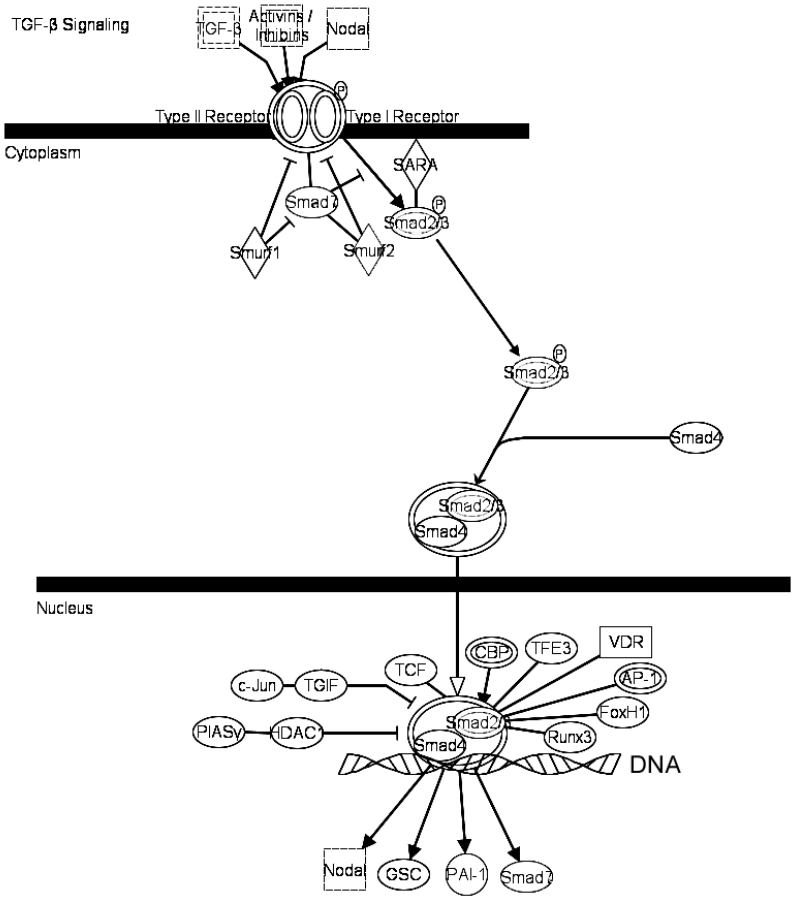


Figure 3-3. Schematic diagram of the TGFβ_1/SMAD3 signaling pathway [32].

The SMAD2/3/4 complex has relatively weak DNA binding affinity as well as less stringent DNA sequence specificity. Greater binding affinity and specificity is afforded by complexing with co-transcriptional activators such as CBP (cAMP-responsive element-binding protein binding protein) and p300. CBP/p300 also has chromatin structure-modifying histone acetyltransferase (HAT) activity [20].

Of the R-SMADs activated through the TGFβ_1 pathway, only SMAD3 and SMAD 4 have significant ability to bind DNA directly. In terms of optimal nucleic acid sequence, SMAD3 and SMAD4 can bind to the sequence 5' – CTCTAGAC – 3'. SMAD3 binds specifically through its MH1 domain to the sequence 5' – GTCT – 3' and its reverse complement 5' – AGAC – 3', the later of which is commonly known as a SMAD binding element (SBE). SMAD3 has also been shown to bind to the sequence 5' – GGCGGG – 3' in the c-Myc promoter [20].

The TGFβ_1/SMAD3 signal transduction/gene regulation pathway also includes a negative feedback loop that quickly terminates pathway activation in the absence of TGFβ_1 stimulation. The two inhibitory SMADs, SMAD6 and SMAD7, are distributed in the cell cytoplasm and are upregulated by the TGFβ_1/SMAD3 signal transduction pathway. These I-SMADs terminate SMAD3 signal transduction by competing with SMAD2/3 for binding to the TβRI receptor. These SMADs interact with the TβRI receptor directly through their MH2 domains [20, 140, 143, 145-147].

Additionally, SMAD6 and SMAD7 interact directly with proteins called SMAD ubiquitination regulatory factor (SMURFs), of which there are two varieties, SMURF1 and SMURF2. These proteins in turn interact with phosphorylated SMAD2 and SMAD3 through E3

ubiquitin ligases to target SMAD2 and SMAD3 for degradation in the proteosome, further suppressing TGFβ_1/SMAD3 pathway activity [20, 140, 145, 148-150].

Finally, in the nucleus, the R-SMADs are continually dephosphorylated and thus, dissociate from their complexes. They subsequently expose a nuclear export sequence (NES) and are shuttled back into the cytoplasm [143, 151].

Aside from negative regulation of the TGFβ_1/SMAD3 pathway by SMAD6, SMAD7, and SMURFs, the pathway has numerous co-repressors of transcriptional activity. Among the most notable of these are the c-SKI and SNO-N proteins. They are proto-oncogene protein products that are found in high levels in many types of human cancer cells. They bind to the MH2 domains of SMAD2 and SMAD3, thereby interfering with binding with transcriptional co-factors and with DNA [152]. They can additionally block the TGFβ_1/SMAD3 pathway by binding to SMAD3 and sequestering it in the cytoplasm [153]. Similarly, c-MYC can inhibit SMAD2 and SMAD3 activity, but only from among a small subset of SMAD3 binding targets such as $p15^{Ink4B}$ and $p21^{Cip1}$ [153].

3.3 INTERACTION OF TGFβ_1/SMAD3 WITH OTHER GENE REGULATORY/ SIGNALING PATHWAYS

Like other cellular and transcriptional factor pathways, the TGFβ_1/SMAD3 pathway does not work in isolation. It is known to interact with numerous other signaling and transcription factor pathways through various mechanisms. One such interaction is through the inhibitory SMAD7. The JAK/STAT signal transduction pathway, which is activated by Interferon-γ,

induces SMAD7. Likewise, TNFα, IL-1, and lipopolysaccharides are well-known activators of NFκB, which in turn also induces SMAD7. In both these cases, we see one mechanism by which pro-inflammatory cytokines or stimulus can suppress activity of the $TGF\beta_1$ pathway [20].

Another interaction between $TGF\beta_1$ and a major signaling pathway involved in cell cycle progression is with mitogen-associated kinase/ MAPK/Erk [144]. MAPK and cyclin-dependent kinases (CDKs) can phosphorylate R-SMADs. Similarly, MAPK kinase 1 (MEKK-1) activates JNK and also phosphorylates SMAD2. JNK in turn phosphorylates SMAD3. Conversely, TGFβ is known to activate MAPK as well [20].

R-SMADs also interact directly with DNA-binding proteins. The phosphorylated R-SMAD complex interacts with the forkhead box transcription factors, FOXO and FOXH1. R-SMADs also interact with zinc-finger DNA-binding proteins such as GATA3,4,5,6 and ZNF198. Finally, the TGFβ/SMAD3 pathway interacts with other signaling and transcriptional activation pathways such as β-catenin and HIF-1A.

3.4 EPITHELIAL-MESENCHYMAL TRANSITION

In most cell types, $TGF\beta_1$ inhibits cell proliferation [154]. $TGF\beta_1$ stimulation of epithelial cells, however, either (a) inhibits cell proliferation, (b) causes cells to undergo apoptosis, or (c) causes them to dissociate from their basement membrane, proliferate, and migrate, taking on many characteristics of certain types of mesenchymal cells [16]. The latter (c) process is called *epithelial-mesenchymal transition* or EMT [12-14]. Mesenchymal cells include progenitor cells of muscle cells, adipocytes, osteoblasts, chondrocytes, and fibroblasts. The

mesenchymal cells that result from EMT closely resemble fibroblasts in morphology and behavior, sometimes with additional motile and contractile abilities characteristic of muscle cells. These resultant fibroblasts with motile/contractile characteristics are known as *myo*fibroblasts. [98, 99, 154].

Epithelial cells normally lie in a polarized, sheet-like structure attached to a basement membrane. $TGF\beta_1$ stimulation of epithelial cells downregulates E-cadherin, causing the cells to lose tight attachment to each other and their basement membrane. The loss of E-cadherin also results in upregulation of c-MYC, cyclin D1, and MMP7, inducing cell proliferation and enhancing cell motility [125]. As the EMT process progresses, cytoskeletal rearrangement takes place, causing the cells to adopt a spindle-like structure characteristic of fibroblasts. The cells also begin expressing mesenchymal markers such as vimentin, N-cadherin, and fibronectin [125, 155].

Prolonged $TGF\beta_1$ stimulation induces the mesenchymal cells to secrete collagens such as COL7A1, decrease protease production, and increase the secretion of protease inhibitors such as TIMPs and plasminogen activator inhibitor 1 (PAI-1) [19, 107, 121, 122]. Eventually the cells may begin expressing alpha smooth muscle actin (αSMA) and transition into motile myofibroblasts that aggressively infiltrate and deposit ECM proteins, particularly collagens [12-14, 97-99, 105-107, 125]. What started as an orderly membrane-bound sheet of polarized epithelial cells transforms into a disorganized collection of motile mesenchymal cells that aggressively infiltrate other tissues, secreting fibrous proteins. While EMT is expected to occur during certain phases of normal embryonic development, in adults it is characteristic of fibrotic diseases as well as neoplastic invasion and metastasis [98, 99, 125].

The TGFβ_1/SMAD3 pathway in particular is implicated as being necessary, although not sufficient itself, in inducing EMT in epithelial cells. Cultured epithelial cells from SMAD3-deficient mice could not be induced to undergo EMT by TGFβ_1 treatment, whereas the same cell type cultured from wild-type mice exhibited EMT [15, 156].

3.5 THE TGFβ_1/SMAD3 PATHWAY AND ITS IMPLICATION IN IPF

Numerous human and animal studies implicate TGFβ_1 in the pathogenesis of IPF. First, strongly increased TGFβ_1 mRNA expression and protein secretion have been found in areas of active extracellular matrix deposition in fibrotic foci of IPF patient lung biopsies [130, 157-161]. Alveolar macrophages secrete TGFβ_1 in both humans and bleomycin-induced mouse model of fibrosis [130, 162].

Further, adenoviral-vector mediated gene transfer of TGFβ_1 into rats induced distorted alveolar and parenchymal structure, increased extracellular matrix deposition, and induced a characteristic “honeycombing” feature in lung architecture that are all hallmarks of pulmonary fibrosis [112, 139]. SMAD3 knockout mice (but not SMAD2) were protected against bleomycin-induced pulmonary fibrosis [141, 163, 164]. Similarly, administration of the TGFβ_1 inhibitor halofuginone in mice markedly reduced ionizing radiation-induced fibrosis [165]. When exogenous SMAD7, which is antagonistic against the TGFβ_1/SMAD3 pathway, was introduced into mice through an adenoviral vector, it protected them against bleomycin-induced pulmonary fibrosis [166]. Also, administration of exogenous TGFβ_1 induced EMT in rat alveolar Type II cells, and this effect was significantly blocked by adenoviral-induced SMAD7 [155]. Finally, in

bleomycin-induced pulmonary fibrosis in rats, SMAD3 clearly decreased in the cytoplasm and increased in the nucleus of lung tissue cells, indicating activation of the TGFβ_1/SMAD3 pathway and nuclear translocation of SMAD3 [167]. Thus, there is ample experimental evidence in animals and humans, both directly and indirectly, that the TGFβ_1/SMAD3 pathway plays a significant role in the pathogenesis of pulmonary fibrosis [122].

4 METHODOLOGICAL ISSUES

4.1 WHAT IS SYSTEMS BIOLOGY?

4.1.1 REDUCTIONISM VERSUS SYSTEMS THINKING

There are two ways to approach investigating any complex system such as a cell or organism. One is the reductionist approach, where the complex system is divided into component parts and those individual parts are studied in isolation. The other approach is holistic, where a global view of the interactions between the parts, as well as the characteristics of the system on the whole, are the objects of study.

The idea of breaking complex systems down into their component parts as a way of understanding the system is both intuitive and reasonable. This is especially true during most of the past two and half millennia, during which natural philosophers were struggling to understand the mysterious workings of nature with very few tools or starting background knowledge. When first approaching a poorly understood system or phenomenon, it is only natural to want to open it up, dissect it, examine its parts, and then try to discern "what makes it tick." As such, the reductionist approach has been enormously successful throughout the history of science, especially in such disciplines as physics and human anatomy.

The systems-level approach has enjoyed far less visibility until more recently. Being inherently complicated by definition, complex systems have generally been difficult to study until the appropriate mathematical, and until recently, computational tools have become available. Some familiar more early disciplines that have evolved a systems-level approach prior to the introduction of computational tools are ecology and physiology.

One of the main questions concerning the relationship between reductionism and holism is whether the properties of one logically entail the other. More specifically, the question is whether a system on the whole has properties that could not have been predicted from a sole understanding of the individual parts. We call this *emergent* properties of the system. A simple example of emergence is that a detailed characterization of an individual bird's flying behavior would not necessarily entail the phenomenon of a group of birds flying coherently in a flock. Therefore, one of the key rationales for studying organisms or populations at a systems level is to be able to observe and study emergent behaviors that would not lend themselves to study through reductionist means.

As a more specific instance of emergent properties in biology, biological systems often include feeback loops that are difficult to identify or study solely using reductionist techniques. The very act of isolating components of a feedback loop destroys the very functional relationships responsible for its behavior. Negative feedback loops in particular give biological systems a degree of resistance to change from external pertubations, or a certain stability or *robustness* [109]. Conversely, positive feedback loops give a system unstability, or a rapid response to external pertubations. In some instances a positive feedback loop makes a subsystem's behavior so rapidly acting toward only one of two extreme states that it can be considered a *bistable* subsystem, acting like a switch with only two states, "on" or "off."

Therefore, a systems-level understanding of a biological system can provide an insight into its behavior not readily discerned using reductionist techniques.

4.1.2 DIFFERENT MEANINGS OF "SYSTEMS BIOLOGY"

Unfortunately there is no one definition of systems biology. Systems biology as a formal discipline is the result of a convergence of several different scientific and engineering traditions and methodological philosophies [109, 168-170].

One interpretation of systems biology is as a comprehensive explanation, description, or model of a complex system. In other words, we can use a systems approach to describe in detail a system's properties and function at the global level. One such description is of a system's structure, where the positions and inter-relationships between the component parts are identified and mapped. Another approach is functional, where the interactions between the component parts, as well as inputs and outputs, are described. A more thorough functional approach quantifies the dynamics of the system, particularly using series of differential equations to describe the system's component's behavior with respect to time.

A crucial element for systems biology is the choice of representation of the system. Just as in mathematics, rigorous rules and definitions of terms, notations, and syntax allow for clearer thinking about the system. One common approach to systems biology is to use tools borrowed from electrical engineering. Here, a biological system is represented by notation inspired by electrical circuit schematics and described and analyzed using similar mathematical tools [168, 171, 172].

Another interpretation of systems biology is as a methodology. The aim of systems biology as a methodological approach is to experimentally characterize the system by combining

and interpreting many individual measurements from throughout the system. In this regard, one can take a top-down approach, where numerous simultaneous measurements are taken throughout the system to produce a "snapshot" of the system's state at one point in time. Dynamic behavior of the system can be captured by taking a series of such snapshots over a time course. Conversely, a bottom-up approach would be to combine accumulated knowledge of individual parts to build up towards a comprehensive model of the entire system.

Similarly, in the systems methodological approach, a suitable representation of measurements and their interpretation is crucial. Systems biology information typically consists of large amounts of data, and often of various types. It must therefore be managed, analyzed, and displayed using computational tools. Most notably, systems biologists require ways to represent and summarize large numbers of variable magnitudes as well as denote the inter-relationships between variables. Some examples of common representations in systems biology are networks [173, 174], heat maps[34], clusters [34], and functional modules (see Figure 4-1) [175-179]. Additionally, to help visually manage large amounts of data, systems biology is often aided by interactive graphs, charts, tables that are linked to databases whose properties and functions can be automated [32, 33].

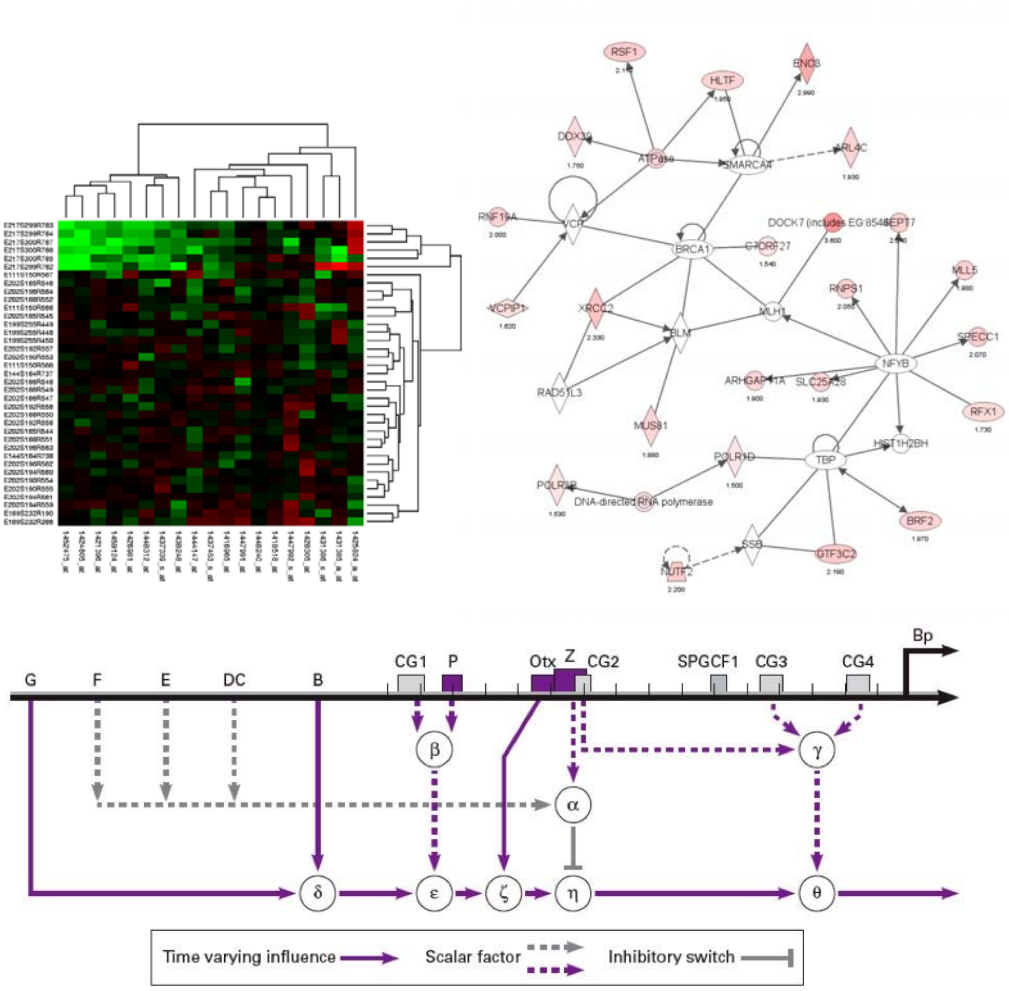


Figure 4-1. Visual representations commonly used in systems biology approaches [180, 181]. "Heat map" visually representing microarray expression intensity levels for individual genes (above left). Interconnected network of genes and/or gene products (above right) [32]. Sea urchin Endo16 transcriptional regulatory network represented as a pseudo-electrical diagram (from Yuh, *et. al*.) (below) [181].

The descriptive and methodological interpretations of systems biology are not mutually exclusive. Ideally, each informs the other to iteratively increase the overall comprehensive understanding of biological entities. Therefore, systems biology can be thought of as an iterative process, in which experimental information informs consequent biological models, and those biological models guide the design of experiments. This approach to biology research is

sometimes referred to as *integrative* biology, and is held by some as synonymous with systems biology [174, 182].

4.1.3 WHY SYSTEMS BIOLOGY NOW?

In the last few years, systems biology has become a "buzz word" of sorts even though the systems-level approach has been around for millennia. What accounts for this robust resurgence of interest in a systems approach to biology? It is the result of the confluence of a number of technological and organizational advancements.

First, it cannot be discounted that the systems approach is much more valuable as a tool when trying to solve difficult and complex problems. In other words, one reason systems biology has become popular recently is that most of the "easy" problems in biology and medicine have already been solved. Single-cause diseases and simpler biological processes were the first to be satisfactorily investigated, analyzed, and understood using reductionist approaches during the last centuries. Many of the remaining problems, such as understanding the etiology and progression degenerative diseases at the molecular and genetic level, for instance, seem to be the result of complex interactions between numerous genes and/or environmental factors. These problems have been frustratingly recalcitrant when addressed through reductionist means.

Consequently, problems involving numerous variables and complex inter-relations are just too difficult to solve by "traditional" scientific approaches of the last century. Until as recently as the 1970s and 80s, scientists commonly logged their measurements by pen and paper and performed their own data analyses by hand. Today, using computational tools for data acquisition, management, analysis, and display is not only more convenient, it is increasingly becoming indispensable simply due to the massive amounts of data involved.

Great advancements in systems biology have not just come from increases in sheer computing power, but also in sophistication of software engineering and the impact of computer technology on the culture of science. The relatively low cost of computing and subsequent introduction of personal computers have fostered an enormous amount of communication and collaboration between biologists, both informally and formally. Large interactive databases are accessible to everyone in the scientific community. This has produced informal but enormous world-wide scientific collaboration. Not only has computer technology promoted collaboration among biologists, but amongst and between numerous other disciplines as well, including computer scientists, engineers, physicists, mathematicians, statisticians, and clinicians, to name but a few [109, 182-184].

In addition to the enormous contribution of computer technology to biology, rapid advancements in high-throughput measurements have also been important to the growth of systems biology. Two notable technologies, automated sequencing and microarray technology in particular have revolutionized the biological sciences [27, 185]. Increasingly, ever more sophisticated imaging sciences, robotics, and nanotechnology are also increasing the efficiency, reliability, accuracy, and through-put capacity of biological measurements [186-192].

Finally, it should be recognized that the formal policies and initiatives of educational institutions, corporations, and governments have been enormously important in promoting the growth of systems biology. Enormous research initiatives such as the Human Genome Project have not only produced massive amounts of biological information in and of itself, but has greatly advanced the development of research technologies and computational tools [193-197]. This could not have been possible without policies that have directed the massive resources of large institutions and governments towards advancing systems biology.

4.2 SYSTEMS BIOLOGY IN THE PRESENT STUDY

As mentioned previously, the etiology of idiopathic pulmonary fibrosis (IPF) is poorly understood. It appears to be an extremely complex, multifactorial disease involving multiple molecular pathways, genetic factors, and most likely environmental and/or pathogenic exposures. As such, it has not lent itself to being studied purely by reductionist approaches. The present study is an application of systems-level techniques to try to elucidate information not easily discerned solely through reductionist techniques.

Two high-throughput methods were used in the study described here: ChIP-on-chip and whole genome gene expression microarrays. The purpose of using these methods was to provide a "snapshot" of the transcriptional state of human alveolar cells during specific activation of the $TGF\beta_1$/SMAD3 pathway in hopes of providing insights not available in a gene-by-gene interrogation.

Like most systems-level experiments, this study uses a combination of reductionist and systems approaches. For one thing, a cell culture model was used, which has a number of drawbacks. Being a single cell type in isolation in a culture dish, it should not be expected to have the same characteristics as that of an alveolar epithelial cell within its natural cellular ecology inside the lung of a human subject. Further, the cell line used was derived from an adenocarcinoma, and is therefore not completely representative of a normal, healthy alveolar epithelial cell.[2] Second, in the series of experiments presented here, the cell systems were stimulated with a single substance, exogenous $TGF\beta_1$. Again, this is a reductionist approach because it aims to measure the effects of a single cytokine in isolation.

[2] Due to practical considerations: human alveolar epithelial cells are nearly impossible to culture to produce the quantities necessary for the types of experiments used here, which is why A549 adenocarcinoma cells were used.

Another reductionist approach in the present study was the confirming of results of the systems level experiments on a gene-by-gene basis. Additionally, I followed up on particular findings with reductionist techniques, isolating single transcription factors and genes for further study (e.g., FOXA2, PINX1).

Despite the reductionist elements of this study, the key systems-level methods used in the present study were:

- Genome-wide transcription factor location analysis (ChIP-on-chip)
- High throughput gene expression analysis (Gene expression microarrays)
- Application of computational tools and databases to analyze data and interpret and display results (Ingenuity Pathways Analysis, Metacore GeneGo) [32, 33]
- An integrative approach to using genome-wide data: comparing, combining, and interpreting a combination of ChIP-on-chip and gene expression data.

5 CHIP-ON-CHIP

5.1 HISTORY, RATIONALE, AND BACKGROUND

The completion of mapping of the human genome through the Human Genome Project (HGP) was a monumentally important scientific achievement, but it resulted in a static map of nucleotide sequences that said nothing about the function of, and interaction between, genes and gene products [193, 194, 197, 198]. An analogy would be that the HGP gave us a phone book with addresses of individual genes, but gave us no information about how the each individual in the society of genes behaves and interacts with others. To understand truly how cells function in both normal and disease states, it is important to understand how these genes interact, particularly how gene transcription is controlled. This field of study in the "post-genomics" era is known as functional genomics.

The human genome consists of an estimated 25,000 genes [199-201]. In any particular cell type, however, only a subset of these genes are ever expressed at any one time [199, 202, 203]. How it is determined that certain genes will be expressed in a cell, or in other words the logic of transcriptional regulation, is one the major frontiers left to be explored in genomics.

The canonical model of eukaryotic gene transcription *cis*-regulation states that particular transcription factor proteins bind to specific short nucleotide sequences that reside inside or upstream of a gene coding region. These short nucleotide sequences are known as regulatory

elements. Most commonly, these regulatory elements, or motifs, are found within a couple thousand base pairs of the transcriptional start site of the gene sequence. This area upstream of the transcriptional start site of a gene is known as its promoter region [203-205].

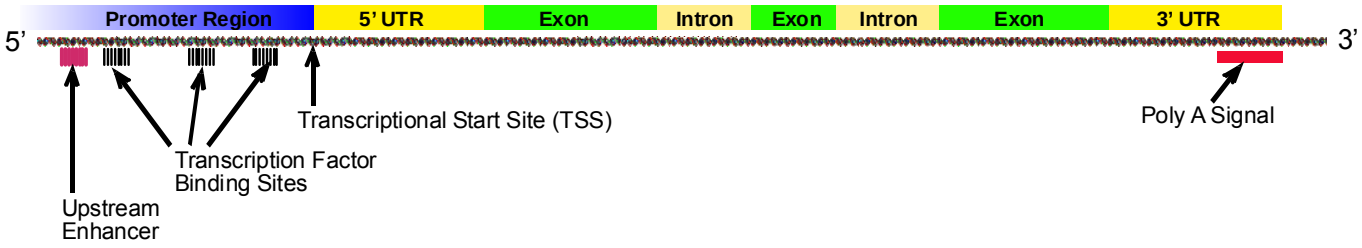


Figure 5-1. Basic anatomy of a gene promoter and coding sequence [204].

When a gene sequence is to be transcribed to mRNA (and ultimately, translated to an amino acid sequence and folded into a protein), a series of events occurs involving both specific and general transcription factor proteins. Certain transcription factors recognize specific DNA sequences and attach. to them. These sequence-specific transcription factors recognize these nucleotide sequences because they have molecular structures that allow them to recognize and fit snuggly within the three-dimensional structure of a particular nucleotide sequence of a DNA double-strand helix [204-206]. Examples of transcription factor motifs that recognize specific DNA sequences are described as helix-turn-helix (HTH), helix-loop-helix (HLX), zinc fingers, and leucine zippers [205]. These motifs recognize and fit snuggly within the major groove of the helical DNA structure (Figure 5-2, Figure 5-3).

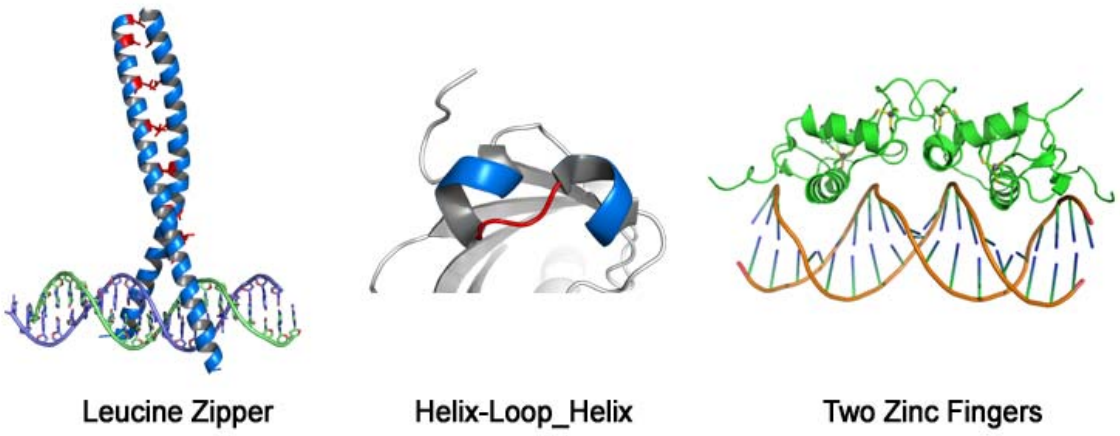


Figure 5-2. Examples of specific transcription factor DNA-binding motifs (www.lbl.gov).

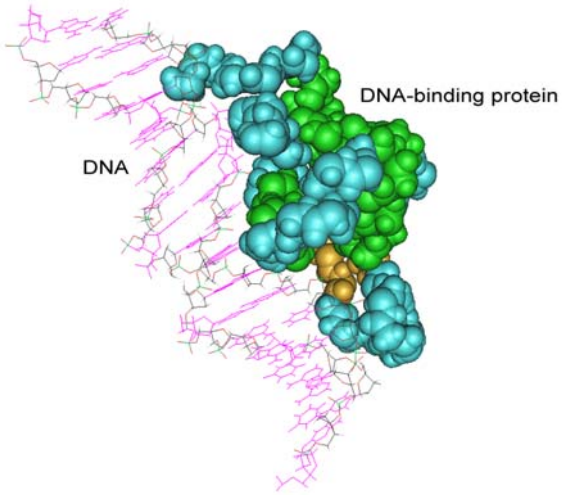


Figure 5-3. Illustration of a specific DNA-binding protein molecular structure binding within the major groove of the helical DNA structure (www.lbl.gov).

Upon attachment of one or more specific transcription factors, other transcriptional machinery elements are likewise recruited to attach to the protein/DNA complex. These are known as general transcription factors and co-factors. They contribute to the formation of an aggregation of functional proteins known as the preinitiation complex (PIC). The preinitiation

complex contains six general transcription factor proteins, TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Once the requisite members of the pre-initiation complex are assembled, they recruit RNA polymerase II (Pol II) which begins the process of transcription. The preinitiation complex and PolII holoenzyme begins sliding down the DNA strand, transcribing the nucleotide sequence of DNA to a single strand of messenger RNA (mRNA). It is this mRNA that will be further processed and serve as a template for translation into an amino acid sequence that will be folded into a protein [204, 206].

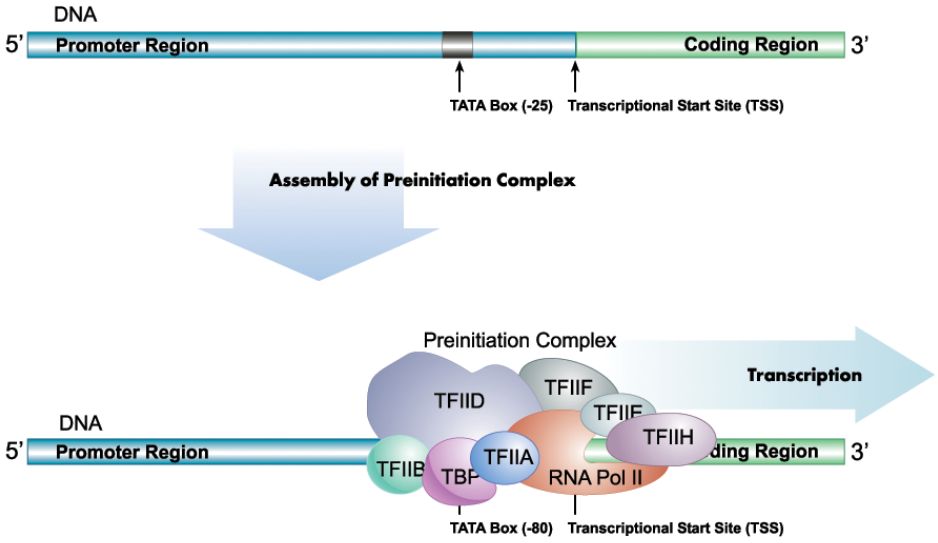


Figure 5-4. Basic anatomy of assembly of a preinitiation complex of specific and general transcription factors.

A great amount of knowledge has been gained about transcriptional machinery in general by studying the process in viral replication, *E. coli*, yeast, and metazoans [205]. However, a major step in understanding the functions of, and functional relationships *between* different genes in an organism, is enumerating lists of target genes of specific transcription factors. We can attempt to predict where specific transcription factors should bind theoretically to specific

nucleotide sequences, or transcription factor binding elements, within the promoter elements of gene coding sequences using statistical and computational models [203, 207-209]. However, computational models of transcription factor binding do not agree completely with empirical data [210]. Thus, computational models may provide a heuristic guide for generating hypotheses as well as an important tool for understanding phylogenetic relationships between species, but they are not a substitute for true empirical information on transcription factor target genes. Gathering actual experimental data on which genes' promoter regions are targets for specific transcription factors is a crucial step in learning about the functional and regulatory relationships between genes. As an example of a practical application of understanding transcriptional regulatory targets, such information may lead to identification of specific gene products as targets for rational drug design to block gene products that are involved in disease processes.

5.2 CHROMATIN IMMUNOPRECIPITATION

One standard experimental method for exploring transcription factor/DNA binding *in vivo* is chromatin immunoprecipitation (ChIP). The concept behind this method is that transcription factor/bound DNA complexes formed *in vivo* during a particular transcriptional state can be preserved by lightly fixing cells in that state with a cross-linking agent. Then the transcription factor/bound DNA complexes can be isolated using an antibody specific to the transcription factor. After transcription factor/bound DNA complexes are isolated and purified, the DNA sequences can be released from the bound protein complex and then identified. Thus, once identified, the DNA sequences can be matched to known gene promoter elements, in turn identifying the genes that had been a target for the specific transcription factor protein when the

original cells were in a particular transcriptional state. ChIP experiments can be performed to provide a snapshot of transcriptional state of cells in particular metabolic, differentiation, pathological, or signaling molecule-induced states. ChIP experiments are not confined to specific transcription factors, but can be performed using antibodies against general transcription machinery such as Pol II or against acetylated or deacetylated histones [211].

In a typical ChIP assay, cultured cells are treated with a low concentration of formaldehyde which lightly fixes them and consequently binds the transcriptional regulatory proteins to the promoter regions of their respective target genes. Cells are collected, lysed, and sonicated to shear chromatin into a distribution of fragment sizes ranging from around ~200bp to ~2000 bp. An antibody against the specific transcription factor protein of interest is then used to isolate transcription factor/promoter fragment complexes. Usually a control pulldown is also performed in parallel with the antibody pulldown that serves as a background reference—this can be whole genome chromatin fragments, a "no-antibody" control, or a "mock" IP using an antibody not expected to have any affinity for the transcription factor of interest. After antibody capture (along with its parallel reference control), the "precipitation" step consists of antibody/transcription factor complexes being immobilized to Protein A or Protein G conjugated to agarose or magnetic beads. Once immobilized, any non-antibody-bound chromatin is washed away in a series of stringency wash steps. The transcription factor/promoter fragment cross-links are subsequently reversed, and the released promoter fragments are then analyzed.

The DNA sequences isolated from ChIP experiments, or ChIP pulldown sequences, can be analyzed and identified in a number of ways. The two main categories of DNA fragment analysis and identification are sequence-based methods and hybridization-based methods [212-217].

The most common sequence-based analysis method is ChIP-PET (Paired End diTags) [216]. PET is a cloning, ditag (see below) concatenation, and sequencing-based method loosely based on the methods of Serial Analysis of Gene Expression (SAGE) [218]. In ChIP-PET, DNA fragments from ChIP pulldown are cloned into a plasmid and transfected into E. coli, producing a plasmid library. Plasmid DNA is extracted, purified, and digested with an MmeI restriction endonuclease, producing 24-mer tags from each the 5' and 3' ends of the sequence. The tags are self-ligated into ditags, concatenated, and amplified again through cloning. The amplified concatemers are extracted, purified, and sequenced. The individual ditag sequences are aligned to known sequences in the human genome. The redundancy of using both the 5' and 3' ends of the sequence provides added confidence against false positives due to experimental or sequencing errors [212, 216]. The main drawbacks of sequencing methods such as ChIP-PET are that they are extremely labor-intensive, expensive, and require specialized laboratory facilities and techniques.

Figure 5-5. Mme1 digestion product showing length of individual sequence incorporated into PET ditags [219].

The hybridization-based method for analyzing ChIP pulldown sequences uses microarrays specifically designed to contain sequences for the promoter regions of genes, i.e., promoter microarrays. ChIP combined with promoter microarrays is known as ChIP-chip or ChIP-on-chip. In ChIP-chip, the promoter DNA fragments resulting from the ChIP antibody pulldown are amplified and labeled with a fluorophore such as Cy-3. A reference control consisting of a no-antibody or mock antibody pulldown is also labeled with a different fluorophore, such as Cy-5. The two labeled materials are then mixed together and hybridized to a

promoter microarray. After hybridization, the arrays are washed in stringency buffers and image analyzed on a laser scanner. The spot intensity ratios between pulldown (IP) and reference (mock) are indicative of the relative amount of promoter fragment sequences originating from the promoter sequences of the transcription factor target genes [211].

The first published report of ChIP-on-chip for identifying target genes of transcription factors was in 2002 by Lee and Young [220]. Lee and Young tagged regulatory proteins in *Saccharomyces cerevisiae* with a c-myc epitope and used an antibody against c-myc to isolate those regulators and identify the DNA sequences bound to them. Subsequently, researchers began using ChIP on mammalian cells using antibodies against native transcription factor proteins [221, 222]. ChIP-on-chip is now a well-established method for globally identifying gene promoters that are bound in a cell's nucleus in a particular transcriptional state [211, 220, 221, 223]. My study used the ChIP-on-chip approach, as opposed to a sequence-based method. Therefore, the remainder of this discussion will focus specifically on methodological issues surrounding the ChIP-on-chip technique.

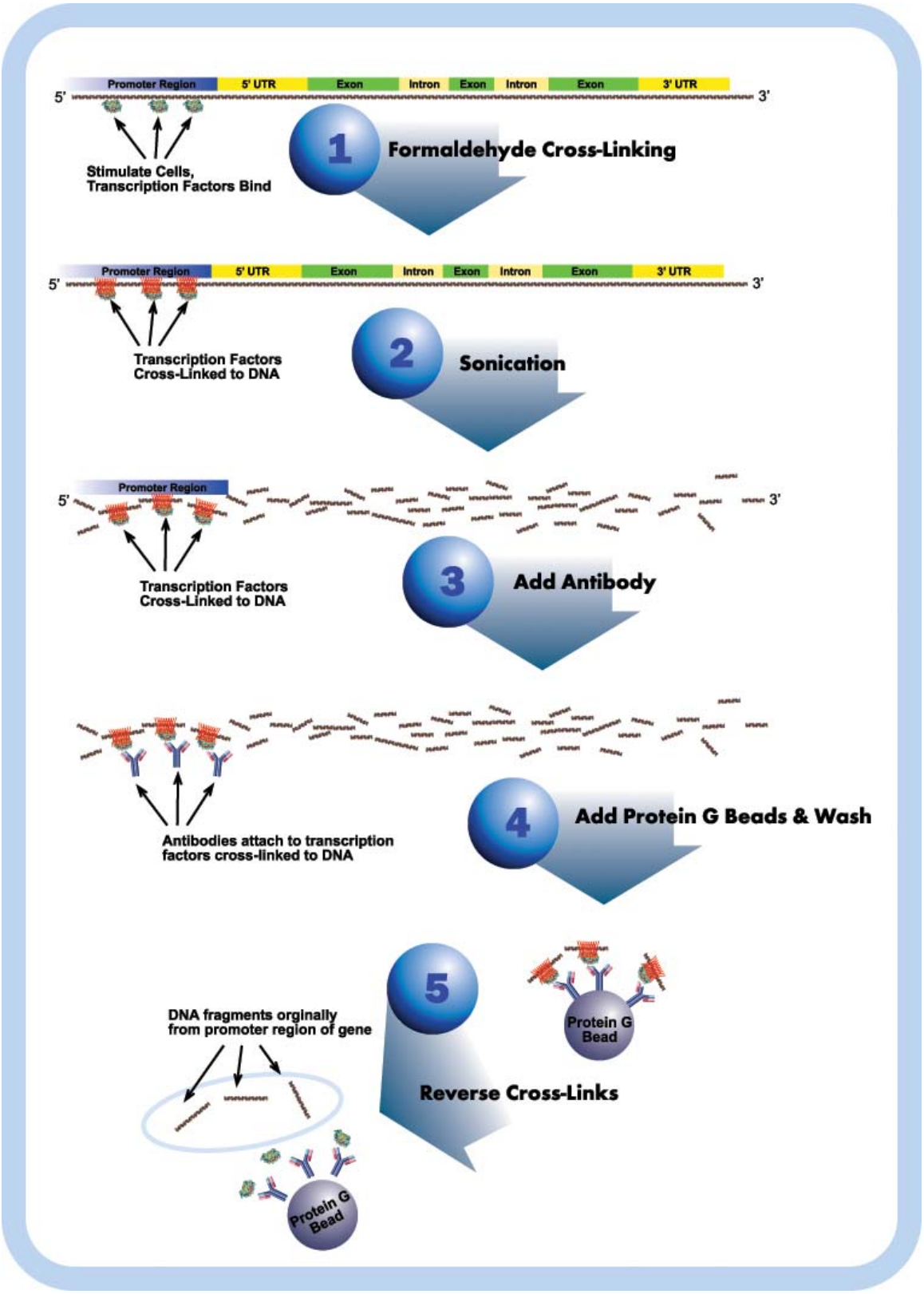


Figure 5-6. Schematic overview of the ChIP procedure.

5.3 CHIP-ON-CHIP METHODOLOGICAL DETAILS AND LIMITATIONS

5.3.1 FIXATION PARAMETERS

Cross-linking agents such as formaldehyde and glutaraldehyde are commonly used in histology and histochemistry to preserve cellular architecture during staining and imaging. Cross-linking agents stabilize the positions of macromolecules by forming bridges between them, giving the cell "stiffness" that resists mechanical deformation and chemical degradation—important features when histology sections are subjected to numerous chemical and mechanical steps. Formaldehyde in particular forms cross-links in cells by reacting with the amino group of the N-terminal amino acid residue and the side-chains of arginine, cysteine, histidine, and lysine residues [224-226]. As such, cross-linking agents such as formaldehyde alter protein conformation. This many not pose as much of a problem for histology studies since cells are imaged at the cellular structure level. However, slight alterations in protein conformation can significantly alter a protein's epitope geometry and biochemical properties [225, 226].

In a ChIP assay, the purpose of cross-linking is to stabilize the structural relation between the pre-initiation complex (containing a transcription factor of interest) and its cognate nucleic acid sequence in chromatin. An insufficient degree of cross-linking does not stabilize the structure enough, while excessive cross-linking causes three problems. First, excessive cross-linking risks changing the transcription factor protein epitope geometry, making its antibody recognition difficult in the subsequent pulldown step. Antibodies are typically made against the native protein species, not formaldehyde cross-linked proteins. Second, excessive cross-linking "overstabilizes" cellular cytoskeletal structure which makes it difficult to separate cellular material from chromatin complexes through cell lysis and ultrasonic disruption. Finally,

excessive cross-linking can make it difficult to expose an epitope or epitopes of the protein of interest so that it can be accessed by antibodies in the subsequent antibody pulldown steps (i.e., "epitope masking") [225, 226].

Consequently, the optimal window for degree of cross-linking may be very narrow and must usually be determined empirically through pilot experiments for particular cell types and particular transcription factors under study. The major parameters to be tested in such pilot studies are formaldehyde concentration, cross-linking time, and cross-linking temperature. Pilot experiments usually are performed at a fixed temperature (e.g., room temperature), and a fixed low formaldehyde concentration (e.g., 1%), while cross-linking time is usually the manipulated experimental variable [211, 214, 221, 222, 227].

5.3.2 SONICATION PARAMETERS

To perform chromatin immunoprecipitation, some method must be used to shear the chromatin into a manageable size distribution as well as separate the chromatin from other cellular components. Through empirical determination of numerous laboratories performing ChIP, the optimal chromatin fragment size distribution is about 200 to 2000 base pairs. There are two major methods of accomplishing this, restriction endonuclease digestion and ultrasonic disruption. Restriction endonuclease digestion effectively shears double-stranded DNA into fragments, but does not separate chromatin from other cellular components. When enzymatic methods are used, they are often used in addition to ultrasonic disruption [211].

Ultrasonic cellular disruption, or sonication, is performed by inserting an ultrasonic transducer probe into the suspension of collected cells that have been treated with formaldehyde. The transducer vibrates at ultrasonic frequencies (~40,000 Hz) and produces cellular disruption

by inducing localized cavitation. Cavitation occurs when a localized increase of enthalpy (thermodynamic measure of internal energy and pressure) causes liquid water to flash to vapor, producing tiny bubbles. When the bubbles move to adjacent areas of lower enthalpy, they collapse asymmetrically at supersonic speeds causing shock waves. The resultant shock waves are violent and energetic enough to break covalent bonds of cellular macromolecules in the vicinity [228].

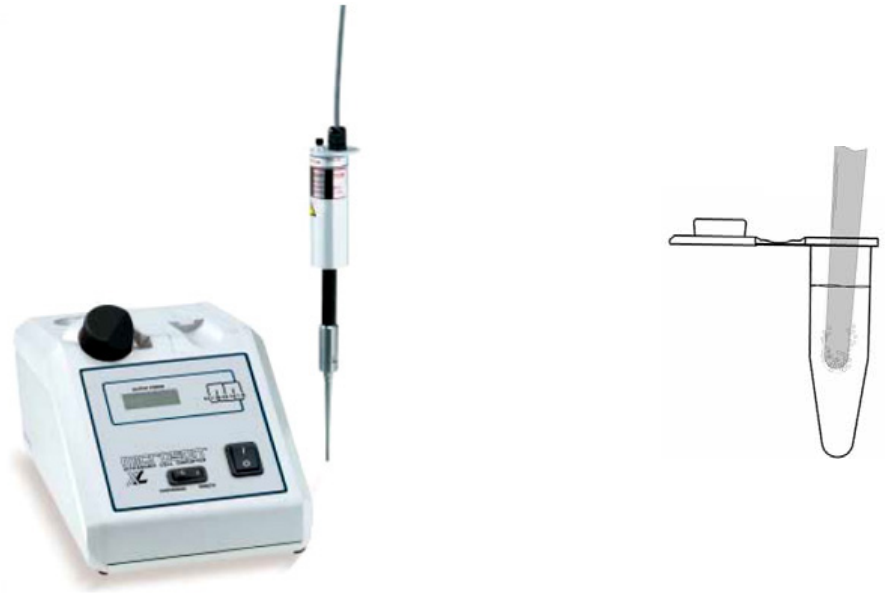


Figure 5-7. Ultrasonic cell disrupter, or sonicator (left). Illustration of sonicator probe immersed in ChIP lysate (right).

As such, sonication produces significant amounts of heat. Therefore in ChIP experiments, sonication is usually done in ten-to-twenty second intervals and placed in an ice bath to cool for perhaps sixty seconds in between [222, 229]. Nonetheless, brief periods of micro-localized heating is enough to fracture or denature proteins, and importantly, this includes the transcription factor(s) of interest. While under-sonication will not produce sufficiently lower fragment size

distribution, over-sonication can seriously degrade the transcription factor(s) of interest as well shift the chromatin fragment size distribution towards too low a size. As in the case of cross-linking, sonication parameters also must usually be determined empirically for a particular experimental system [222].

I have found that a number of sonication parameters play a key role in determining the fragment size distribution, and ultimately, the success of a ChIP experiment:

- Power setting
- Sonication make, model, and probe type
- Volume of lysate
- Geometry of microcentifuge or centrifuge tube
- Depth of probe tip

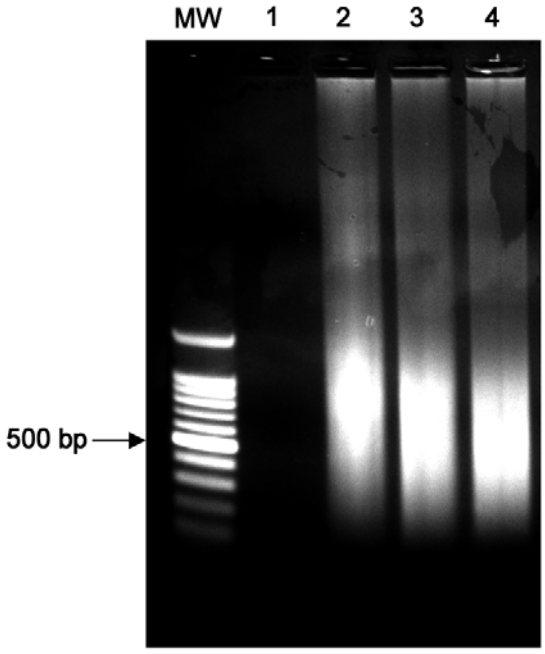


Figure 5-8. Agarose gel showing size fragment distributions after sonication.
MW: molecular weight marker; Lane 1: no sonication; Lane 2: 2x10 seconds; Lane 3: 3x10 seconds; Lane 4: 4x10 seconds.

5.3.3 ANTIBODY

Another potential source of variation in ChIP is the antibody used for the chromatin pulldown step. Ideally, the antibody should recognize and bind the appropriate epitopes on the transcription factor of interest. However, as mentioned previously, antibodies are typically made against native proteins [230]. Cross-linking of the transcription factor protein during formaldehyde treatment can alter the protein's conformation, reducing or eliminating antibody recognition of the cognate epitope(s). Just as important, the epitopes of the transcription factor of interest must be presented on an outer surface so that the antibody has physical access to it. A transcription factor buried deep inside a large complex of transcriptional machinery proteins may not be physically available for antibody recognition. Therefore, there is no guarantee that even if an antibody is available for some particular transcription factor of interest, that it will work successfully for ChIP [214, 223, 225, 226].

Even if an appropriate ChIP antibody is found, the sensitivity and specificity of antibodies can vary between manufacturers and even between batches from a single manufacturer. The choice of antibody is limited to what is available. Again, the choice of antibody from among available antibodies should be determined empirically by their ability to bring down the intended transcription factor with minimal background. Both polyclonal and monoclonal antibodies have been successfully used for ChIP [211, 222].

5.3.4 STRINGENCY WASHES

Once the chromatin fragments in the sonicated cell lysate are bound by an antibody specific to the transcription factor(s) of interest, the antibody/TF complex must be separated

from non-specific background chromatin and cellular debris. Protein A and/or Protein G are most commonly used to bind antibodies.

Protein A and Protein G are surface proteins found in the cell wall of bacteria. Protein A is found in *Staphylococcus aureus*, while Protein G is found in Strepotococcal bacteria. Each has a very high affinity for the Fc heavy chain of IgG_1 and IgG_2 in humans as well as most common animal sources of antibodies [231, 232]. Since their binding affinities can be manipulated by altering salt concentration and pH, both are commonly used in Sepharose columns to select and purify antibodies. In ChIP protocols, Protein A and/or Protein G conjugated to agarose or magnetic beads are most commonly used to bind chromatin/Ab complexes.[3]

After Protein A/G binding, non-specific chromatin and cellular debris must be removed by a series of washes. Because Protein A/G as well as the agarose or magnetic beads themselves can bind some proteins and nucleic acids non-specifically, ChIP protocols typically recommend somewhere between four and eleven washes [233, 234]. Sometimes these washes start with lower stringency, low-salt washes and progress to high-salt, higher-stringency washes [234]. The last wash nearly always consists of a standard Tris-EDTA buffer (TE) to remove traces of previous wash buffers whose constituents may interfere with subsequent steps. There is no clear cutoff for the number of washes—more washes improve specificity but reduce overall signal. Therefore, for a particular ChIP/antibody experiment system, the number of bead washes must be determined empirically from pilot experiments [233, 235]. Non-specific binding as well as

[3] The term "immunoprecipitation" in "ChIP" is a historical misnomer held over from traditional protein isolation experiments. These experiments used ionic conditions to truly chemically precipitate Ab-protein complexes, which were subsequently isolated into a pellet by centrifugation. In ChIP, Ab-protein complexes are not actually precipitated at all, but rather "captured." The title "Chromatin Immunocapture" or "ChIC," despite being a more accurate title, has not gained popularity.

over-washing-induced signal reduction can add to the variability and lack of consistency in ChIP results between separate experiments.

5.3.5 CROSSLINK REVERSAL

After antibody pulldown, the next step in the ChIP procedure is to reverse the formaldehyde-induced cross-links by applying a combination of heat (60ºC) and high salt concentration. Published protocols vary, but experience shows that when samples are heated more than 8-12 hours the signal intensity of subsequent microarrays is reduced [229, 233, 235].

5.3.6 FRAGMENT AMPLIFICATION

Potentially the greatest source of bias and variability in the ChIP-on-chip procedure is introduced during ligation-mediated PCR [223]. One of the drawbacks of ChIP is that it results in a very small amount of material. Promoter microarrays, on the other hand, require a large quantity of starting DNA, usually on the order of micrograms. The amount of amplification necessary is in the range of one thousand to a million fold. Since the sequences of the ChIP nucleic acid fragments are unknown, the most common solution is to amplify them through ligation-mediated PCR [236, 237].

Before ligation-mediated PCR can be done, ChIP nucleic acid fragments must be blunt-ended using a DNA polymerase. This is necessary when the chromatin has been sheared by sonication because the DNA phosphate backbone will be broken randomly in any number of places. This results in the double-stranded DNA fragments having 5’ or 3’ overhangs of varying lengths. After the fragments are blunt-ended, blunt-end oligonucleotide linkers or adapters of

known sequence are ligated onto each end of the ChIP fragments. The resultant ligation product is then amplified using PCR with primers complementary to the linker sequence.

There are some drawbacks to ligating linkers onto DNA fragments resulting from ChIP. First, ligation efficiency and specificity is severely limited because the fragments must be blunt-ended. Blunt-end ligation does not provide the same yield as ligation performed on complementary sequences of overhangs (i.e., "sticky ends") [237]. Second, the desired ligation product is the target DNA fragment flanked on each side by linkers. However, linker ligation also produces such unwanted products as linker-linker multimers, linkage between target fragments, and linker ligated to only one side of the target sequence (see figure 5-9).

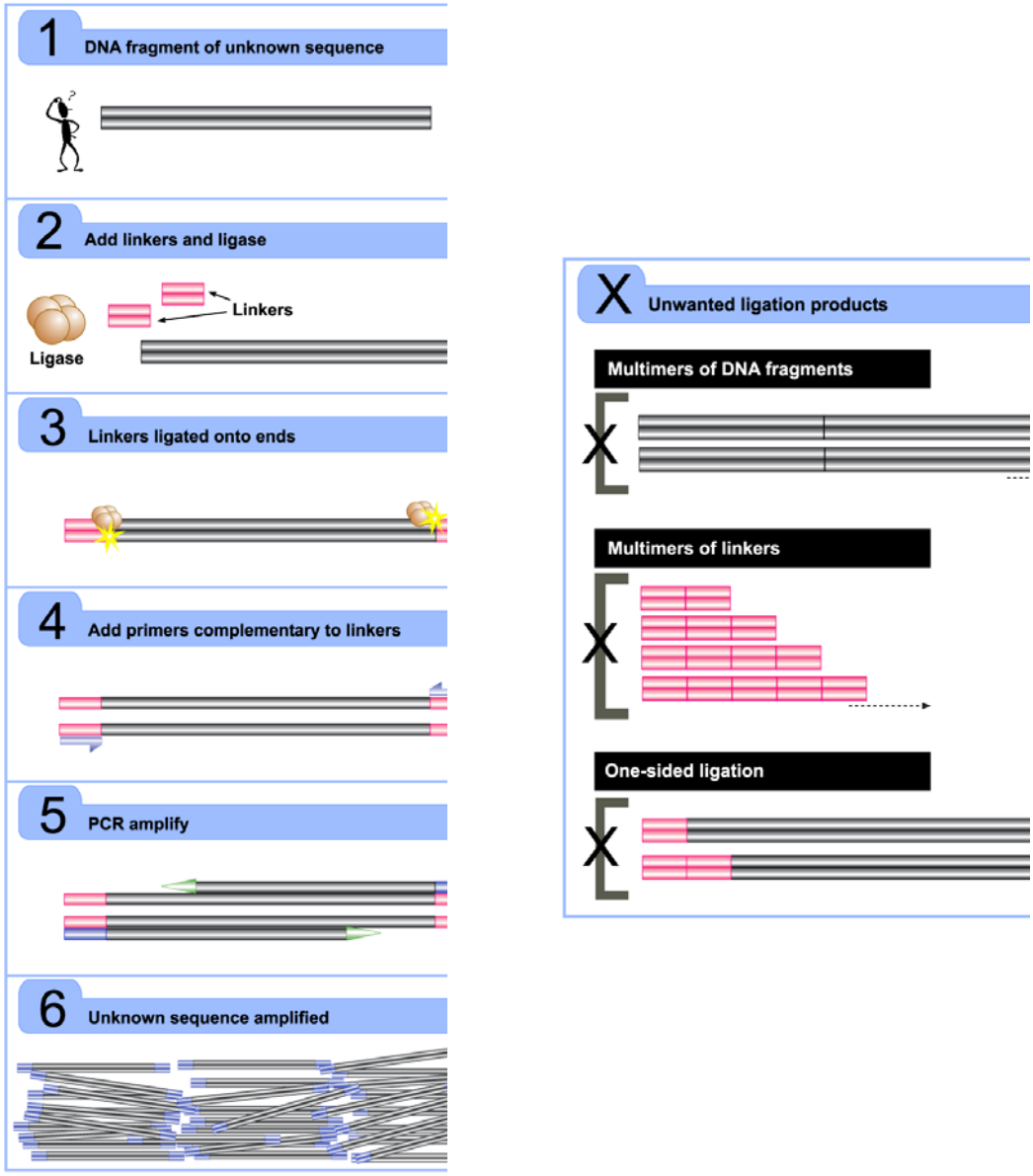


Figure 5-9. Schematic overview of ligation-mediated PCR (LM-PCR) process (left). Examples of unwanted ligation products that reduce sensitivity, specificity, and yield of ChIP-chip when performed using LM-PCR amplification (right).

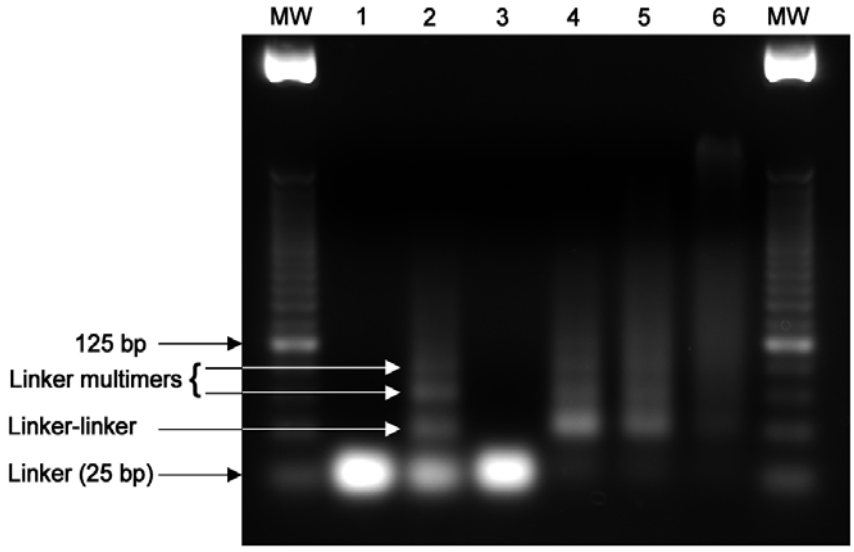


Figure 5-10. Test of ligation reaction conditions using self-ligation of a 25 base-pair blunt-end dsDNA linker oligonucleotide. 25 bp molecular weight marker flanks lanes; 3% agarose gel. Lane 1: No ligase negative control, room temperature 10 minutes; Lane 2: 2000 units T4 ligase, room temperature 10 minutes; Lane 3: No ligase control, 16°C 30 minutes; Lane 4: 500 units T4 ligase, 16°C 30 minutes; Lane 5: 1000 units T4 ligase, 16°C 30 minutes; Lane 6: 2000 units T4 ligase, 16°C 30 minutes.

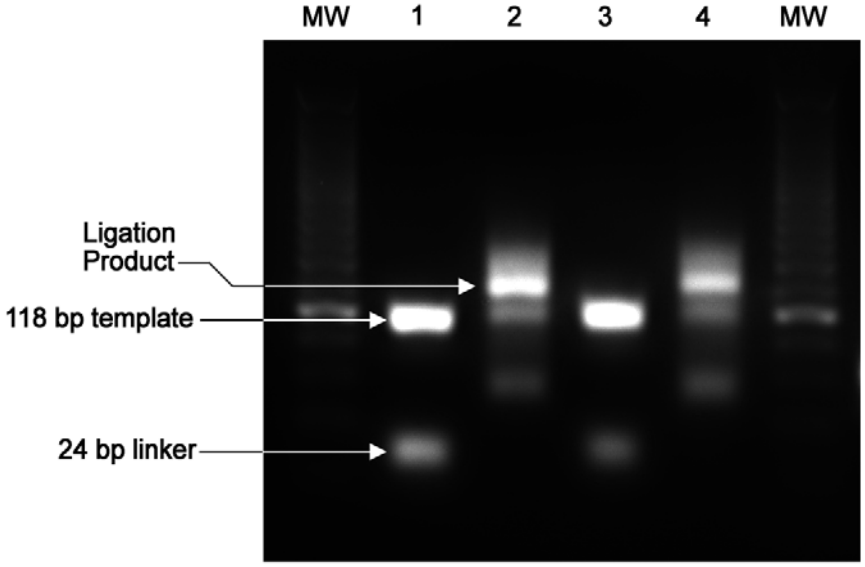


Figure 5-11. Test of ligation reaction conditions using 118 bp β-actin test sequence. MW: molecular weight marker; Lane 1: 24bp linker + 118 bp template, no ligase; Lane 2: 24bp linker + 118 bp template, 500 units T4 ligase; Lane 3: 24bp linker + 118 bp template, no ligase; Lane 4: 24bp linker + 118 bp template, 1000 units T4 ligase.

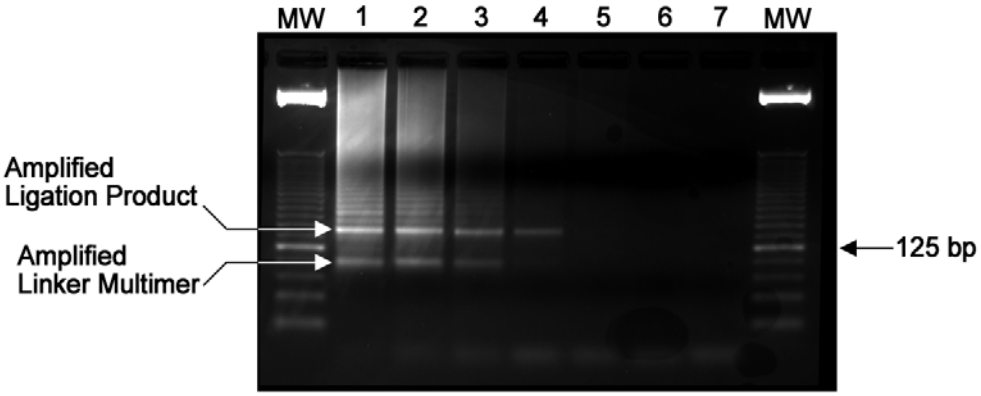


Figure 5-12. Test of ligation-mediated PCR using 24 bp linkers ligated to118 bp β-actin test sequence. MW: 25 bp molecular weight marker. Lane 1: 100 ng template; Lane 2: 10 ng template; Lane 3: 1 ng template; Lane 4: 100 pg template; Lane 5: 10 pg template; Lane 6: 1 pg template; Lane 7: 100 fg template.

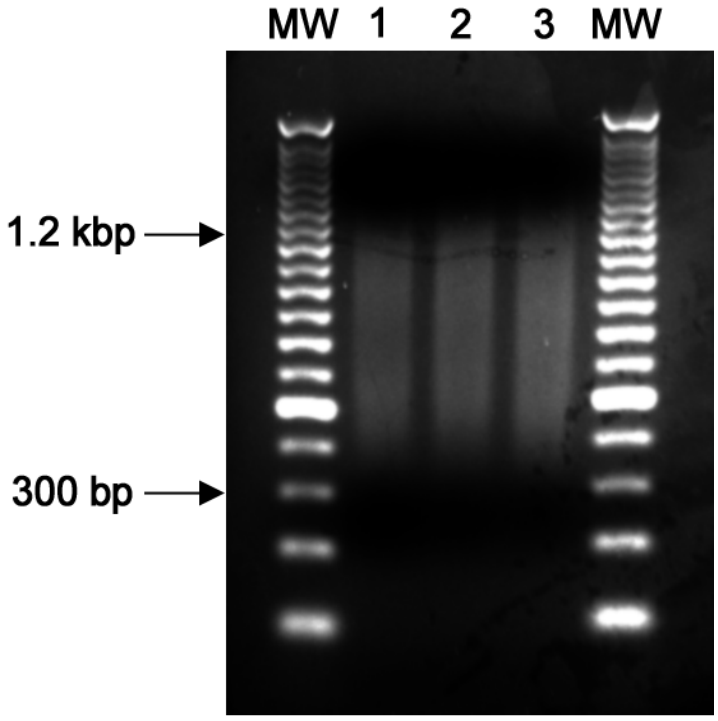


Figure 5-13. Agarose gel evaluation of ligation-mediated PCR of actual ChIP pulldown procedure. Lane 1: SMAD3 IP from TGFβ-stimulated A549 cells; Lane 2: SMAD3 IP from non-stimulated control A549 cells; Lane 3: Mock IP from non-stimulated control A549 cells.

LM-PCR is also known to introduce bias based on sequence and input fragment size [203]. Ligation efficiency is skewed toward shorter DNA fragments, and subsequent PCR amplification exponentially increases this bias [238]. Additionally, *Taq* polymerase commonly used in PCR amplification misincorporates nucleotides on the order of 10^{-4} errors per nucleotide, and additionally lack proofreading ability [239]. Alternative, higher-fidelity heat-resistant DNA polymerases derived from sources such as *Pfu* and *Vent* have lower error rates and therefore might be preferable to traditional *Taq* polymerase [240-242].

There are some alternatives to LM-PCR as a DNA amplification technique. Most are whole genome amplification methods developed and used for such applications as SNP detection or comparative genome hybridization (CGH) [243]. One non-PCR-based whole genome amplification technique used successfully to amplify very low quantities of genomic DNA is called multiple displacement amplification (MDA) [244-246]. In this method, random hexamer primers are annealed to denatured, ssDNA fragments. Phi29 DNA polymerase is used to fill in the complementary DNA strand, subsequently displacing the adjacent hexamer primer and complementary strands. It has been used to amplify genomic DNA up to 1000-fold without introducing sequence bias However, despite these attractive attributes, my pilot experiments have shown that Phi29-based MDA does not work well for amplifying short DNA fragments such as in the case of ChIP.

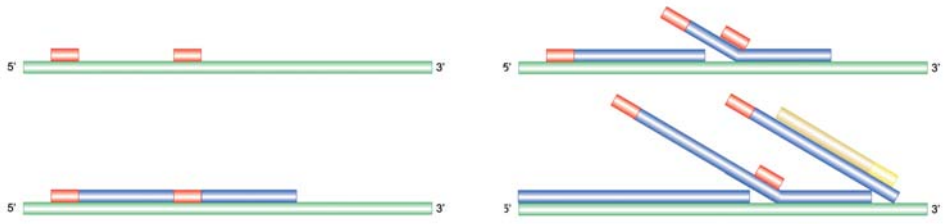


Figure 5-14. Multiple strand displacement amplification (MDA) process

Finally, T7-based linear amplification of DNA (TLAD) uses Alu I restriction endonuclease digestion and a terminal transferase to add a polyT tail on the 3' end [247]. A primer containing a 3'polyA tract and 5' T7 promoter is annealed to the polyT tail. Taq polymerase is used to synthesize the second strand. The dsDNA is then placed into an in vitro transcription reaction followed by reverse transcription, producing numerous cDNA copies of the original sequence. While this technique has the advantage that it does not introduce sequence and length-dependent biases, it is not commonly used due to the time-consuming, cumbersome protocol [248].

5.3.7 FLUOROPHORE LABELING

Once DNA fragments are amplified, they must be labeled with a fluorophore for microarray analysis. The three most common DNA fluorophore labeling methods are direct labeling, indirect labeling, and dendrimer labeling (Genisphere, Inc., Hatfield, PA) [249, 250]. The direct and indirect methods involve denaturing of the input DNA into single strands. The ssDNA is either reverse transcribed into complementary RNA (cRNA) or used as a template for synthesizing complementary DNA (cDNA). The direct labeling method incorporates Cy3-dUTP or Cy5-dUTP fluorophore directly into either the newly-synthesized cRNA or DNA. This is often done by end-labeling or random-primed, Klenow-based extension [233]. In the Klenow-based extension, an equimolar mix of dATP, dCTP, and dGTP is added along with a reduced molarity of dTTP. The deficit of dTTP is made up with either Cy5- or Cy3-dUTP, which will be incorporated into the new strand by the Klenow fragment DNA polymerase.

The indirect labeling method incorporates a modified (amino-allyl) dUTP during reverse transcription or 2nd strand cDNA synthesis. The fluorophore is then subsequently covalently

bonded to the (amino-allyl) dUTP in the resultant cRNA or cDNA. Dendrimer labeling uses a multi-armed molecular (dendritic structure) whose arms contain fluorophores. In dendrimer labeling, the reverse transcriptase or cDNA synthesis reaction is primed with an oligonucleotide containing a dendrimer “capture” sequence. The cDNAs containing the “capture” sequences are hybridized to the dendrimers whose arms contain multiple (~200) fluorophore molecules. The cDNA/dendrimer/fluorophore complex is then hybridized to the microrarray [250].

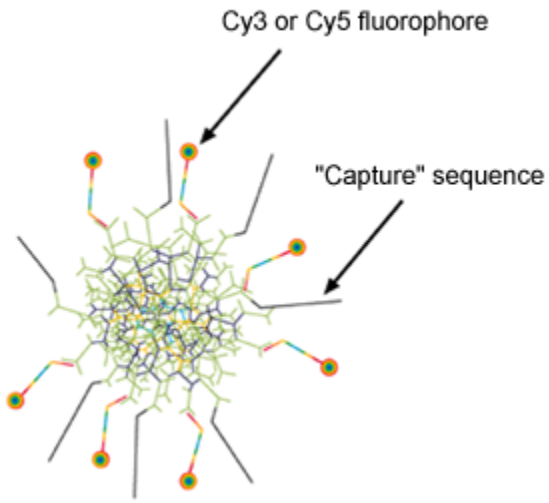
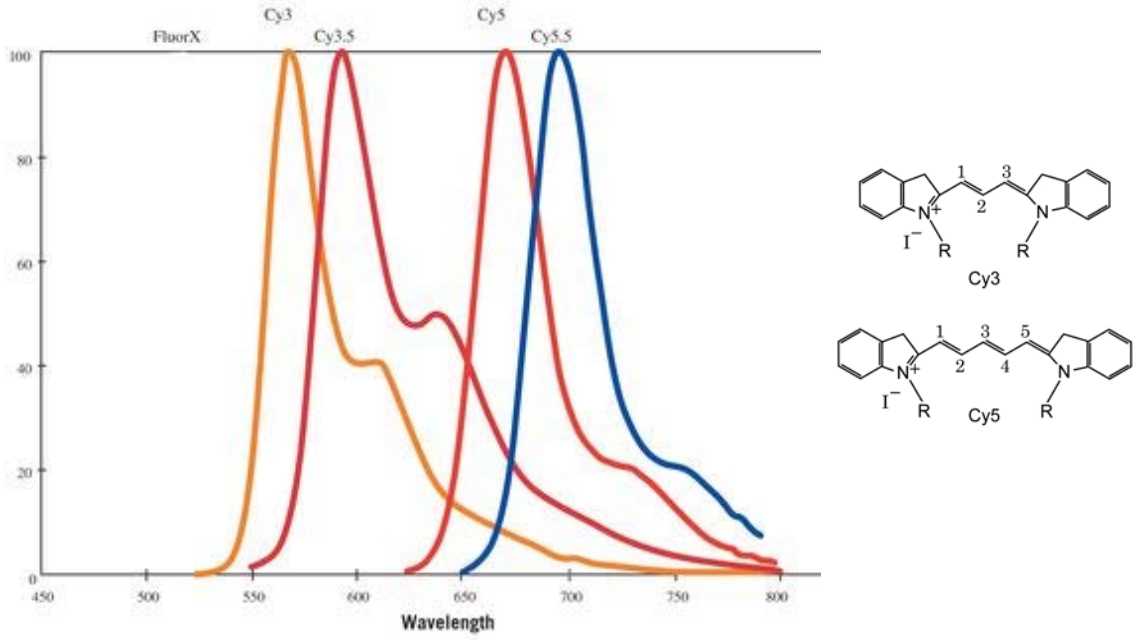


Figure 5-15. Dendrimer (Genisphere, Inc.)



A.1.3

Figure 5-16. Cy-3 and Cy-5 normalized excitation and emission spectra (left) and Cy3/Cy5 molecular structures (right).

There are several limitations of using fluorophores, in this case Cy-3 and Cy-5 in particular. First, since their emission spectra overlap, there is a small amount of signal confounding between probe/target complexes using the different dyes. Also, Cy-3 and Cy-5 are different sized molecules (see Figure 5-11) and may incorporate into newly synthesized DNA at slightly different rates during the labeling reaction step. Therefore the labeling efficiency between the two dyes can introduce dye bias. In addition, in direct and indirect labeling of DNA, dye incorporate is sequence-dependent, i.e., fluorophores are incorporated in some but not all of the T nucleotide positions. More significantly, both dyes, but especially Cy-5, are subject to degradation by airborne ozone. Thus, the signal intensity of Cy-5 probe/target complexes on microarrays can diminish significantly in a matter of minutes. Since microarrays are laser-scanned one at a time, Cy-5 signal can vary significantly from one scanned array to the next. Even worse, the Cy-5 signal can fade from one side or corner of a slide to the other. To help

combat this problem, microarray manufactures often offer anti-oxidant treatments for their arrays (e.g., Agilent's proprietary Stabilization and Drying Solution), which appears to be somewhat but not completely effective.

5.3.8 PROMOTER MICROARRAY DESIGN, ANALYSIS, AND LIMITATIONS

There are a number of microarray platforms, although most are gene expression arrays. ChIP-chip by definition requires microarrays that cover significant portions of the promoter regions of the human genome. Currently, two manufacturers dominate the ChIP-chip promoter microarray market.

Agilent whole genome human promoter arrays contain an average four to six separate probes 60-mer oligonucleotide probes spaced at approximately 100 to 300 base pair intervals. The probes cover ~2000 base pair upstream to ~800 base pair downstream of the transcriptional start site for 18,002 transcriptional start sites representing 17,917 RefSeq genes. The probe sequences have been designed specifically to limit cross-hybridization and other artifacts. First, the probes are designed to avoid areas of repetitive nucleotide sequences. Also, the predicted melting temperature of the probes are optimized to be as close as possible for all probes [251]. The arrays also contain approximately 2000 positive and negative control probes. Of these 663 probes are includes that have sequences unique to *Arabidopsis thaliana* genes. They serve as negative controls for non-specific hybridization. The arrays also contain 616 "gene desert" negative controls identified from intergenic regions 1 Mb or greater which are assumed to not be bound by promoter-binding transcriptional regulators [252].

NimbleGen human promoter arrays contain 50-75-mer probes with approximately 100bp spacing between probes. They have several designs based on RefSeq sequences that cover ~3000 to ~5000 base pair promoter regions from among ~11,000 to ~18,000 separate genes [253].

5.3.9 PROMOTER MICROARRAY ANALYSIS ISSUES

The ultimate purpose of any microarray measurement is to gain biological knowledge. Ideally then, the fluorescence intensity measurements of target/probe complexes on promoter microarrays will reflect actual biological information from the input sample. However, this information is often embedded into and obscured by numerous sources of random and systematic errors. These errors are unavoidable and generated by experimental procedures and measurement techniques. Ideally, the purpose of any data manipulation steps on raw microarray data is to systematically minimize these errors so that the information best reflecting true biological reality shines through. In essence, microarray data is a collection of numbers that must be "calibrated" to the biology.

Two general issues arise with microarray data: one is calibrating the data within a microarray to minimize the effects of biases or errors; the other is calibrating differences between microarrays. This calibration is known broadly as normalization. For within array normalization, the goal is to remove systematic bias. For between array normalization, the goal is to make the between-array data comparable by standardizing their ranges. Within-array and between-array bias can come from a number of known and unknown sources. These include: Total fluorescence intensities differ between arrays, both due to features of the arrays themselves and due to the laser scanner used. Also, in dual-dye systems, one dye may be stronger than the other across-the-board, or be stronger than the other as a function of total fluorescent intensity.

This may vary within an array and also between separate arrays. Finally, there will always be at least a small amount of background or residual fluorescence on arrays, and this background can be different depending on area within the same area, and between separate arrays.

For instance, consider two arrays. Array one's intensity values of one array range from 100 to 1000 and array two's intensity values range from 0 to 10,000. The number 1000 on the first or 10,000 on the second are somewhat arbitrary and therefore meaningless without comparison to a biological standard. Therefore, the array values are normalized to each other so that the ranges can be made comparable.

In the absence of a true biological standard, there are several conventional choices for normalization methods for microarray data. One normalization method is simply to scale the array data globally by some constant factor so that they each have same mean or median. Another method is to adjust the intensity values of array data by some non-linear function so that their ranges agree. Examples of this are quantile normalization or lowess ("Robust Locally Weighted Regression and Smoothing Scatterplots") [254]. In quantile normalization, each data set is ranked in order of intensity values and then divided into quantiles. The quantiles are then individually adjusted to the average for that quantile. Quantile normalization forces each individual array dataset to have the same intensity distribution. Lowess normalization, on the other hand, is similar in concept except that it adjusts each data point individually instead of by quantile. Lowess forces the data at every intensity value to have an equal mean.

To assess arrays for dye bias on dual-dye arrays, we usually make an MA plot, where M = log(R/G) and A = log ($\sqrt{R \cdot G}$). If the relative intensity of each dye does not change as a function of overall spot intensity, then the scatterplot should generally follow a horizontal line. If

the plot shows a trend toward turning up or down, this indicates a need to use a normalization technique that ideally flattens and rotates the trend toward horizontal.

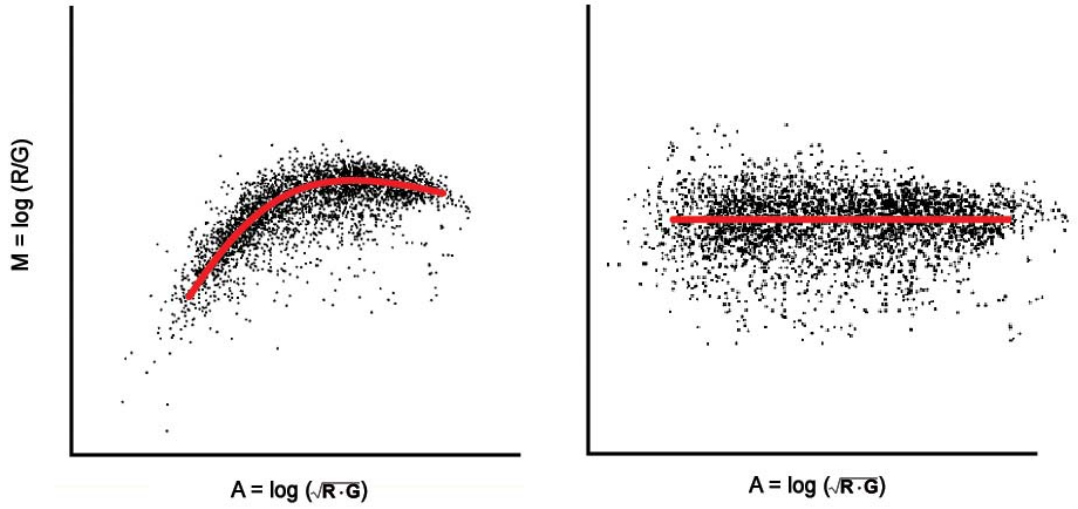


Figure 5-17. MA plot of microarray data, un-normalized data (left) and after lowess normalization (right).

The conventional approach to processing raw microarray data is discussed in more detail in the following chapter Gene Expression Microarray Technology and Methods. However, the following major steps apply to both gene expression microarrays and ChIP promoter microarrays:

- Microarray image quality assessment and filtering
- Feature (spot) identification and background subtraction
- Log_2 transform
- Normalization
- Analysis

Traditional normalization techniques were developed for gene expression microarrays, not promoter microarrays. One major assumption in gene expression microarrays is that there will be an approximately equal number of down-regulated and up-regulated genes. With promoter microarrays used in ChIP-chip, there is either promoter binding or no binding. Therefore, in a scatter plot of intensity values one would expect to see an asymmetric bulge or hump of data points corresponding to binding events. Finally, ChIP-chip usually uses dual-dye arrays, where the 2nd dye is a reference such as mock IP, no-antibody IP, or whole cell extract.

Additionally, tiled promoter microarrays use information from adjacent probes. Clusters of probe binding events within the same gene promoter region are more informative than single probe binding, as in the case of gene expression arrays.

With respect to mock IP/no-antibody IP/wce reference, the common approach is to subtract the reference from the IP signal. This, however, can introduce noise

As mentioned previously, promoter array probes are usually spaced ~100 to ~300 bp apart. Since promoter array target sequences typically range from ~300 to ~2000 base pairs, their positions can be expected to overlap a number of probes. True TF binding events, therefore, should be indicated by target hybridization to several adjacent promoter array probes. Promoter array analysis usually makes use of this fact that clusters of probe binding, especially those that center around a single point, normally indicates a true TF binding event. Most promoter microarray analysis techniques therefore plot probe intensities (y-axis) versus genomic position (x-axis), and use a "sliding window" to detect the presence of clusters of probe binding, and especially clusters of intensity values that loosely follow a characteristic "peaked" shape. This is illustrated in Figure 5-18 below.

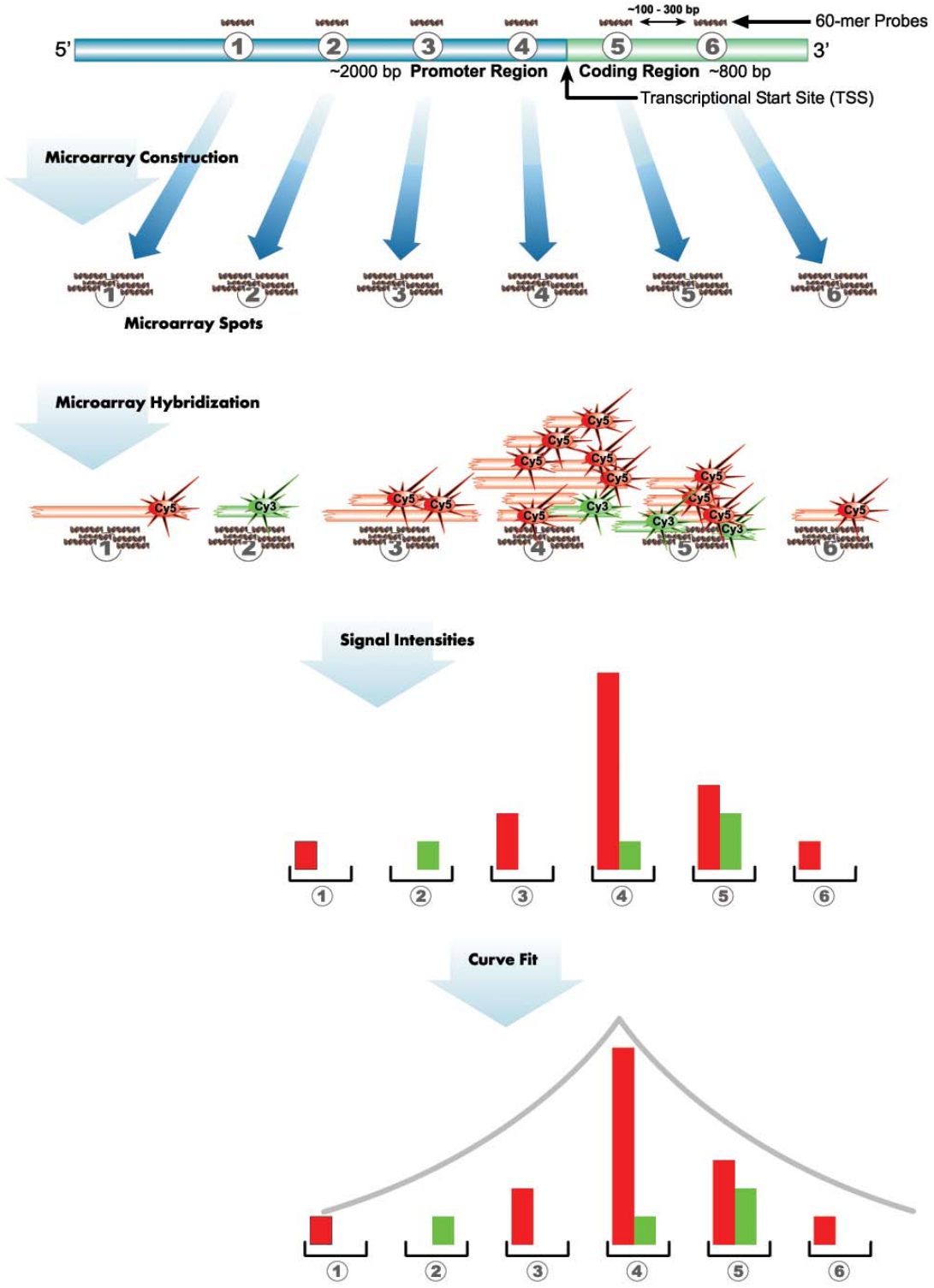


Figure 5-18. Promoter microarray design and interpretation.
Top: 60-mer probes are designed from sequences spaced at ~100 to ~300 bp intervals spanning the proximal promoter region and immediately downstream of the

transcriptional start site (TSS). Microarray Construction: Microarrays are spotted with 60-mer probes designed as described. Microarray Hybridization: ChIP is performed, yielding nucleic acid fragments labeled with Cy5 (IP; red) and Cy3 (Mock IP reference; green). These fragments are hybridized to probes on the arrays. Signal Intensities: The relative signal intensities reflect the amount of bound target to each probe, and can be plotted as a function of chromosomal position. Curve Fit: Non-specific target binding should be randomly distributed, whereas true promoter fragment target binding should roughly correspond to a peaked distribution surrounding the actual binding site. A curve fit to this model of promoter target binding is used to distinguish true binding events from non-specific signal (noise).

6 GENE EXPRESSION MICROARRAY TECHNOLOGY AND METHODS

6.1 CONCEPTUAL ORIGIN AND EVOLUTION OF THE MICROARRAY

As is well known, protein production in cells is mediated through translation of messenger RNA. Quantifying levels of each type of mRNA provides two kinds of information: indirect information about the protein-mediated molecular activities within cells, and direct information about transcriptional regulation *per se*. Both kinds of information are extremely important to life science and medical researchers. Traditional molecular biology approaches toward mRNA quantification, such as northern blotting, measure no more than dozens of mRNA species at one time. In northern blotting, RNA is first separated by size by gel electrophoresis, transferred and immobilized to a porous membrane, denatured, and detected by hybridization with a complementary nucleic acid probe conjugated to a molecular indicator. The molecular indicator is usually radioactive (e.g., end-labeling by P^{32}), colorimetric, or phosphorescence/fluorescence. When hybridization is complete, non-specific and excess indicator is washed away and the blot is imaged. The resultant band(s) location on the blot provides information about the size of the mRNA species, while band intensity serves as a measure of the original mRNA species quantity [255-258].

While northern blotting is a very specific measurement due to analysis by both nucleic size and sequence specificity, the process is laborious, cumbersome, and expensive. A less

laborious approach is to omit the electrophoresis step and simply perform hybridization, indicator development, and imaging. This type of nucleic acid detection and quantification is known as a "dot blot."

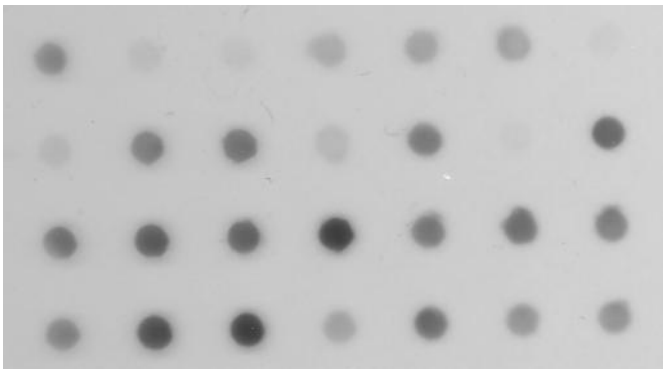


Figure 6-1. Example of a "dot blot."

From the mid-1980s through the mid-1990s, new technologies for measuring large numbers of mRNA sample levels grew out of the dot blot concept [259-265]. What restricts the number of samples that can be run at one time on northern blotting is the requirement for electrophoresis, which takes up a great deal of space on a two-dimensional membrane. When this requirement is removed (e.g., dot blot), the number of samples that can be run at one time is limited by the size of each individual dot. Thus, one can easily analyze dozens or hundreds of mRNA samples on a suitable dot blot using suitably small dots arrayed into an efficient geometric arrangement.

Since cells from *E. coli* through humans contain many thousands of mRNA transcripts[4], it is even more attractive to researchers to be able to run mRNA from an entire transcriptome at once, i.e., thousands of mRNA transcripts. This massively parallel, "whole transcriptome" approach provides a comprehensive snapshot of the transcriptional activity of cells at one point

[4] At the time of this writing, the number of known human RNA transcripts is 48,400 [199].

in time. One limitation of the dot blot array, even in miniaturized form, is that the biological samples must be spotted individually, and the array is probed by the probe(s). By reversing the target/probe relationship, the microarray was born.

The nucleic acid microarray was pioneered by Mark Schena and Pat Brown at Stanford [27, 266, 267]. The concept of the microarray is to immobilize hundreds or thousands of sequence-specific probes on a substrate, and then allow the nucleic acid target (i.e., the biological sample) to hybridize to the stationary probes. The target, not the probe, is labeled with an indicator (typically fluorescence). By miniaturizing the size of the probe footprint (dot) to the order of microns, and immobilizing thousands of probe dots in an efficient spatial geometry on the substrate, one can simultaneously measure thousands of mRNA transcript levels. Also, since DNA is more chemically stable than RNA, DNA probes are usually used on the arrays. Thus, this massively parallel mRNA measurement technology became known as DNA microarrays.

There are a number of general technological variations of the DNA microarray. The first microarrays used full-length cDNA probes (variable length, and up to100s of base pairs), often generated from EST mRNA and IMAGE clones [268, 269]. Arrays were often custom-made by researchers in their own labs [30, 270]. This type of DNA microarray design suffers from a number of limitations, including wide variations in melting cDNA probe melting temperatures (resulting in non-specific hybridization bias based on variable probe length and C/G content) and propensity for cross-hybridization artifact [271].

The first major commercial manufacturer of DNA microarrays was Affymetrix. To avoid the non-specific hybridization problems associated cDNA, Affymetrix introduced two novel features. One was to use ~25 base pair oligonucleotide probes instead of full-length cDNA. Also, to help distinguish between specific and non-specific hybridization, Affymetrix arrays use two

kinds of probes for every mRNA species measurement: perfect homology 25-mer probes ("perfect match" or PM) and a probe with a single nucleic acid mismatch ("mismatch" or MM). In principle, by subtracting probe intensity of the MM probe from the PM probe, one can subtract out any effect of non-specific probe hybridization. In practice, users and development engineers at Affymetrix found that this scheme did not work as well in the real world, and retreated to using solely the PM probe values .

Most commercial microarray manufacturers now use oligonucleotide probes of a constant length or within a small defined range (e.g., 60-mer). Oligonucleotide probes are often carefully designed from curated genomic sequence databases to meet a number of criteria, such as minimal nucleotide repeats, similar G/C content, similar melting temperatures, minimal complementarity/self-annealing, etc [272]. The purpose of this is to reduce artifact and bias due to non-specific probe/target hybridization and probe cross-hybridization.

6.2 OVERVIEW OF MICROARRAY WORKFLOW

In principle, the concept behind DNA microarray measurements seems like it would be a very robust technology. In practice, however, the process is lengthy and the technology complex, resulting in numerous places for introduction of biases, errors, and outright mistakes. While microarray manufacturers have made great technological advancements in producing arrays with better reliability, accuracy, and reproducibility, it is instructive to understand their design features and workflow, and therefore some of their limitations.

6.2.1 RNA EXTRACTION

The first step in doing microarray-based measurements is to extract and purify mRNA from the biological samples of interest. Among the number of RNA isolation methods, the most commonly used is based on the guanidinium isothiocyanate method of Chomczynski and Sacchi [273]. Guanidinium salts are powerful chaotropic agents that denature proteins and protect RNA against degradation. It is usually used as a mixture with phenol, which is sold under the trade name Trizol® (Invitrogen). Cells or tissues are placed in the guanidinium/phenol solution and agitated, lysing the cells and breaking up and solubilizing their constituent parts. The mixture is centrifuged and the mixture separates into an upper aqueous phase, an insoluble interphase, and a lower organic phase. RNA partitions in the upper aqueous phase, while lipids, proteins, and DNA partition into the interphase and lower organic phase. The aqueous phase is then recovered and the RNA concentrated and purified from it by isopropanolol or ethanol precipitation.

An alternative RNA purification and concentration method is known as a spin column. The spin column is a small microcentrifuge tube containing a tiny silica disk that tightly binds nucleic acid at lower pH, and releases it at high pH. The tube has an upper chamber connected through the silica flow-through disk to a lower collection tube. The aqueous solution containing RNA to be purified is mixed with an acidic solution, loaded into the upper chamber, and centrifuged through the silica disk. Then a wash buffer containing ethanol is forced by centrifugation through the disk to remove impurities. Finally, a small volume of basic pH buffer is centrifuged through the disk into a new, clean microcentrifuge collection tube. RNA-purification spin columns let short RNA strands (<200 bp) flow through, so they may not be suitable for experiments dealing with short RNA species such as microRNAs.

6.2.2 RNA AMPLIFICATION AND LABELING

The amount of RNA from biological samples is usually not enough (on the order of nanograms) for the process of microarray hybridization (5-20 μg), especially when replicate microarrays are performed. The RNA sequence must therefore be amplified. Most microarray RNA amplification protocols are based on the original technique of Van Gelder and Eberwine which uses an oligo(dT) primer attached to a T7 RNA polymerase promoter [274]. The sample mRNA, which contains a 3'-end poly(A) tail, binds the complementary oligo(dT) primer. Moloney Murine Leukemia Virus (MMLV) reverse transcriptase and dNTPs are added to synthesize 1^{st} and 2^{nd} strands of complementary DNA sequence (cDNA). This cDNA, which contains the T7 promoter at the 3' end of the original sample mRNA, is then used as a template for generating multiple copies of antisense cRNA with T7 RNA polymerase in an in vitro transcription reaction (IVT). For direct Cy3 or Cy5 labeling, Cy3 or Cy5 conjugated to CTP is added so that it will be incorporated into the cRNA strands as they are synthesized. Such a reaction will amplify input mRNA in excess of 100-fold [275]. As an alternative to direct labeling, indirect labeling incorporates amino allyl UTP into cRNA during the IVT. The amino molecule allyl UTP is then chemically coupled to an indicator such as a fluorescent dye. Since the T7 RNA polymerase incorporates allyl UTP at a higher rate than the much larger CTP-Cy3 or CTP-Cy5 molecule, indirect labeling produces a much higher labeling density than direct labeling .

6.2.3 HYBRIDIZATION

Once a sufficient amount of fluorescently-labeled cRNA (or cDNA) is obtained, the samples are placed on the microarray surfaces so that any sequences in the sample complementary to the probes on the microarray will anneal. The annealing process is an equilibrium process whose kinetics are governed by concentration of reactants and products as well as the applicable rate constants k_a (association) and k_d (dissociation) [276].

Equation 6-1. $$[\text{probe}] + [\text{target}] \; k_a \underset{}{\overset{k_d}{\longleftrightarrow}} [\text{probe} \bullet \text{target}]$$

Equation 6-2. $$k_d = Ae^{\frac{-E_a}{RT}}$$

Further, the association and dissociation constants have a temperature dependence described by the Arrhenius equation (Eq. 6-2), where A is the Arrhenius constant, E_a is activation energy, R the gas constant, and T temperature (Kelvin) [276]. In the case of oligonucleotide annealing, the activation energy is influenced by the ionic properties of the solvent used in the hybridization solution.

Too low a hybridization temperature will promote association, meaning an increase in mismatch sequence annealing in addition to perfect match annealing. Too high a hybridization will promote dissociation, improving sequence annealing specificity but reducing total number of hybridizations and hence the fluorescent signal. There is therefore no perfect annealing temperature, only a tradeoff between signal strength and annealing specificity. Thus choosing an annealing temperature is an optimization involving how much loss of signal and how much non-specificity one is willing to tolerate. The effects of temperature on annealing can also be

influenced by the ionic properties of the hybridization solution (i.e., salt concentration, pH), and therefore are also parameters to be optimized in relation to annealing temperature.

Viscosity of the hybridization fluid also plays a role in the hybridization kinetics. Targets must come into close proximity to the probes. This can be achieved by diffusion, but is often encouraged by mechanical agitation or rotation of the array. The hybridization fluid under motion is then governed by principles of fluid dynamics. Higher fluid viscosity increases the static fluid layer at the boundary surface of the microarray which has two opposing effects. First, target velocity must be low enough to allow annealing to probes. Second, if target velocity is too low, then targets will not circulate extensively enough to find probes with a perfect complementarity match. Therefore, as in the case of temperature, an optimization balance must be struck between too high and too low a hybridization fluid viscosity.

To reduce non-specific hybridization caused by repetitive nucleotide sequences inherent in the human genome, a hybridization solution often contains competitor sequences containing repeat DNA. Even though the probes are designed to not contain nucleotide repeats, the target RNA or DNA will likely contain some. Therefore, non-labeled competitor repetitive nucleotide DNA such as human Cot-1 DNA is often added to the mixture.

The hybridization reaction will eventually reach equilibrium asymptotically, as all probe sequences have the chance on average to contact all complementary target sequences. The hybridization process is multifactorial and complex, and therefore in practice parameter optimization must be determined empirically (usually by the manufacturer) for a particular microarray platform.

6.2.4 MICROARRAY WASHING

After microarrays are hybridized for a sufficient amount of time, remaining unbound targets must be washed off. Additionally, a proper wash or series of washes will help remove any non-specific binding to the array probes or intervening blank areas by fluorescently-labeled target (i.e., "background"). The wash steps can also be designed as a series of washes in incrementally lower ionic strength solutions to help encourage dissociation of partially mis-matched targets from probes. As shown in Eq. 6-1, removal of reactants will shift equilibrium toward dissociation, so the duration of microarray washes plays a significant factor in the balance between probe/target specificity and signal strength.

A typical microarray wash sequence is therefore: initial wash in a low stringency sodium chloride-sodium citrate buffer solution containing a small amount of ionic detergent. This is followed by one or more washes in higher-stringency diluted buffer of the same constituents. Often one or more finishing washes are done in an organic solvent to remove traces of water and any lipid-soluble residues.

6.2.5 MICROARRAY SCANNING

Microarrays need to be imaged in a way that measures the presence and intensity of signal from the probe/target complexes. Most commonly, the target label is a fluorophore such as Cy3 or Cy5. Imaging is accomplished through a laser-based excitation scan and emission measurement by a sensitive photomultiplier tube. Cy3 has its peak excitation at 550 nm and emits maximally at 570 nm, in the red part of the visible spectrum [277]. Cy5 has its peak excitation at 649 nm and emits maximally at 670 nm, in the far red part of the visible spectrum

[277]. Often, the laser scanning is done in a confocal manner, meaning the beam and photomultiplier tube imaging system are both focused at a particular point on the surface of the microarray slide. The system in tandem then scans the surface of the slide in a raster pattern. A false-color image of the slide is reconstructed from the raster pattern. This image shows a characteristic pattern of circular features or dots containing the probe/target/fluorophore complexes. Non-specific target/fluorophore binding will show up as artifact features or a widespread "haze" that constitutes the microarray background. Similarly, defects such as scratches or foreign body contaminants (e.g., lint) can be seen on the image as well. Background and artifacts are confounding intensity signals whose effects must be minimized with image analysis and statistical techniques.

In early versions of microarray image analysis programs, individual target/probe spots had to be identified on a grid by a human operator. Most recent microarray platforms offer an integrated system of microarray, laser scanner, and image analysis program designed to work together. If the layout and dimensions of the probe grid is known and can be aligned through some kind of registration marking, then the image analysis program can automatically identify each probe feature. Automated scanning, feature recognition, and feature measurement greatly facilitates the high-throughput nature of microarray experiments.

Since the probe features consist of (ideally) a circular, uniform deposit of a particular species of probe, the common algorithm for measuring overall feature intensity is to integrate individual signal intensity measurements over a series of 5 or 10 micron increments. Sophisticated image analysis programs can also evaluate the quality of the features (e.g.., loss of signal in the spot center, or "donuting"). Image analysis programs also measure signal background intensity adjacent to each probe feature, and subsequently subtract that value from

the feature signal intensity to yield a net signal. Once all the probe feature signal intensities are measured, the information is stored in a large flat-file database.

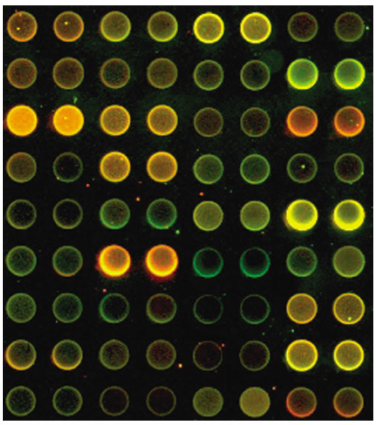


Figure 6-2. Example of microarray artifacts ("donut" spotting, background spots).

6.2.6 ARRAY NORMALIZATION

Microarrays measure mRNA levels indirectly through a chain of biochemical steps culminating in a measure of fluorescence intensity. As such, they do not measure mRNA directly, but are surrogate measures of the mRNA levels in the originating biological samples. Therefore they are not usually calibrated to known biological standards. While the absolute value of array intensity values from an array may not immediately be interpreted for biological meaning, the *relative* relationship between intensity values may reflect the biological reality of the input sample. Individual intensity values in large microarray raw datasets therefore cannot be assumed to be comparable until suitable corrections have been made for systematic errors and biases [29]. We refer to this process as normalizing the data. Within and between arrays there may be many sources of systematic bias. For instance, there may be slight differences in starting

mRNA amounts, differences in efficiency of fluorophore incorporation, differences in array washes can make arrays differ in their absolute intensity values, etc. These sources, however seemingly slight, can end up having profound effects in the range and distribution of intensity values from a particular microarray.

Microarray data therefore undergo a series of steps aimed at removing meaningless artifact and improving the ability to interpret the data in a biological context. In general these steps often include:

- Image quality assessment, feature (spot) identification, and filtering—The microarray image is assessed for excessive background, image artifacts, and spot quality (i.e., circularity, evenness of spot intensity). Spots that do not meet pre-set criteria for quality are often removed from the analysis (i.e., filtering) as their intensity values represent artifact, not biological meaning.
- Background subtraction—Background fluorescence intensity is measured adjacent to each probe spot and subtracted from the spot intensity to avoid misinterpreting background as biologically-relevant signal.
- Log_2 transform—As mentioned previously, absolute fluorescence values are essentially meaningless without reference to another value. Commonly, the reference value is provided within each spot by using two-dye arrays, with one dye corresponding to a reference (e.g., control) and the other a separate condition of interest (e.g., treated). In this case, the ratio between dyes (e.g., R/G), not either value separately, is the value to be interpreted. Similarly, in a single-dye array, the ratio between spots corresponding to the same gene (e.g., $R_{1array1}/R_{1array2}$) in two separate arrays conveys biological meaning

(assuming they are normalized). Thus, with microarrays, it is nearly always ratios between two numbers that are interpreted.

Simple ratios, however, introduce a difficulty in interpreting data. A two-fold increase between two gene expression values would be 2, whereas a two-fold decrease would be 0.5. A decrease appears to be less of a difference to the human eye, even though the relative change is equal in magnitude albeit in opposite in direction. This is solved by transforming the data to a base 2 logarithm. Thus, $\log_2(2) = +1$, while $\log_2(0.5) = -1$, giving up- and down-regulated gene expression values a pleasing symmetry that is easier to interpret visually. Log transformation offers other advantages as well, such as allowing faster and more efficient computation of large datasets: multiplying ratios simply becomes addition of their log values.

- Normalization—a brief overview of normalization issues was presented in section 5.3.9 in the previous chapter.
- Analysis—see next section.

6.3 STATISTICAL EVALUATION

Normalized data are used in a variety of ways to gain biological knowledge. While gene expression microarrays are intended to measure mRNA levels, the values are relative to each other, not to a series of known biological reference standards. Therefore, microarray data is commonly used to derive differential expression values between the same gene transcript species under different conditions, disease states, or experimental treatments.

Differential gene expression information can be used in a number of ways to try to glean biological insight. First, they can be evaluated through statistical hypothesis-testing procedures to find statistically-significant up- or down-regulation of particular genes. When replicate array

measurements are used, replicate data points for each gene can be used in parametric or non-parametric statistical tests. Both statistical significance of change, as well as relative degree of change (i.e., fold-ratios) can be interpreted for biological meaning. Since thousands of statistical hypothesis tests are performed at once in this case, a large number of significant changes can be expected simply due to chance variation. Therefore, when performing thousands of simultaneous hypothesis tests on microarray data the p-values must be corrected for multiple hypothesis testing. The most stringent method is simply to divide the single-experiment p-value criterion for significance (e.g., 0.05) by the total number of tests (e.g., 10,000). In this example, called Bonferroni correction, only gene transcript changes whose test p-value are less than 0.05/10,000, or $p < 0.000005$, would meet the criterion for significance. Many statisticians feel this is an overly strict method of evaluating statistical significance, and so a number of less stringent alternative methods such as False Discovery Rate (FDR) have been developed [278]. FDR controls the expected proportion of false positives rather than setting a strict p-value cutoff such as 0.05. Therefore, if 100 genes are found to be significant, with an FDR of 0.3 we know that 70% of the findings are likely to be true and 30% may be false positives. This may be beneficial when we are looking to discover new genes, where too strict a cutoff would force us to overlook potentially important findings. The FDR setting for analyzing microarray results can therefore be tailored to how many false positives we are willing to tolerate in the specific context of our experiment.

Aside from hypothesis testing, one can also use computational methods for pure discovery, or in other words to *mine* data for biological meaning. This is normally accomplished through classifying gene expression measurements according to some property of the

measurements. The two categories of gene classification methods are class discovery, and class prediction [279-282].

Class discovery groups gene expression values by degree of similarity. This type of analysis is known as clustering. Clustering uses properties of the value to partition data points into subsets such that each subset of data shares some common trait. The key decisions in choosing a clustering analysis are what features or measures to use, what similarity measure to use, and what method for clustering values to use [279, 280].

The features on which data may be clustered can be fluorescence intensity values only, or intensity values in relation to experimental treatment group, array replicates, etc. For a similarity measure, some metric is chosen as a way to establish the "distance" between numbers. One example of the choice of metric is simply Euclidean distance, which is a familiar direct distance measure. However, other non-Euclidean measures can also be used such as Manhattan distance measures or Mahalanobis distance [279, 280]. These other non-Euclidean distance measures have other desirable mathematical properties depending on the context, such being scale-invariant [281, 282].

There are also numerous algorithms for choosing how to group similar items together. These often differ by whether one desires the data to segregate automatically according to certain criteria (hierarchical clustering, self-organizing maps), or segregate according to some pre-established rules (*k*-means clustering, where the number of clusters *k* is pre-determined) [279, 280].

Class prediction is usually pursued for more practical purposes, such as for in clinical settings to help predict prognosis or response to therapies [283-286]. The idea behind class prediction is that automated methods of recognizing patterns in data are used to classify and

predict properties of new data sets based on previous data. Specifically, this is referred to as supervised learning, and requires one or more training sets of data where the input data and their correct classifications are both given as inputs. The algorithm then adapts or "learns" how to classify future data [281, 282].

Differential gene expression values can also be inspected visually. In this case the microarray intensity values are converted to color intensities and plotted in a two-dimensional array known as a heat map. Typically, a heat map will list individual genes vertically, and individual arrays horizontally. The horizontal list will usually group the columns by replicates, and then by condition or treatment of the source biological sample. By using heat maps with clustering on one or more experimental or biological attributes, interesting trends or features of biological significance may be discovered from microarray data.

6.4 MAJOR LIMITATIONS OF ARRAYS

Despite an impressive example of high-throughput technology, gene expression microarrays have a number of limitations that need to be acknowledged. First, by their very nature microarrays depend on hybridization of complementary sequences between targets and probes. As mentioned previously, this is a probabilistic process that inherently ensures some proportion of non-specific, i.e., mis-match, hybridizations. Further, the nature of probes and targets also makes it likely that there will be some number of species that will hybridize specifically to a small region of the correct sequence contained on another probe giving rise to cross-hybridization.

The miniaturization process that gives microarrays its ability to measure thousands of gene expression levels simultaneously also introduces limitations. With probe spots on the order of hundreds of microns, the effect of eneven backgrounds and handling artifacts are magnified. For instance, a stray piece of lint on a macroscopic dot blot has a negligible effect on signal intensity—the same lint on a microarray may completely obliterate the signal from dozens of probes.

Unless done carefully, the spotting of probes on an array can introduce significant sources of error. The net number of probe molecular in each spot should ideally be equal, and the probes should each have equal physical access to interrogation by labeled targets in the hybridization mix.

Similarly, dyes should ideally be incorporated equally in each target. Since the labeling methods incorporate a mix of dye-dUTP and dTTP in the polymerization process, there is inherently a stochastic component as to how many dye-UTP molecules are incorporated at a particular site. More significantly, since different targets have different sequences by definition, they also will each contain a different number of possible dye-UTP incorporation sites. Thus, different targets can each be expected to have different fluorescent intensities that is sequence-dependent, not copy-dependent. In two-color arrays, probes containing Cy3 and Cy5 dyes compete for occupancy in the same probes—the size differential caused by different sizes of Cy3 and Cy5 molecules introduces an inherent hybridization bias.

Fluorescent dyes used in microarrays also are vulnerable to bleaching by room light as well as oxidation by atmospheric ozone, even at low levels. Any difference in handling or the amount of time arrays are exposed to light or room air can introduce array-to-array differences in fluorescence signal. When arrays are laser-scanned, slight differences in laser intensity or

photomultiplier tube sensitivity over time can also introduce signal biases over time—this highlights the importance of scanner manufacturers usually recommending scanners be allowed sufficient warmup and stabilization time prior to use.

Each of these technical issues with microarrays can introduce systematic and random errors into microarray data. Unfortunately, these errors are multiplicative, not merely additive. Many of these technical errors can be minimized through thoughtful design and precision manufacture of microarrays. Similarly, the effects of these sources of error can also be minimized during execution of microarray experimental protocols [287].

7 DETAILED EXPERIMENTAL METHODS

7.1 METHODS OVERVIEW

The purpose of the study presented here was to globally identify the gene targets of the TGFβ/SMAD3 transcriptional regulatory pathway, and more specifically, how this gene regulatory pathway might be involved in the pathophysiology of pulmonary fibrosis. Therefore, this study used the combination of two discovery-based approaches rather than a purely (or single) hypothesis-driven approach. The strategy was to combine two high-throughput discovery methodologies, ChIP-on-chip and gene expression microarrays, to give a clearer, more comprehensive picture of global transcriptional regulation of TGFβ/SMAD3 gene targets in A549 alveolar epithelial cells. Since the information gained from high-throughput measurement techniques is ordinarily too vast and complex for proper interpretation solely by the human eye, the second part of the strategy was to integrate and analyze the data using systems biology computational tools. Thus, the main methods presented here describe the methods of chromatin immunoprecipitation, promoter microarrays, and gene expression microarrays.

We ordinarily wish to spot-check the results of high-throughput measurements, as well as pursue new hypotheses generated from interpretation of the data using traditional molecular biology techniques. To verify ChIP pulldown, gene-specific PCR on single promoter targets is commonly used. The concept behind this is that PCR primers are designed for the promoter

regions of specific genes expected or known to be targets of the transcription factor under study (in this case, SMAD3). Ideally, the PCR primers will straddle a known or predicted transcription factor binding sequence. The presence of a PCR amplification product in comparison to a mock/no-antibody IP control on an agarose gel indicates specific pulldown of a portion of the expected promoter sequence. Thus, gene-specific PCR serves as a supporting validation of the quality of a ChIP experiment.

Similarly, information derived from a ChIP-on-chip experiment provides indirect evidence that the transcription factor under study binds to the promoter region of identified target genes. To provide an independent confirmation that the transcription factor protein under study indeed binds specifically to the identified target DNA sequence, an electromobility shift assay (EMSA, or sometimes called "gel shift") is used. The concept behind EMSA is that a short labeled DNA sequence (20-100 bp) containing the putative transcription factor binding site is added to the transcription factor protein (either purified recombinant protein or nuclear lysate which contains the native protein, or both), run on a resolving electrophoretic gel, and the protein-DNA interaction is indicated by retardation of the bound complex in the gel as compared to a lane containing the DNA alone. Specificity of protein-DNA binding is further confirmed by adding incremental amounts of non-labeled target DNA of the identical sequence in a series of other gel lanes. Specific protein-DNA interaction is indicated by a "competing out" of labeled DNA binding as evidence by a reduction in band intensity proportional to the amount of non-labeled competitor added. Further verification of specificity of protein-DNA binding is gained by (in yet another gel lane) also adding an antibody specific to the transcription factor protein. The motility of the resultant heavy complex of target DNA/transcription factor protein/antibody is

significantly reduced in the electrophoretic gel, producing a very highly retarded band. This is known as a “supershift” assay.

Finally, to confirm the results of gene expression microarrays, individually selected gene expression activity is measured using standard real-time quantitative PCR. Thus, the methods described here consist of the techniques associated with chromatin immunoprecipitation, promoter microarrays, gene expression microarrays, gene-specific PCR, EMSA, and quantitative real-time PCR. Additionally, the methods involved in data analysis and visualization are also described.

7.2 CELL CULTURES

Cryopreserved A549 lung alveolar epithelial carcinoma cells (CCL-185) were obtained from the American Type Cell Culture collection (ATCC; Manassas, VA). Cells were grown in F12-K culture medium (ATCC; 30-2004) supplemented with 10% fetal bovine serium (ATCC; 30-2020) and subcultured at 80-90% confluency. Prior to all experiments, cells were serum-starved for 18-24 hours. Cryopreserved primary Small Airway Epithelial Cells (Clonetics SAEC; CC-2547) were obtained from Lonza, Inc. (Walkersville, MD). Cells were grown in serum-free Small Airway Medium with supplied supplements (Clonetics BulletKit®, CC-3118).

7.3 CHROMATIN IMMUNOPRECIPITATION

The ChIP procedure was performed according to the protocol originally described by Weinmann et al., with the following modifications: 1 x 10^7 A549 cells were treated with TGFβ_1 (2 ng/ml) for up to 2 h. Cells were cross-linked with 1% formaldehyde for 12 min at room temperature (RT), after which glycine (125 mM) was added to quench the formaldehyde. The cells were washed twice with ice-cold PBS and lysed in 500 µl cell lysis buffer [50 mM Tris-HCl, pH 8.0; 1% Triton X-100; 10 mM KCl; supplemented with Complete protease inhibitor cocktail (Roche Diagnostics, Basel, Switzerland)]. Nuclei were pelleted at 5000 rpm for 5 min at 4°C, and resuspended in 400 µl of nuclear lysis buffer (50 mM Tris- HCl, pH 8.0; 10 mM EDTA; 0.1% SDS; supplemented with Complete). The samples were sonicated 3 x 10 s to yield sheared DNA fragments between 200 and 700 bp, and lysates were clarified by centrifugation (13,000 rpm, 10 min, 4°C). Samples were then incubated with 25 µg of anti-Smad3 antibody or control IgG (both from Upstate/Millipore) for 1 h at 4°C. To reduce nonspecific association, 30 µg of sonicated salmon sperm competitor DNA and 50 µg of BSA (both from Promega, Madison, WI, USA) were added to each sample. Immunoprecipitation was carried out using 50 µl of 50% (v/v) Protein A/G PLUS-Agarose beads (Santa Cruz Biotechnology, Santa Cruz, CA) at 4°C overnight. The immune complexes were washed as follows: three times with low-salt wash buffer (10 mM Tris-HCl, pH 8.0; 0.1% SDS; 0.1% sodium deoxycholate; 1% Triton X-100; 1 mM EDTA; 140 mM NaCl), 3 times with high-salt buffer (same as low-salt wash buffer, but with 500 mM NaCl), 2 times with LiCl wash buffer (10 mM Tris-HCl, pH 8.0; 250 mM LiCl; 1% Nonidet P-40; 1% sodium deoxycholate; 1 mM EDTA), and 2 times with TE buffer (20 mM Tris-HCl, pH 8.0; 1 mM EDTA). Elution was performed twice at 65°C for 15 min, first

with 200 μl of 1.5% SDS solution, and then with 250 μl of 0.5% SDS solution. Immunoprecipitated DNA-protein complexes were then reverse cross-linked at 65°C overnight and purified by phenol-chloroform extraction and ethanol precipitation with 30 μg glycogen (Roche Diagnostics). The resultant purified DNA was dissolved in 20 μl of water.

7.4 PROMOTER MICROARRAYS

Purified nucleic acid for array hybridization was blunt-ended using T4 DNA polymerase and ligated to linkers (sense strand: oJW102: 5'-GCGGTGACCCGGGAGATCTGAATTC ; anti-sense strand: oJW103: 5'-GAATTCAGATC) using T4 DNA ligase. Ligation products were amplified using a two-stage (15 cycle followed by dilution and input to a 25 cycle reaction) *Taq* polymerase-based PCR and purified using Qiagen PCR reaction purification columns. Amplified DNA was labeled through a random-primed, Klenow-based extension protocol derived from Invitrogen's BioPrime® Array CGH Genomic Labeling kit. Immunoprecipitated nucleic acid was labeled with Cy5 and mock IP (anti-flag) was labeled with Cy3 (Cy-5 and Cy-3 dyes; PerkinElmer Life and Analytical Sciences, Boston, MA). Dye incorporation was verified by Nanodrop spectrophotomer measurement (Nanodrop Technologies, Wilmington, DE). Labeled amplified DNA (Cy5 and Cy3) was combined and hybridized to Agilent 44K two-array whole genome promoter sets (G4112F; Agilent Technologies, Inc., Santa Clara, CA) for 40 hours at 65°C. Arrays were then washed in a series of sodium chloride-sodium citrate (SSC) buffers and acetonitrile, and treated with Agilent stabilization and drying solution for 30 seconds. Arrays were then immediately scanned on a GenePix 4000B scanner in two-color array mode (Cy5/Cy3) yielding an intensity ratio of Cy5 (IP) to Cy3 (mock IP) for each probe.

7.5 CHIP-ON-CHIP PROMOTER MICROARRAY ANALYSIS

Agilent 44K whole genome promoter arrays contain probes that cover 2000 base pair upstream to 800 base pair downstream of the transcriptional start site for 44,000 published RefSeq genes. The probed areas contain on average four to six separate 60-mer sequences spaced at approximately 300 base pair intervals.

For Agilent promoter microarray analysis, we used a model-based algorithm developed by Kaplan and Friedman [288]. The algorithm uses the length distribution of input sonicated target DNA fragments to predict the shape of the frequency distribution of overlapping targets over the series of 60-mer probes in each promoter sequence on the array. The method then uses this predicted shape to discriminate actual binding events from random binding events (i.e., noise), as well as estimate the location of the true binding event.

The estimated shape of the binding curve is modeled by the following equation,

Equation 7-1
$$F(\Delta_x) \propto \int_{l=\Delta_x}^{\infty} (l - \Delta_x) c(l) dl$$

where x is the location of the probe on the promoter sequence and l is the length of the target DNA fragment. The probability of a probe located Δx bases away from the binding location is proportional to the integral over all fragment lengths (of length larger than Δx) of the number of possible alignments of the target DNA fragment bound to the reporting probe, multiplied by the relative abundance of DNA fragments of such length, denoted by *c(l)*.

Once the shape of a binding event is estimated, relative enrichment is estimated by measuring relative peak height. For this, Kaplan and Friedman have developed an iterative algorithm to identify all significant binding events that appear in the ChIP-chip data. Briefly, this

is done by using a "sliding window" approach that identifies stretches of enriched probes and attempts to explain (at least part of) their values using the shape of a peak. For each peak, the algorithm enumerates and selects the most probable values for center position and peak height (enrichment), and then computes the statistical significance of this peak.

The statistical significance of a binding event is estimating by computing an empirical log-likelihood ratio (LLR) p-value. They compute the likelihood of the set of probes *S* given the null model L_0 that assumes the values are normally distributed around the median enrichment ratio of the array. They then compute the likelihood of the same probes given a peak model with center *x* and height *alpha*, denoted by L_{peak}. They use the log-likelihood-ratio (LLR) L_{peak}/L_0 to score the significance value of the peak by computing 1000 shuffling-based LLR scores as follows: they replaced the measured enrichment values for each probe in *S* with a randomly chosen probe from the array, find the optimal peak height as described above, and then calculate the log likelihood ratio for this set. Finally, they calculated the empirical LLR-based p-value of the original peak by computing the percentile of the rank of the true LLR score among the 1000 shuffling-based scores. If the *p*-value was significant (i.e. falls below 0.01), and the peak's height exceeded 1.5, they call it a "binding event." This model-based approach also allows us to integrate data from different replicates, calculating the likelihood of the peak based on all intensity values of its probes.

Kaplan and Friedman analyzed the SMAD3 ChIP-on-chip data for peaks with and without $TGF\beta_1$ stimulation. Each peak was assigned an enrichment value and a *p*-value (the statistical significance of seeing such a peak at random). To differentiate the true target genes of SMAD3 with and without $TGF\beta_1$ stimulation, we analyzed the ChIP-chip data, and identified

genes whose promoter was bound by SMAD3 in at least two of the three array replicates. For this, we used a p-value threshold of 0.01 in each of the two replicates.

7.6 GENE-SPECIFIC PCR VERIFICATION

A portion of LM-PCR amplified immunoprecipitation product was used as input for a separate gene-specific PCR reactions (25 cycles) to verify enrichment of promoter regions of the known TGFβ_1-responsive genes PAI-1 and SMAD7 as well as the FOXA2 promoter sequence. PCR was performed using *Taq* polymerase (Invitrogen Corporation, Carlsbad, CA) in 15 µl reactions according to the manufacturer's protocol.

7.7 GENE EXPRESSION MICROARRAYS

For gene expression measurements we used Agilent 4 x 44K whole human genome microarray kits (G4112F; Agilent Technologies, Inc., Santa Clara, CA) according to the manufacturer's protocol. Briefly, 500 ng of total RNA was amplified using an Agilent Low Input Linear Amplification and Labeling kit and resultant cRNA was labeled with cyanine-3 (cy-3, 10 mM; PerkinElmer Life and Analytical Sciences, Boston, MA). Cy-3 labeled probes were purified using Qiagen RNeasy Mini kit (Qiagen, Hilden, Germany) per the manufacturer's protocol. Sufficient yield and dye incorporation were confirmed using a Nanodrop spectrophotometer (Nanodrop Technologies, Wilmington, DE). Arrays were hybridized for 17 hours at 60°C under continuous rotation at ~ 20 RPM. The gasket slide coverslips were removed and the slides

washed for one minute in Agilent Wash Buffer 1 (6x sodium chloride/sodium phosphate/EDTA (SSPE) + 0.005% N-laurylsarcosine). The slides were then washed in Agilent Wash Buffer 2 (0.06x SSPE + 0.005% N-laurylsarcosine) for 1 minute followed by 1 minute in acetonitrile and then 30 seconds Agilent Stabilization and Drying solution. Arrays were scanned using the Agilent DNA microarray scanner.

7.8 EXPRESSION MICROARRAY DATA ANALYSIS AND STATISTICS

DNA microarray feature intensities were measured using Agilent Feature Extraction software version 9.5.2. There were three replicates each of four time points (0, 2 hour, 12 hour, 24 hour) of TGFβ_1 stimulation, each for vehicle-only control (DMSO) and for SIS3 treatment. This yielded 24 microarrays total. The three replicates of the 2 hour time point in the SIS3 treatment came out completely blank due to some unknown technical or manufacturing error and were therefore excluded from normalization or further analysis.

Background-subtracted signal intensities of the remaining arrays were log-base 2 transformed and then normalized across arrays by cyclic lowess in the R statistical package (see Appendix B). Since array data often contains multiple (and variable numbers of) probes per gene, the probe intensities were averaged and combined into individual gene intensity values. Individual gene intensities across arrays (*i.e.,* row) were geometric mean normalized to the first time point (0 hr control).

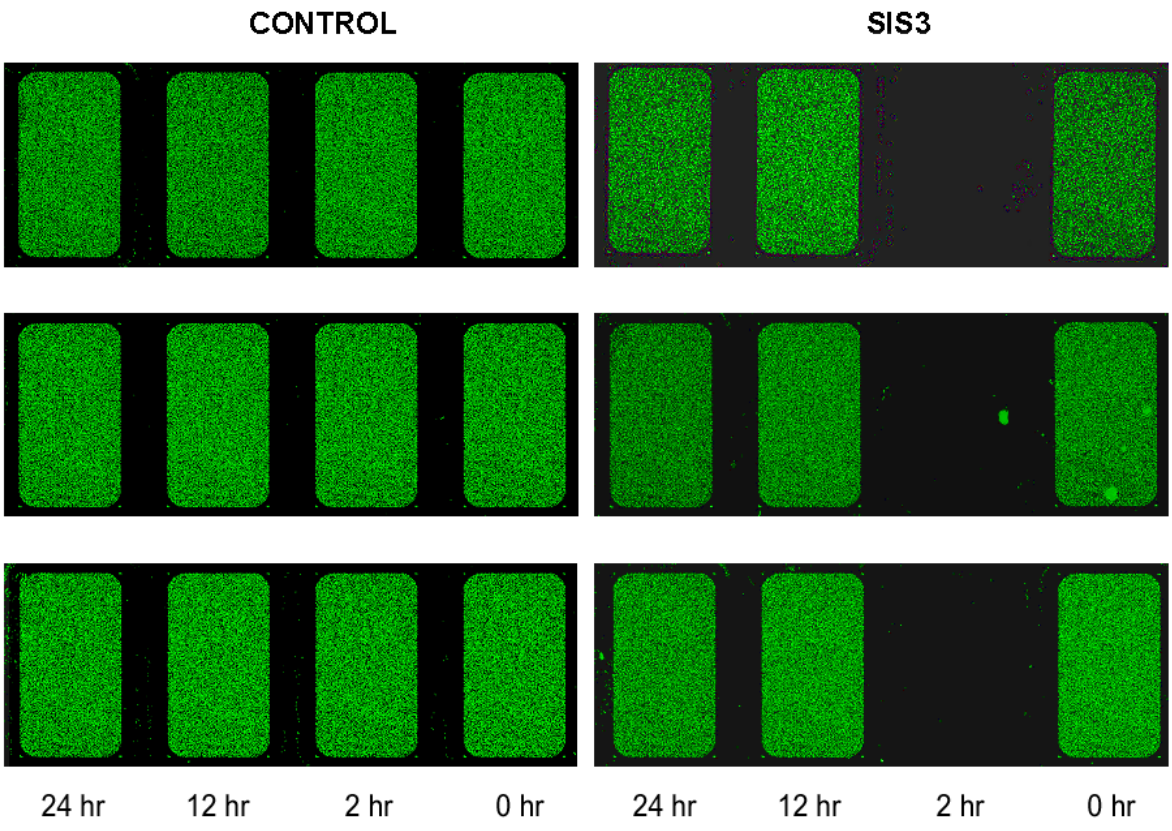


Figure 7-1. Gene expression microarray images.

To assess data quality, a boxplot of signal intensities was created. From visual inspection arrays Control, 12 hour, replicate 1 (C12.1) and control, SIS3, 12 hour, replicate 1 (S12.1) showed significant variability due to unknown technical problems and were excluded from further analysis. The raw data from the remaining good microarrays were again normalized as described and this data were used for all subsequent analysis.

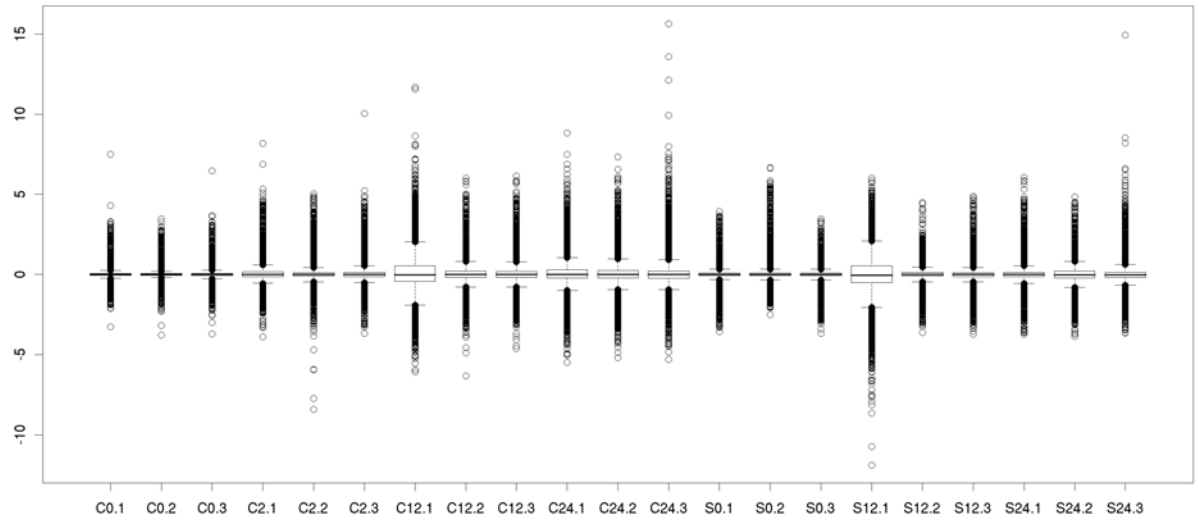


Figure 7-2. Boxplot of 21 microarrays after lowess normalization.

The data were analyzed using three separate software packages: first, by permutation test between separate time points in the R statistical programming environment (www.r-project.org/); next by the Significance Analysis of Microarrays package from Stanford (www-stat.stanford.edu/~tibs/SAM/) [289]; and finally in the Short Time-series Expression Miner (STEM) program developed by Jason Ernst and Ziv Bar-Joseph at Carnegie Mellon University (www.cs.cmu.edu/~jernst/stem/) [290, 291]. Since the features of STEM are ideally suited to the time series nature of the gene expression microarray data, and also minimizes the effect of the missing time point of the SIS3 microarray data. Therefore it was the main program used for further analysis.

STEM is a Java program developed for clustering, comparing, and visualizing time-series microarray data having eight or fewer individual time points. STEM first compares expression behavior of genes against all possible behaviors. It then clusters the gene expression behaviors it sees into groups of similarly behaving genes. Next, it performs hypothesis testing on genes in

each cluster to identify significantly regulated genes. The null hypothesis is that the probability of observing any value at a point in time is independent of its future and past values, i.e., gene expression values fluctuate randomly. The program uses a permutation test to quantify the expected number of genes that would have been assigned to each model if the data were random. Thus, a gene expression profile deemed as significant would generate an established pattern similar to other genes in its group and distinctly different from random deviation. The resultant p-values are then Bonferroni corrected [290, 291].

7.9 ELECTROMOBILITY SHIFT ASSAY

Cultured A549 lung alveolar epithelial carcinoma cells (ATCC; CCL-185) at 60-70% confluence were treated with 2 ng/ml recombinant human $TGF\beta_1$ (catalog number 240 B/CF; R&D Systems, Minneapolis, MN) for 60 minutes. Nuclear proteins were isolated using a micropreparation technique published previously [292]. Briefly, A549 cells at ~80% confluence were scraped into 1.5 ml of cold phosphate-buffered saline and pelleted for 10 seconds. The cell pellet was resuspended in 400 μl cold Buffer A (10 mM HEPES-KOH pH 7.9 at 4°C, 1.5 mm MgCl2, 10 mM KCl, 0.5 mM dithiothreitol, 0.2 mM PMSF). The cells were then allowed to swell on ice for 10 minutes, vortexed for 10 seconds, centrifuged for 10 seconds, and the supernatant fraction discarded. The pellet was resuspended in 100 μl of cold Buffer C (20 mM HEPES-KOH pH 7.9, 25% glycerol, 420 mM NaCl, 1.5 mM MgCl2, 0.2 mM EDTA, 0.5 mM dithiothreitol, 0.2 mM PMSF) and incubated on ice for 20 minutes. Cellular debris was removed

by centrifugation for 2 minutes at 4°C and the supernatant containing nuclear proteins was flash-frozen in liquid nitrogen and stored at -80°C.

Nuclear extracts at 1:10 dilution and recombinant full length SMAD3 protein (catalog number sc-4709; Santa Cruz Biotechnology, Santa Cruz, CA) were incubated with 5'-end Cyanine-5 labeled probe and/or non-labeled competitor oligonucleotide for 20 minutes at room temperature in a binding buffer consisting of 20% glycerol, 5 mM $MgCl_2$, 2.5 mM EDTA, 25 mM DTT, 200 mM NaCl, 50 mM Tris HCl pH 7.6, and 0.25 mg/mL poly(dI-dC). The oligonucleotides (5'- Cy5-GATTGCTGGTCGTTTGTTGTGGCT - 3', 5' - AGCCACAACAAACGACCAGCAATC - 3') were synthesized by Integrated DNA Technologies (Coralville, IA), and consisted of a sequence obtained from the HNF3β/FOXA2 promoter sequence (chr20:22,512,934-22,512,958). Supershift assay was performed by additionally incubating nuclear extract with 0.4 μg rabbit polyclonal antibody to SMAD3 (Catalog Number ab28379; Abcam, Cambridge, MA) prior to incubating with oligonucleotide. The protein/DNA complexes were run on a 6% native polyacrylamide gel and visualized on a Typhoon 9400 imaging and documentation system using Cyanine-5 dye excitation and fluorescence settings.

7.10 SIS3 INHIBITION OF SMAD3 ACTIVITY

Specific Inhibitor of SMAD3 (SIS3, Catalog Number 566405; EMD Chemicals, Inc., San Diego, CA) is a potent, specific inhibitor of $TGF\beta_1$/ALK-5 phosphorylation of SMAD3 while having no effect on SMAD2, p38 MAPK, ERK, or phosphoinositide 3-kinase (PI3K) signaling [293]. Cultured A549 cells at 30-50% confluence were treated with 10 uM SIS3 in dimethyl

sulfoxide (DMSO), or DMSO (vehicle-only) 30 minutes prior to $TGF\beta_1$ treatment. Cells were treated with 2 ng/mL recombinant $TGF\beta_1$ (Catalog Number 240 B/CF; R&D Systems, Minneapolis, MN) for 0, 2, 12, and 24 hours. Total mRNA was extracted using Trizol (Invitrogen Corporation, Carlsbad, CA) according to the supplier's protocol for adherent cultured cells.

7.11 QUANTITATIVE REAL-TIME PCR

A549 cells (Catalog Number CCL-185; ATCC) and Small Airway Epithelial Cells (Catalog Number CC-2547, Lonza, Inc.) were grown to 80-90% confluence. Cells were treated with 2 ng/mL recombinant $TGF\beta_1$ (Catalog Number 240 B/CF; R&D Systems, Minneapolis, MN) for 0 (control), 2, 12, and 24 hours. Total mRNA was extracted using Trizol (Invitrogen Corporation, Carlsbad, CA) according to the supplier's protocol for adherent cultured cells. Singly-purified total mRNA was normalized to 600 ng and reverse-transcribed using random hexamer priming with a SuperScript kit (Invitrogen Corporation, Carlsbad, CA). Quantitative PCR was performed using FAM/MGB TaqMan Gene Expression Assays (Applied Biosystems) on an ABI Prism 7900HT thermocycler/measurement instrument. Cycling parameters were *Taq* activation 95°C for 12 minutes, then 40 cycles cycling between denaturing at 95°C for 15 seconds and annealing at 60°C for 1 minute. The ABI Taqman primers used are listed in Table 1 To evaluate relative mRNA expression of FOXA2 and PAI-1 (Serpine1), we used GAPDH as a reference gene. Relative changes in transcript levels of FOXA2 and PAI-1 (serpine1) as compared to controls are expressed as ΔΔCt values ($\Delta\Delta Ct = \Delta Ct_{treated} - \Delta Ct_{control}$) using ABI Sequence Detection Software v2.2.2.

7.12 INGENUITY PATHWAYS FUNCTIONAL ANALYSIS

7.12.1 NETWORK GENERATION

A data set of signficantly bound (ChIP) or up/down-regulated (expression) genes containing gene identifiers and corresponding binding/expression values was uploaded into in the application. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. These genes, called focus genes, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus genes were then algorithmically generated by Ingenuity Pathways Analysis based on their connectivity.

Genes or gene products are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least one reference from the literature, from a textbook, or from canonical information stored in the Ingenuity Pathways Knowledge Base. The intensity of the node color indicates the degree of up- (red) or down- (green) regulation. Nodes are displayed using various shapes that represent the functional class of the gene product.

7.12.2 FUNCTIONAL ANALYSIS OF A NETWORK

The Functional Analysis of a network identified the biological functions that were most significant to the genes in the network. The network genes associated with biological functions and/or diseases in the Ingenuity Pathways Knowledge Base were considered for the analysis.

Fischer's exact test was used to calculate a p-value determining the probability that each biological function assigned to that network is due to chance alone.

7.12.3 CANONICAL PATHWAY ANALYSIS

Canonical pathway analysis identified the pathways from the Ingenuity Pathways Analysis library of canonical pathways that were most significant to the data set. A data set of signficantly bound (ChIP) or up/down-regulated (expression) genes containing gene identifiers and corresponding binding/expression values was uploaded into in the application and associated with a canonical pathway in the Ingenuity Pathways Knowledge Base. The significance of the association between the data set and the canonical pathway was measured in two ways: 1) A ratio of the number of genes from the data set that map to the pathway divided by the total number of genes that map to the canonical pathway is displayed. 2) Fischer's exact test was used to calculate a p-value determining the probability that the association between the genes in the dataset and the canonical pathway is explained by chance alone.

8 EXPERIMENTAL RESULTS

This chapter will present detailed experimental results from ChIP-on-chip and gene expression microarrays, followed by a systems-level integrated analysis of this data. From this systems-level analysis arise three novel findings that may provides clues to transcriptional regulatory mechanisms involved in the pathogenesis of IPF. The following descriptions of ChIP-on-chip and gene expression results, respectively and in combination, serve to confirm activation of the TGFβ_1/SMAD3 signal transduction and gene transcriptional regulatory pathways in the experimental system. The first discussion is of the three specific findings, transgelin, FOXA2, and PINX1.

8.1 SPECIFIC FINDINGS FROM SYSTEMS-LEVEL ANALYSIS OF THE TGFβ_1/SMAD3 PATHWAY

Three previously unknown SMAD3 target genes were identified from the ChIP-on-chip experiments. Each is discussed in this section, along with their respective supporting experimental results.

8.1.1 TRANSGELIN (TAGLN)

From analysis of ChIP and gene expression data were derived three findings relative to the pathogenesis of IPF. First, transgelin (TAGLN), a marker of EMT and cell mobility, was found and confirmed to be strongly up-regulated in alveolar epithelial cells in response to direct activation from the TGFβ_1/SMAD3 pathway. In the follow up experiments, TGFβ_1 was confirmed to induce upregulation of both transgelin mRNA and protein in alveolar epithelial cells. In bleomycin-treated mice and IPF patients alike, transgelin was also shown to be increased in type II alveolar epithelial cells by qRT-PCR and immunohistochemistry. SiRNA inhibition of transgelin suppressed TGFβ_1-induced migration of A549 and primary type II alveolar epithelial cells. This gene identification and detailed follow up work was performed by the lab of our collaborator, Oliver Eickelberg, MD of Geissen University and published separately in a recent article in FASEB (Yu, H., Konigshoff, M., Jayachandran, A., Handley, D., Seeger, W., Kaminski, N., Eickelberg, O.) [35].

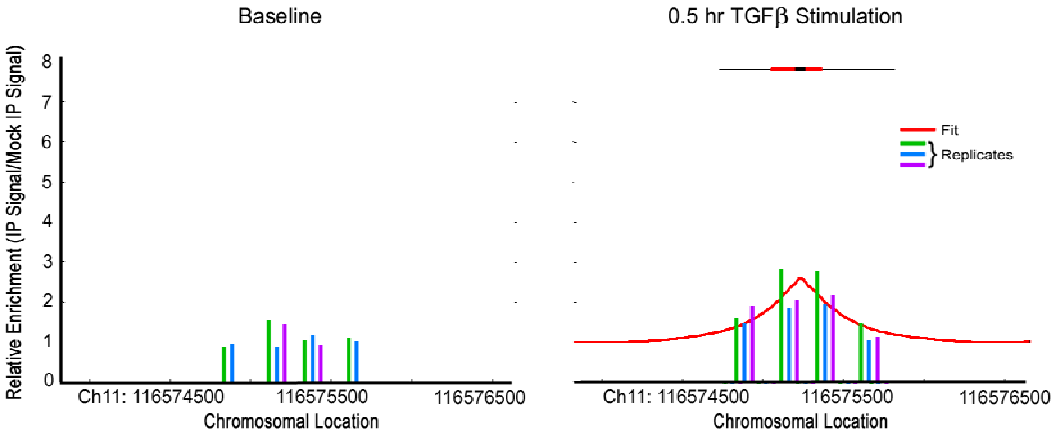


Figure 8-1. ChIP promoter binding profile of transgelin, baseline (left) and after 30 minutes 2ng/ml TGFβ_1 stimulation (right). Each bar height indicates respective array

signal intensity for that probe. Values from the three promoter array replicates are shown (green, blue, purple, respectively). If the binding was statistically significant, the binding curve (red) is also included and shows the fitted peak shape. The region shown maps to Chromosome 11q23.2

8.1.2 FORKHEAD BOX A2 (FOXA2)

We identified a novel transcription factor, Forkhead Box A2 (FOXA2, also known as HNF3β), as one of the key genes that is transcriptionally regulated directly through the c. FOXA2 is a highly conserved member of the winged helix nuclear factor gene family and plays a major role in embryonic development as well as proper function of the mature lung [38], [294]. Its role in alveolar epithelial cell function, in particular surfactant production, established FOXA2 as key gene that may be involved in the pathogenesis of IPF.

The following experimental results help establish the $TGF\beta_1$/SMAD3 as a transcriptional regulator of FOXA2. These include (1) the binding graph of the FOXA2 promoter with and without $TGF\beta_1$ stimulation, (2) electromobility shift assay (EMSA) demonstrating direct and specific binding of SMAD3 to the promoter of FOXA2, (3) quantitative real-time PCR of FOXA2 after $TGF\beta_1$ stimulation with and without specific SIS3 inhibition of the $TGF\beta_1$/SMAD3 pathway, and (4) gene expression data showing down-regulation of FOXA2 by the $TGF\beta_1$/SMAD3 pathway that is largely abrogated by SIS3 treatment. Further, the gene expression data shows a $TGF\beta_1$/SMAD3 pathway specific effect on surfactant proteins A,B,C, and D. Since surfactant proteins in alveolar epithelial cells are known to be mediated through the FOXA2 transcription factor, we believe this provide strong evidence for the direct role of $TGF\beta_1$/SMAD3/FOXA2/surfactants in the pathogenesis of IPF. This will be discussed in greater detail in the following chapter.

8.1.3 FOXA2 CHIP SMAD3 BINDING

ChIP-on-chip analysis identified the FOXA2 promoter as significantly bound after TGF TGFβ_1 stimulation, but not at baseline (Figure 8-35). The maximum peak height before TGFβ_1 stimulation is 1.41 and after TGFβ_1 stimulation is 2.62.

8.1.3.1 FOXA2

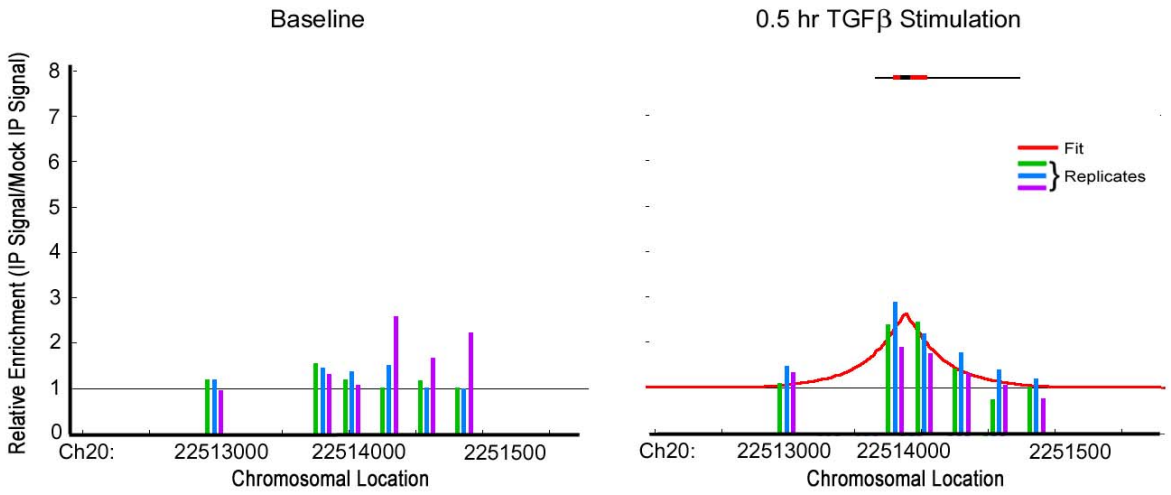


Figure 8-2. ChIP promoter binding profile of FOXA2, baseline (left) and after 30 minutes 2ng/ml TGFβ_1 stimulation (right). Each bar height indicates respective array signal intensity for that probe. Values from the three promoter array replicates are shown (green, blue, purple, respectively). If the binding was statistically significant, the binding curve (red) is also included and shows the fitted peak shape. The region shown maps to Chromsome 20p11.

8.1.3.2 EMSA

Electromobility shift assay (EMSA) using both recombinant SMAD3 protein and nuclear extract from A549 cells stimulated with TGFβ_1 shows specific binding of the protein to the promoter region of the FOXA2 gene (Figure 8-36 below).

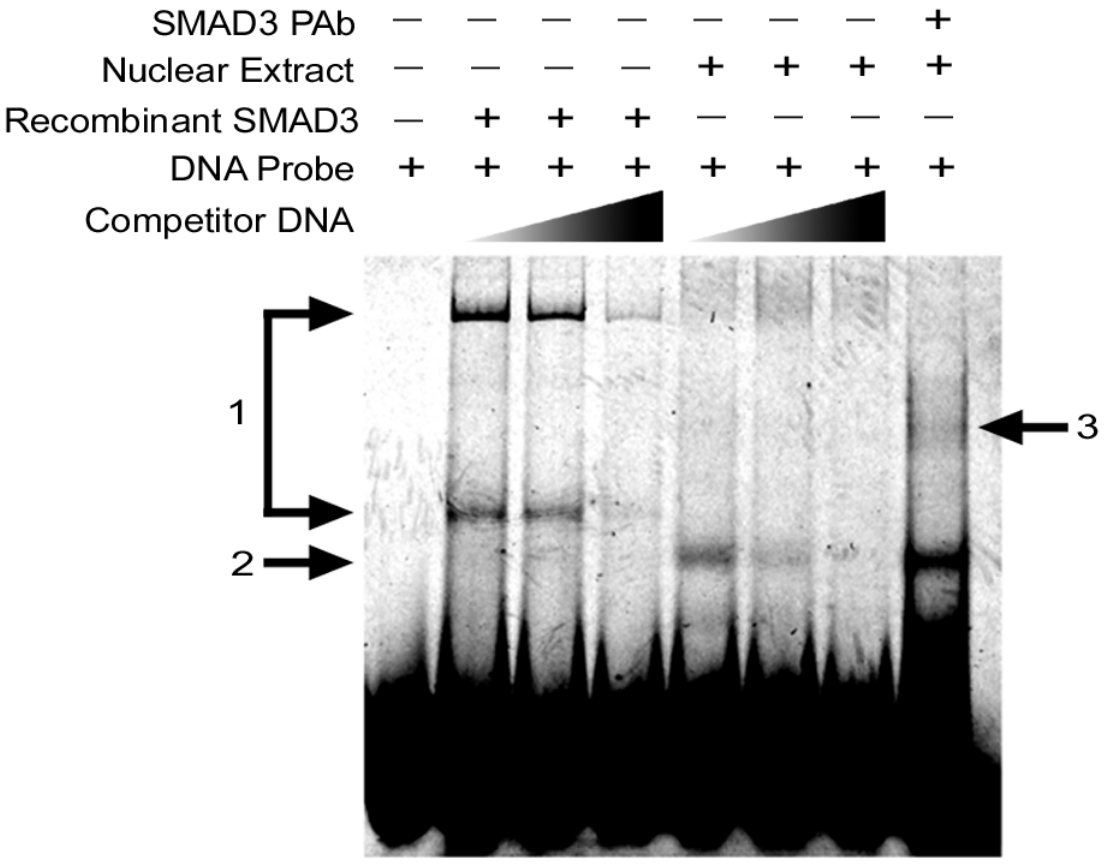


Figure 8-3. Electromobility shift assay shows specific binding of the SMAD3 protein (lanes 2-4) and nuclear extract from TGFβ_1-stimulated A549 cells (lanes 5-7). Lanes 3/6 and 4/7 contain non-labeled competitor FOXA2 promoter sequence DNA, 40 ng and 200 ng, respectively. Lane 8 contains a PAb against SMAD3 and has a supershift band (3).

8.1.3.3 FOXA2 QRT-PCR

Quantitative real-time PCR (qRT-PCR) was also performed to assess levels of FOXA2 mRNA after 2, 12, and 24 hours of stimulation with 2 ng/ml exogenous $TGF\beta_1$, respectively. This was performed in A549 cells receiving 2 μl of DMSO containing a potent and specific SMAD3 inhibitor, SIS3, or a vehicle-only control (DMSO). The results show that FOXA2 is significantly repressed (approximately 70% decrease in mRNA levels) at 2 hours. This effect is largely abrogated by SIS3 treatment, suggesting the effect is mediated specifically and directly through the $TGF\beta_1$/SMAD3 pathway.

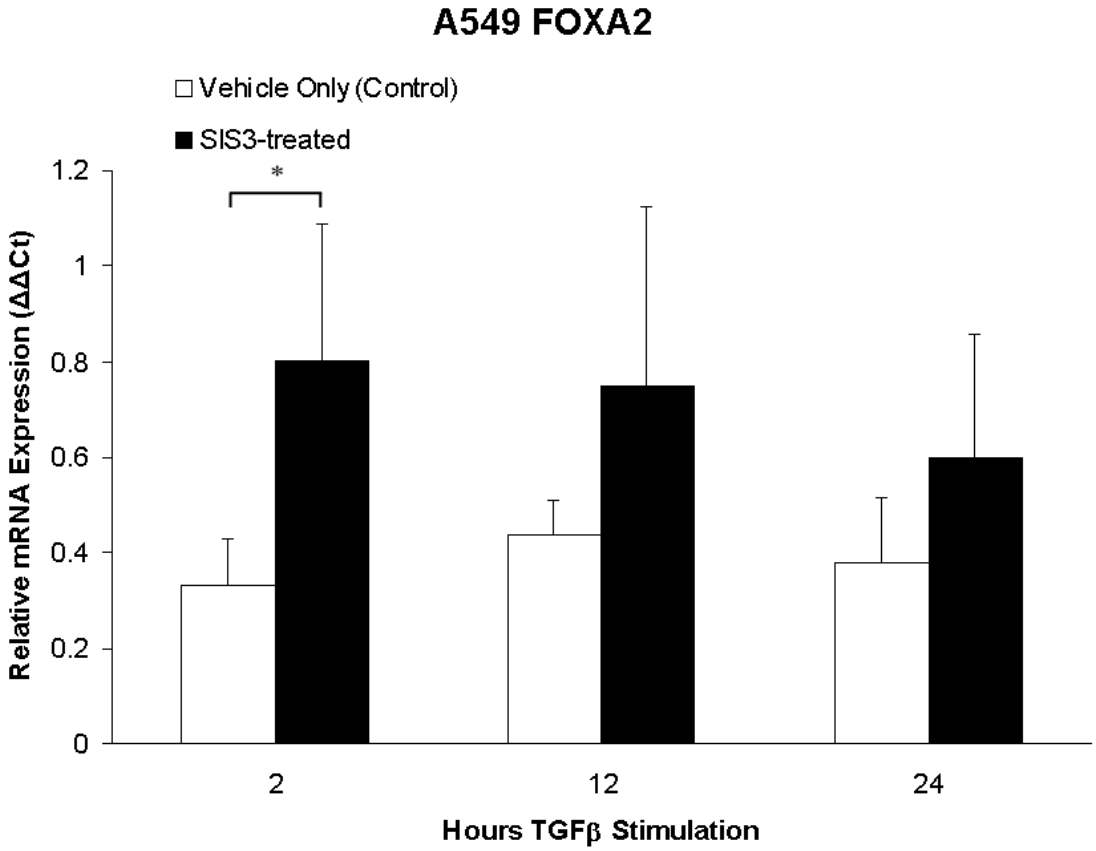


Figure 8-4. Quantitative real-time PCR of FOXA2 levels in A549 cells with and without SIS3 treatment.

To assess whether this effect was specific to the A549 cell line, FOXA2 mRNA levels were also measured in primary human small airway epithelial cells (SAEC). FOXA2 mRNA levels were measured in relation to the known TGFβ_1-responsive gene, Serpine1 as a verification of TGFβ_1/SMAD3 pathway induction. The results show that in relation to no TGF$\beta_1$treatment, FOXA2 levels are repressed by about 70-80% at 2, 12, and 24 hours. Conversely, Serpine1 levels increase steadily and monotonically by over 2-fold during the same time series (Figure 8-34). The qRT-PCR results in both A549 alveolar epithelial cells and primary SAECs suggests that TGFβ_1 represses mRNA expression of FOXA2 in pulmonary epithelial cells.

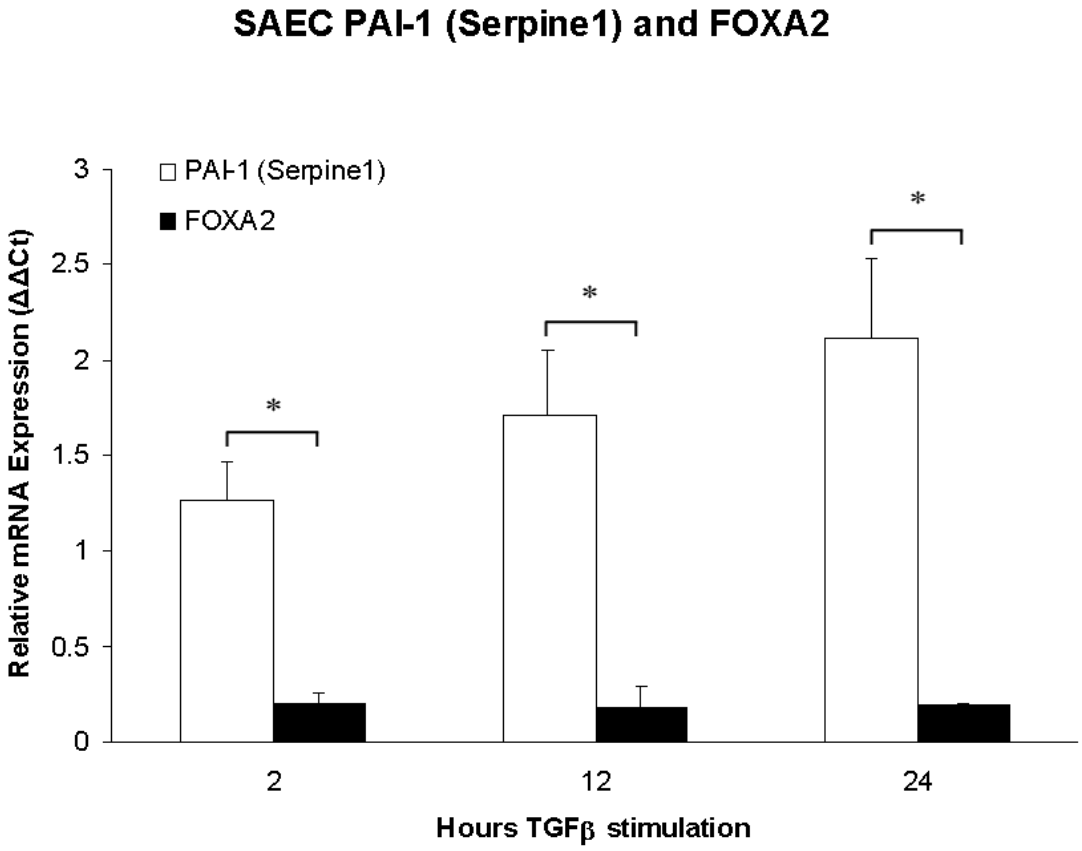


Figure 8-5. Quantitative real-time PCR of FOXA2 and Serpine1 levels in Small Airway Epithelial Cells (SAEC) at 2, 12, and 24 hours TGFβ_1 treatment in relation to control (no TGFβ_1).

8.1.3.4 CHIP AND GENE EXPRESSION MICROARRAY DATA ON FOXA2

The specific results from both ChIP-on-chip and gene expression microarray data are consistent with each other, as shown in Figure 8-35 below. On the left top portion of the figure we see evidence of significant, strong binding of SMAD3 to the promoter of FOXA2. Similarly, FOXA2 gene expression microarray levels are strongly decreased after TGF$\beta_1$treatment (top, right). This effect is largely abolished by SIS3 treatment, which is in complete agreement with the qRT-PCR results shown above. This suggests regulation of FOXA2 directly by SMAD3.

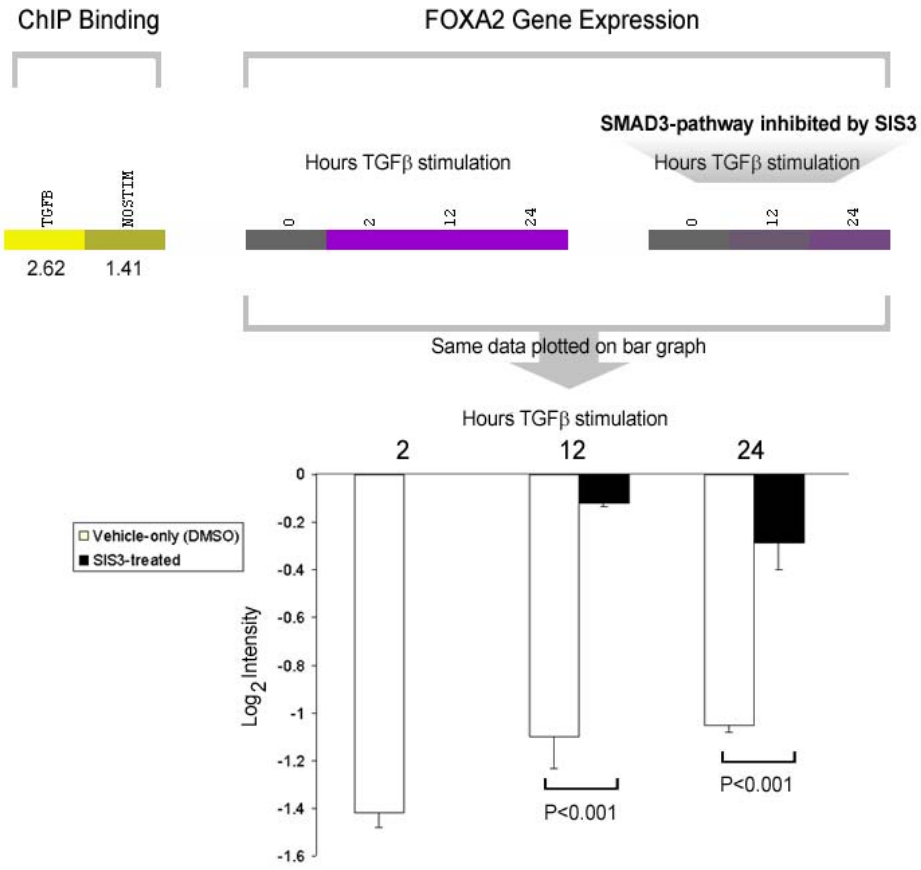


Figure 8-6. Combined ChIP binding values (left, top) with gene expression microarray values (right, top). The microarray expression values are plotted in a bar graph (bottom) and show significant repression (white bars) of FOXA2 during a time course of TGFβ_1 treatment that is largely abolished by SIS3 treatment (black bars).

8.1.3.5 SURFACTANT GENE EXPRESSION HEAT MAP

The effect of TGF$\beta_1$treatment on expression of various surfactant genes (A1, B, C, D) are seen in the gene expression microarray results. The following heat map summarizes the findings. The microarray contains probes for three transcript variants of Surfactant A1. The first one is strongly upregulated after TGFβ_1 treatment; SIS3 largely reverses this upregulation. The second variant is also strongly upregulated by TGFβ_1 treatment and is only slightly affected by SIS3. The third variant shows a variable time course of up- and down-regulation, whose pattern appears completely inverted with SIS3 treatment. TGFβ_1 treatment appears to strongly down-regulate Surfactant B (*SFTPB*); this effect appears reversed (*SFTPB* is up-regulated) with SIS3 treatment. Finally, Surfactant D is first down-regulated at 2 hours, then up-regulated at 24 hours. With SIS3 treatment, gene expression appears to be up-regulation at 12 and 24 hours. FOXA2 is a known transcriptional regulator of surfactants in alveolar Type II epithelial cells [].

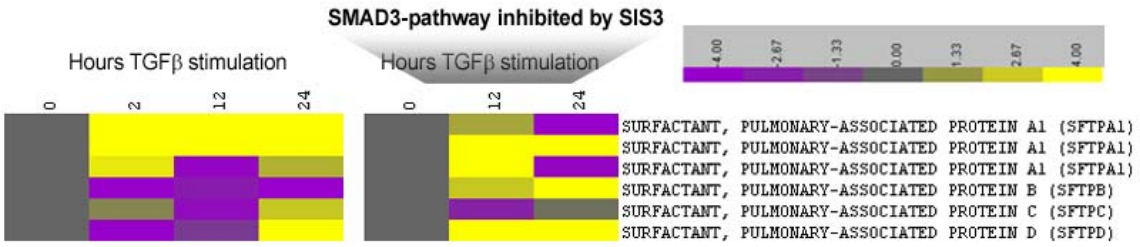


Figure 8-7. Heat map illustration of gene expression microarray results for Surfactant A1, B, C, and D, in a time course treatment of TGFβ_1 at 2, 12, and 24 hours (left; vehicle-only control) and 12 and 24 hours (right; SIS3 treatment).

8.1.4 PIN2-INTERACTING PROTEIN 1 (PINX1)

Finally, SMAD3 was identified as binding to the promoter of the gene for PIN2-interacting protein 1 (PINX1). PINX1, a potent inhibitor of telomerase reverse transcriptase (hTERT), was also found by ChIP/gene expression network analysis to be involved in the protein-interaction network of *hTERT* (Sections 8.2.4.2-8.2.4.5) [46, 47, 49].

8.1.4.1 PINX1 CHIP BINDING CURVE

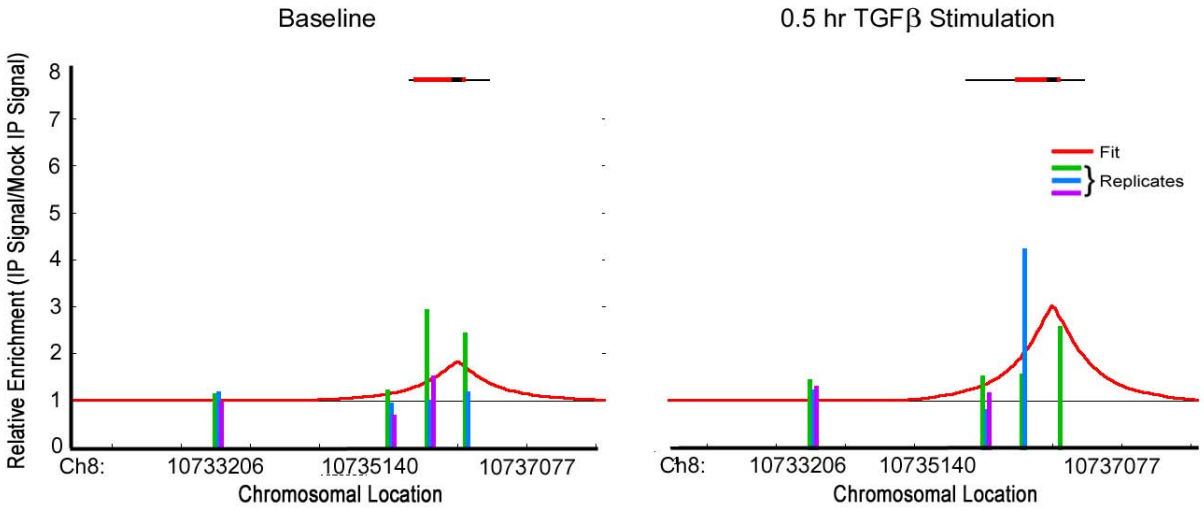


Figure 8-8. ChIP SMAD3 promoter binding curves showing baseline (no $TGF\beta_1$; left) and after $TGF\beta_1$ treatment (right). This suggests specific and strong binding of SMAD3 to the promoter of PINX1 both at baseline and after SMAD3 phosphorylation/nuclear translocation. Each bar height indicates respective array signal intensity for that probe. Values from the three promoter array replicates are shown (green, blue, purple, respectively). If the binding was statistically significant, the binding curve (red) is also included and shows the fitted peak shape. The region shown maps to Chromosome 8p23.

8.1.4.2 CHIP SMAD3-BOUND GENES IN TERT NETWORK

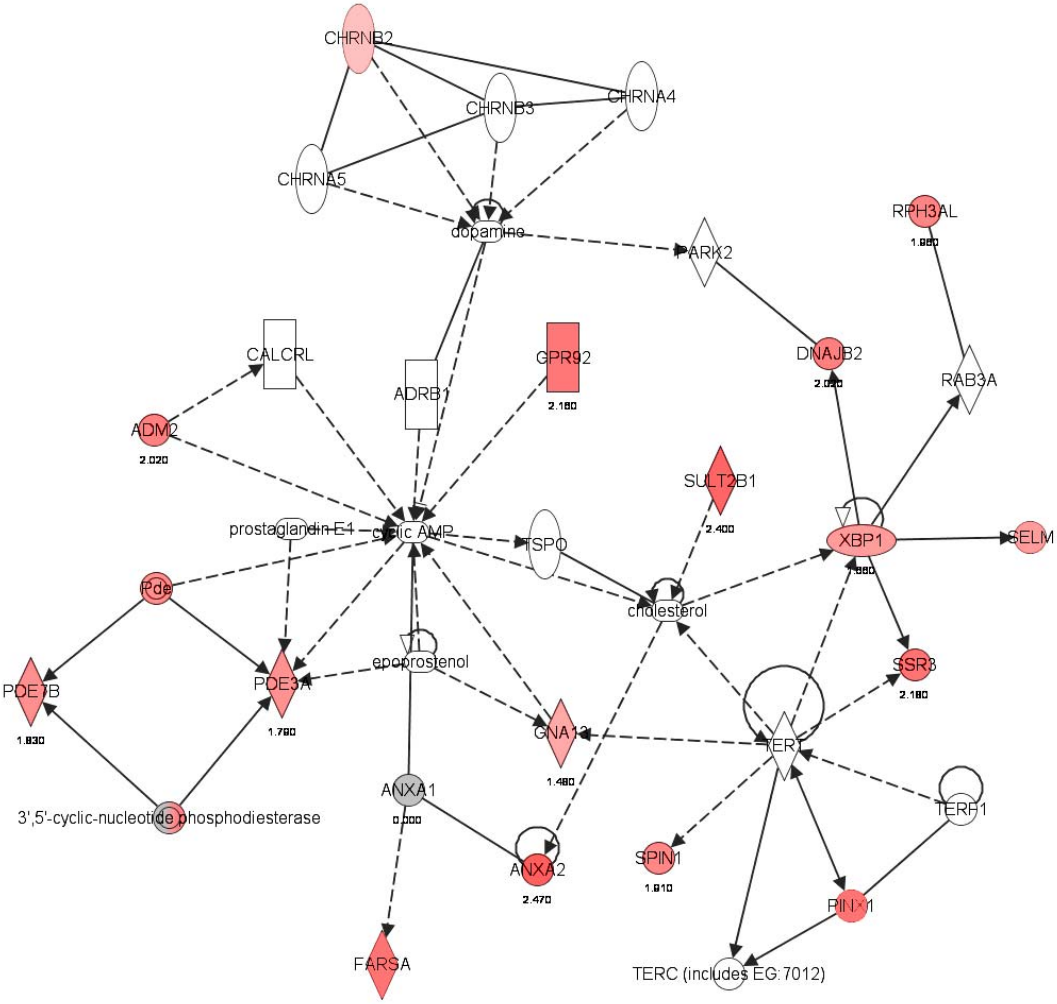


Figure 8-9. ChIP SMAD3-bound target genes illustrated in the TERT interaction network. Red denotes bound target gene; Color intensity depicts binding peak height.

8.1.4.3 GENE EXPRESSION IN TERT NETWORK AFTER TGFβ$_1$ TREATMENT

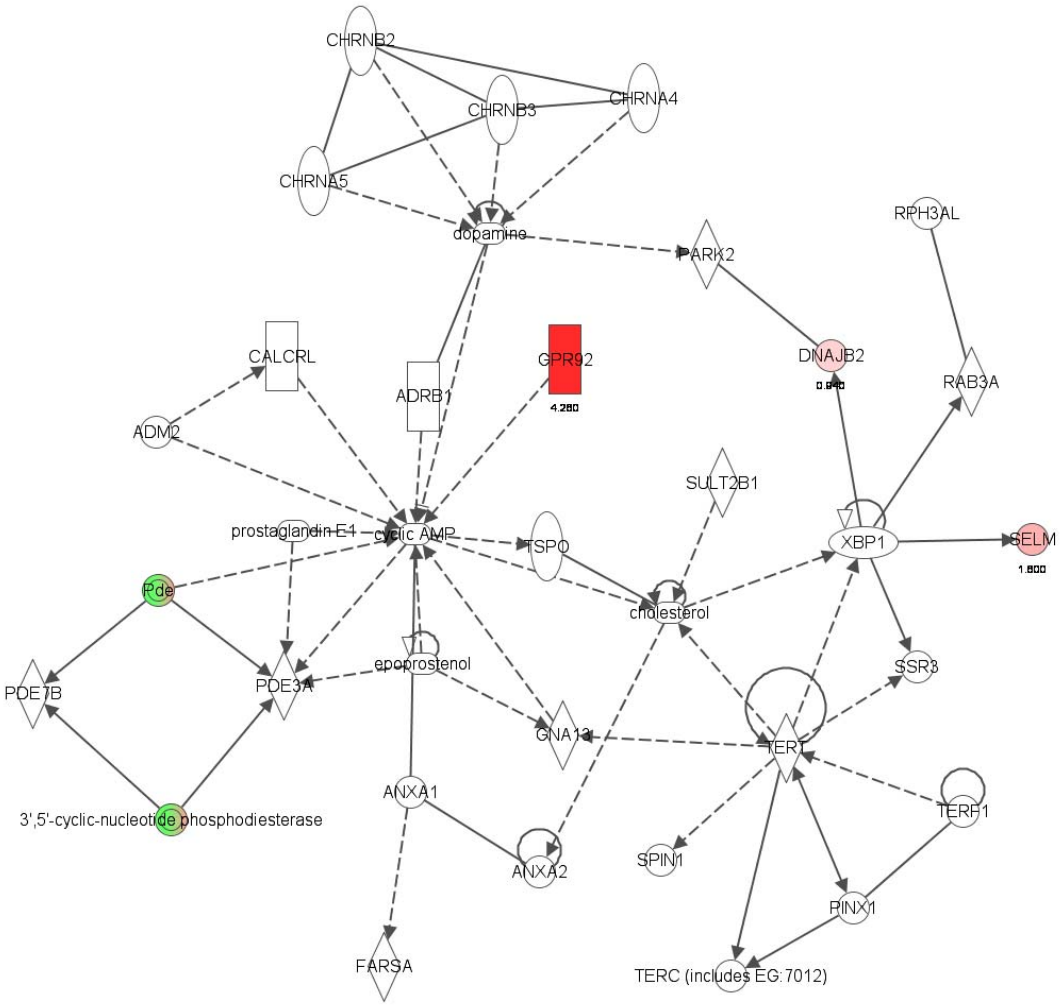


Figure 8-10. Gene expression values illustrated in the ChIP SMAD3-bound target genes illustrated in the TERT interaction network. Significant gene expression profiles after TGFβ$_1$ stimulation; red denotes up-regulation, green denotes down-regulation. Color intensity depicts relative gene expression levels.

8.1.4.4 COMBINED CHIP BINDING AND GENE EXPRESSION IN TERT NETWORK AFTER TGFβ$_1$ TREATMENT

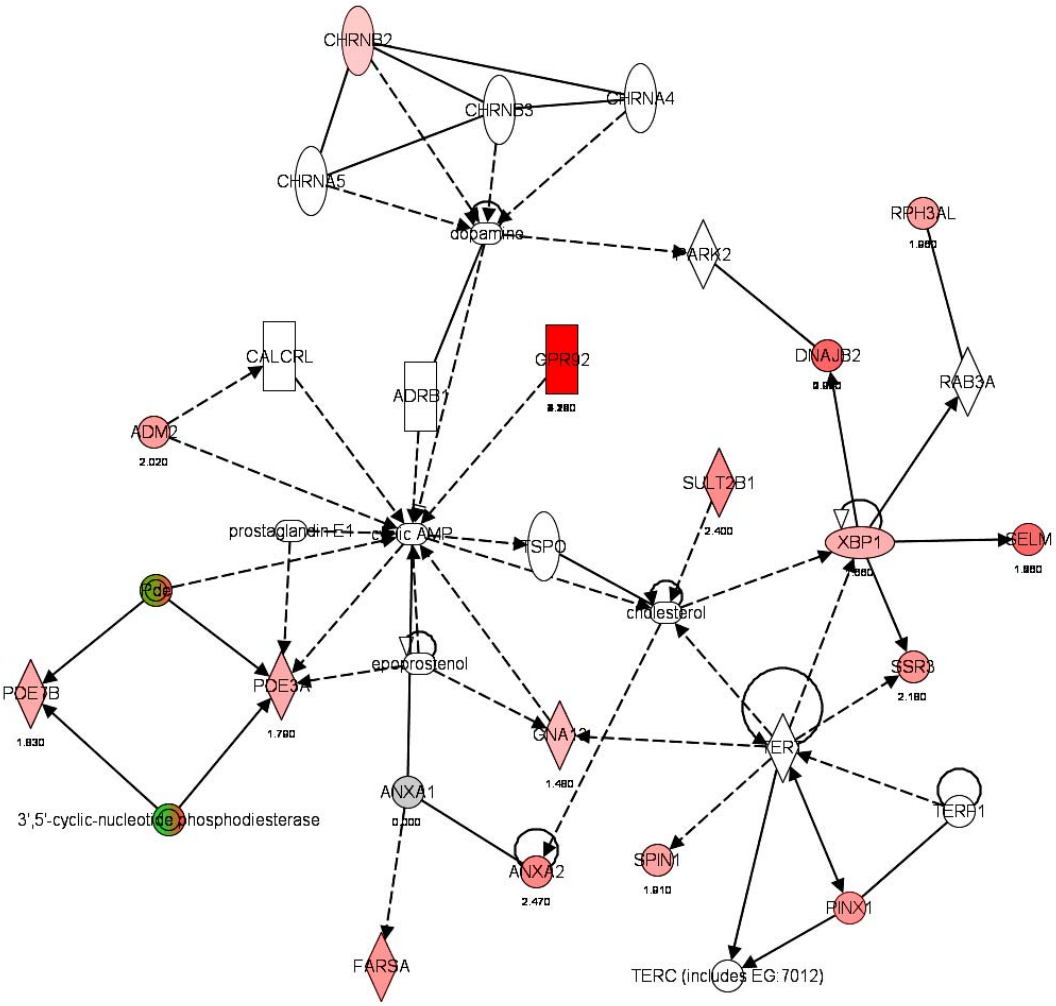


Figure 8-11. Combined ChIP target genes and gene expression data in the same TERT interaction network illustration, showing that a majority of network members are either bound by SMAD3 or significantly up- or down-regulated by TGFβ$_1$.

8.1.4.5 GENE EXPRESSION IN TERT NETWORK AFTER TGFβ$_1$ AND SIS3 TREATMENT

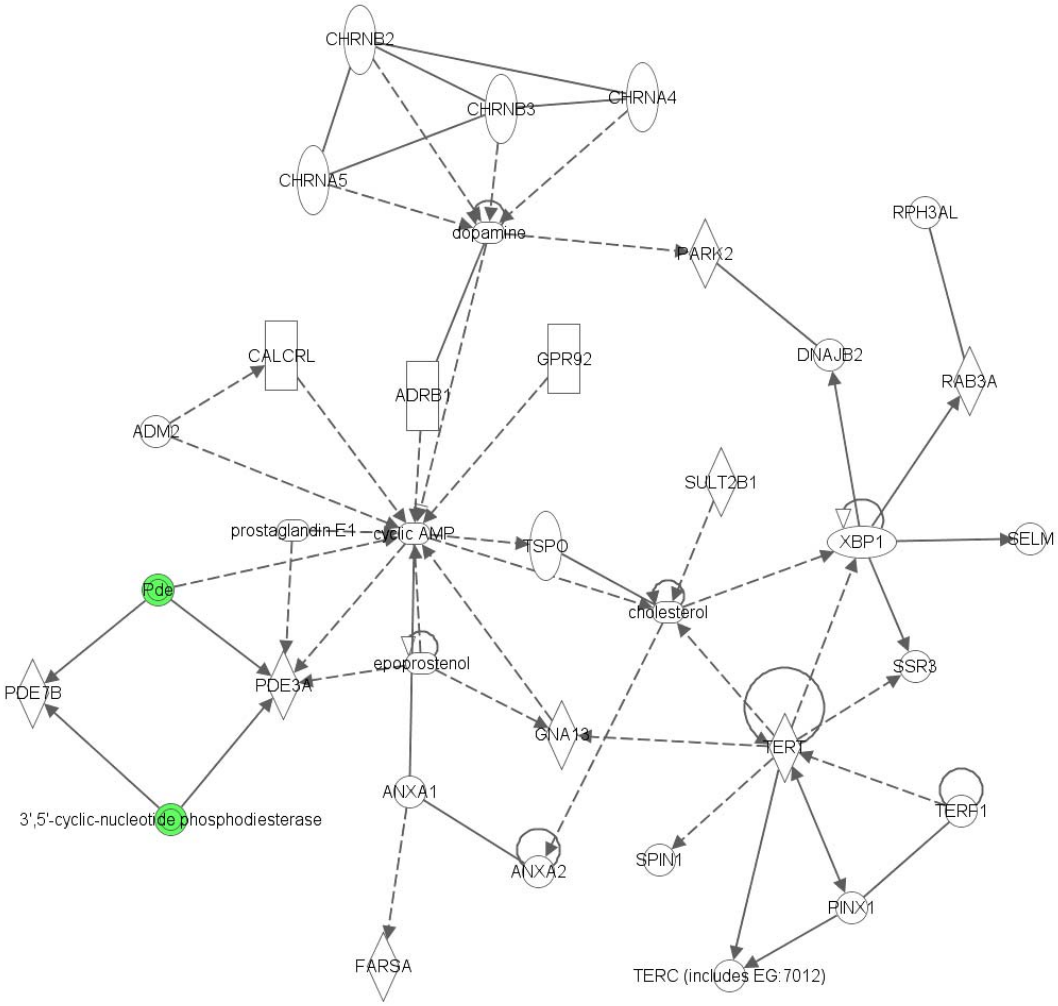


Figure 8-12. Gene expression data in the same TERT interaction network illustration with SIS3 treatment and TGFβ$_1$. Demonstrates that the gene expression profiles seen in the previous network illustrations are likely specifically SMAD3-mediated.

8.2 CHIP-ON-CHIP RESULTS

As described in detail in the methods section, ChIP was performed using an antibody against SMAD3 on A549 cells in two conditions: treated with 2 ng/mL TGFβ_1 and non-treated (control). ChIP-enriched DNA fragments were amplified, labeled, and hybridized to promoter microarrays in triplicate for each condition. Binding peaks were identified by the model-based method of Kaplan and Friedman [288]. The complete lists of found gene whose promoters were identified as bound to SMAD3 (significant binding, peak height of at least 1.5) are listed in Appendix C. Since some of the promoter regions listed on the array gene list were sometimes common to more than one gene, they were hand-curated to ensure the list correctly listed binding for individual, distinct genes. After curation, 350 genes met the binding criteria before TGFβ_1 stimulation, and 474 after 30 minutes TGFβ_1.

8.3 EXPERIMENTAL VERIFICATION OF KNOWN TARGET BINDING

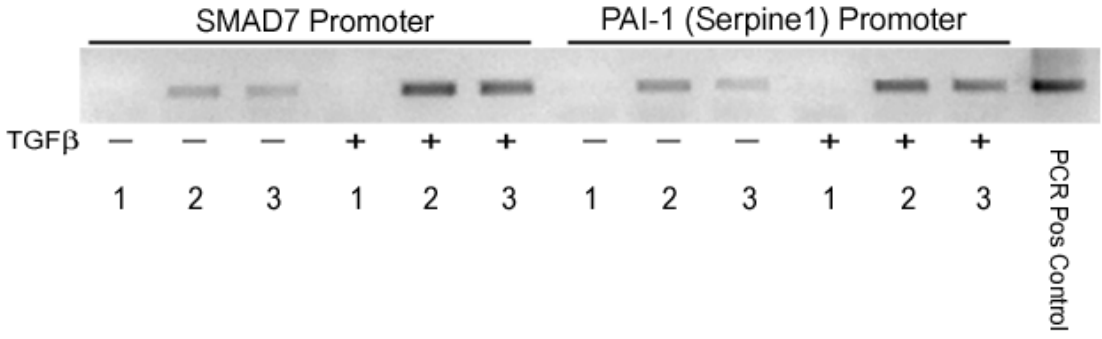


Figure 8-13. Gene-specific PCR of SMAD7 and Serpine-1 promoters. (1) Mock IP (anti-flag Ab); (2) anti-SMAD3 Ab (Upstate Biosciences); (3) anti-SMAD2,3 Ab (BD Biosciences). Data from Oliver Eickelberg, M.D., University of Geissen.

8.4 CHIP PROBE BINDING CURVES OF SELECTED BOUND GENES

The following figures show the intensities of the labeled targets bound to probe clusters in the promoters of their respective genes. Each bar height indicates respective array signal intensity for that probe. Values from the three promoter array replicates are shown (green, blue, purple, respectively). If the binding was statistically significant, the binding curve (red) is also included and shows the fitted peak shape. The bound promoter regions in the following figures show many expected species (serpine1, collagen 7A1, etc.) which serves as a validation of the method.

8.4.1.1 SERPINE1

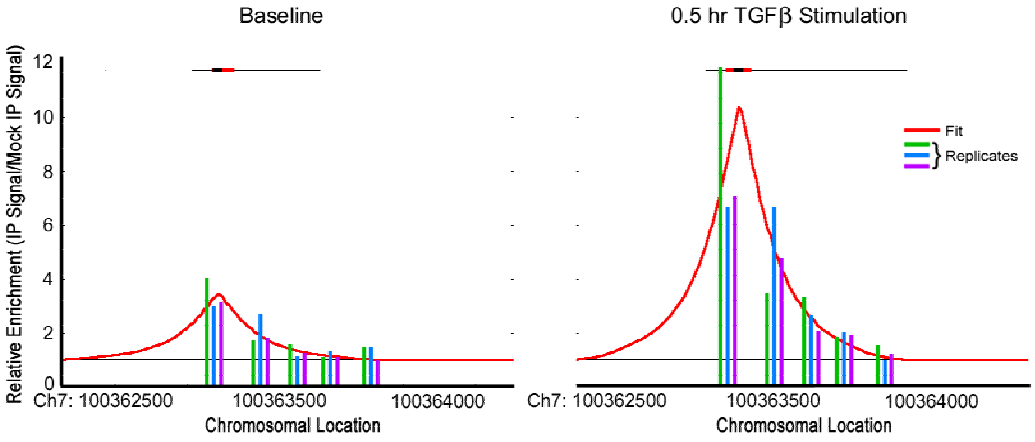


Figure 8-14. Serpine1 (PAI-1) is a well-recognized highly-responsive TGFβ_1-induced gene. The left panel shows baseline promoter binding of SMAD3 in the absence of exogenous TGFβ_1 stimulation. The right panel shows highly increased Serpine1 promoter binding after 30 minutes 2 ng/mL TGFβ_1 stimulation.

8.4.1.2 COLLAGEN 7A1

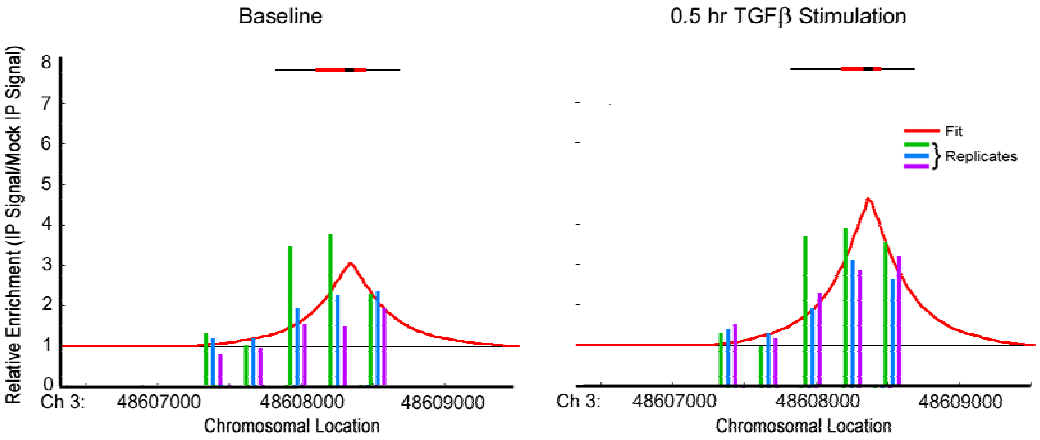


Figure 8-15. Collagen 7A1 is a component of extracellular matrix and a known TGFβ$_1$-induced gene. The left panel shows baseline promoter binding of SMAD3 to the collagen 7A1 promoter in the absence of exogenous TGFβ$_1$ stimulation. The right panel shows highly increased collagen 7A1 promoter binding after 30 minutes 2 ng/mL TGFβ$_1$ stimulation.

8.4.1.3 SMAD6

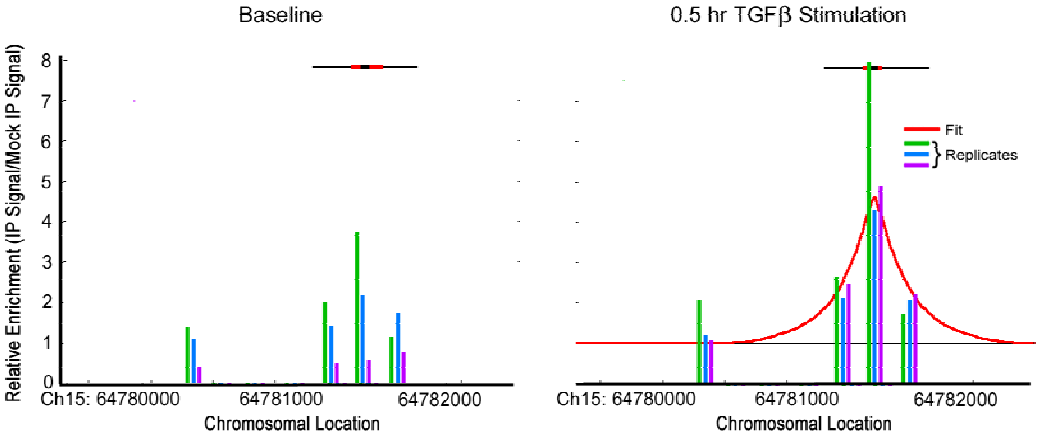


Figure 8-16. SMAD6 is an inhibitory SMAD protein involved in terminating TGFβ activation and is also a known TGFβ-induced gene. The left panel shows baseline promoter binding of SMAD3 to the SMAD6 promoter in the absence of exogenous

TGFβ$_1$ stimulation. The right panel shows highly increased SMAD6 promoter binding after 30 minutes 2 ng/mL TGFβ$_1$ stimulation.

8.4.1.4 SMAD7

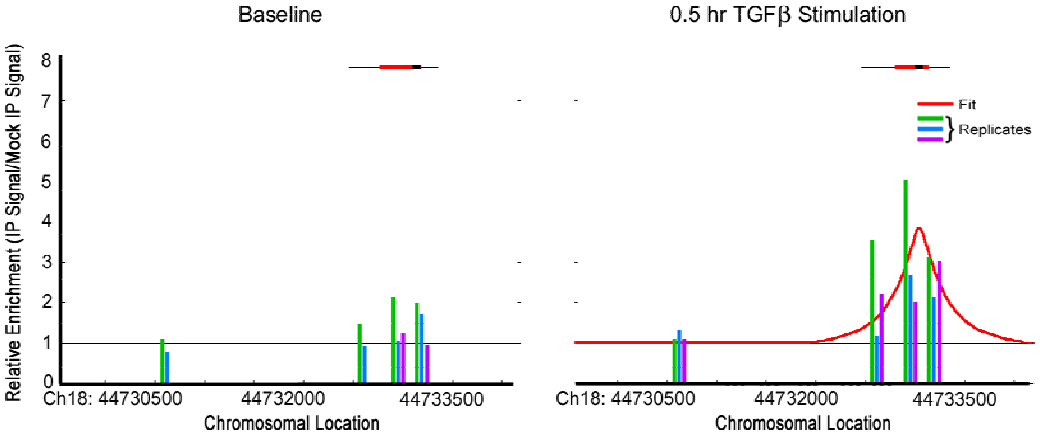


Figure 8-17. SMAD7 is another inhibitory SMAD protein involved in terminating TGFβ$_1$ activation and a known TGFβ$_1$-induced gene. The left panel shows baseline promoter binding of SMAD3 in the absence of exogenous TGFβ$_1$ stimulation. The right panel shows highly increased Serpine1 promoter binding after 30 minutes 2 ng/mL TGFβ$_1$ stimulation.

8.4.1.5 TGFβ$_1$

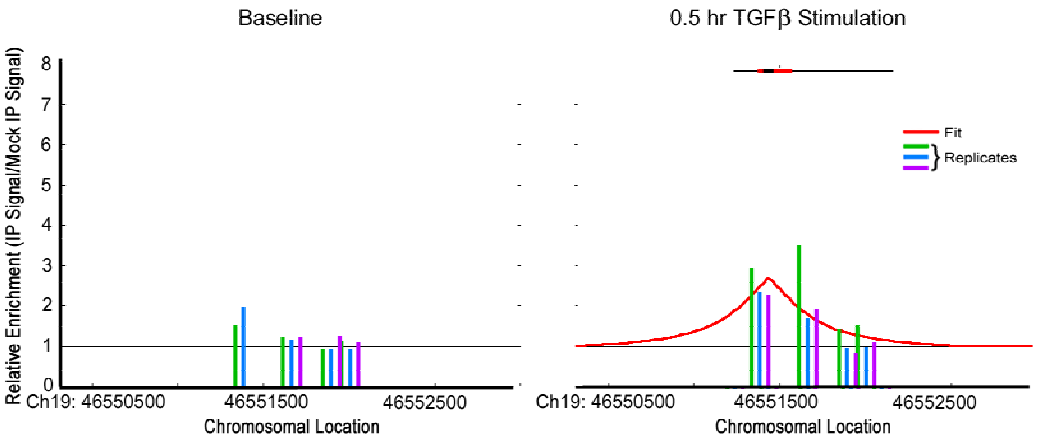


Figure 8-18. TGFβ$_1$ is known to auto-induce. The left panel shows baseline promoter binding of SMAD3 to the TGFβ$_1$ promoter in the absence of exogenous TGFβ$_1$

stimulation. The right panel shows highly increased TGFβ_1 promoter binding after 30 minutes 2 ng/mL TGFβ_1 stimulation.

8.4.1.6 LATENT TRANSFORMING GROWTH FACTOR BINDING PROTEIN 3 (LTBP3)

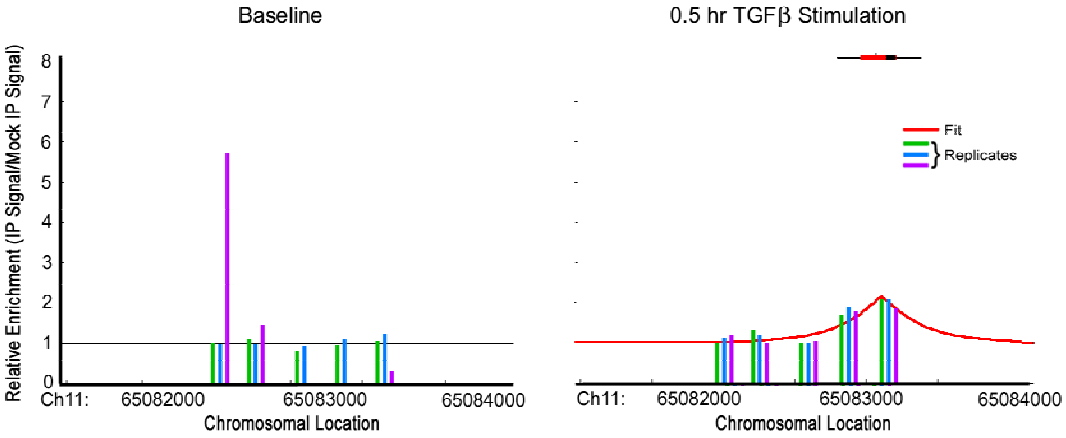


Figure 8-19. Latent Transforming Growth Factor Binding Protein 3 (LTBP3) is a protein responsible for binding TGFβ_1 in the extracellular space in inactivated form (see Chapter 3). The left panel shows baseline promoter binding of SMAD3 to the LTBP3 promoter in the absence of exogenous TGFβ_1 stimulation. The right panel shows highly increased LTBP3 promoter binding after 30 minutes 2 ng/mL TGFβ_1 stimulation.

8.4.2 INGENUITY PATHWAYS FUNCTIONAL GENE GROUPING, CHIP-ON-CHIP: BIOLOGICAL FUNCTION

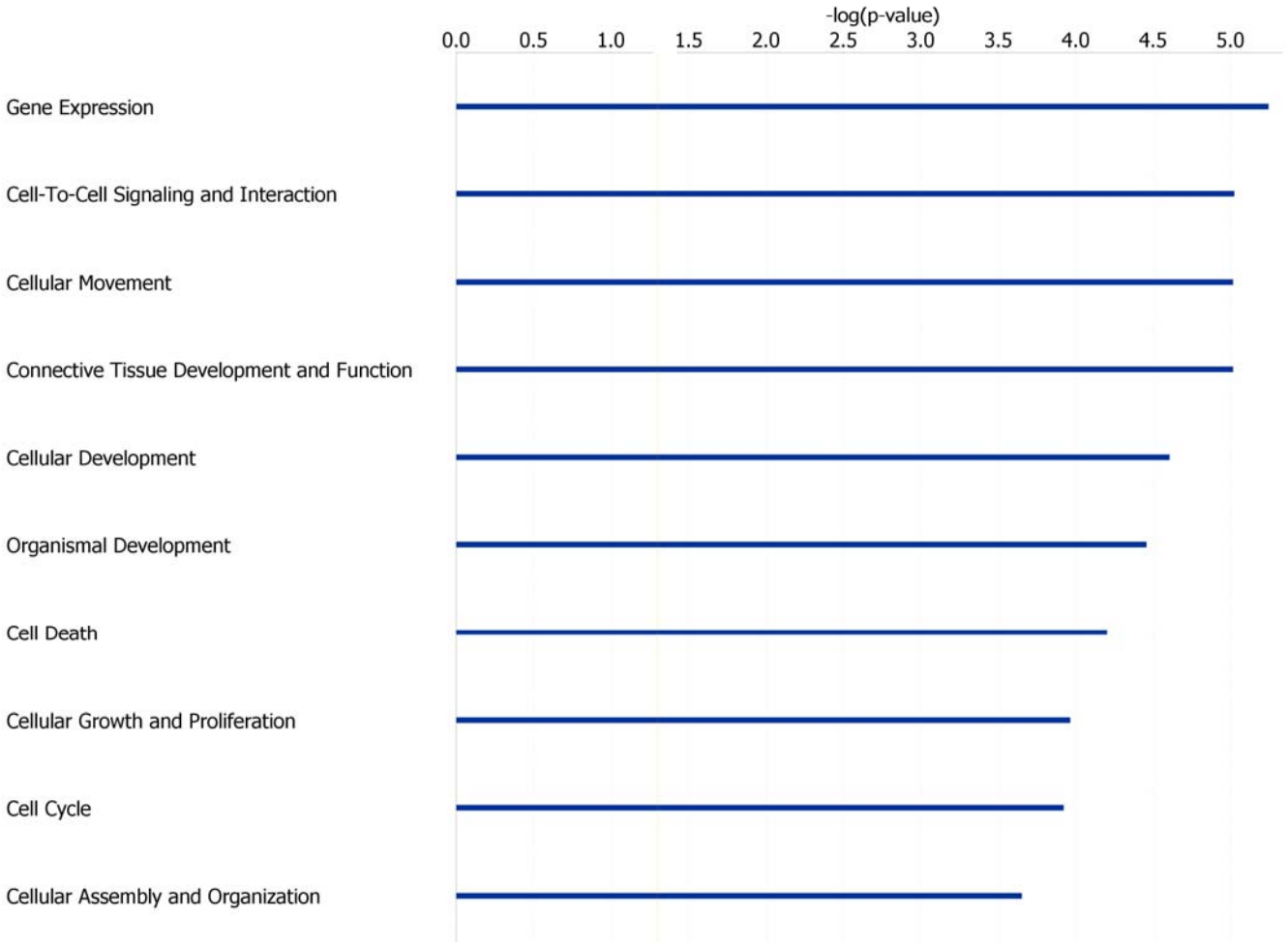


Figure 8-20. Functional grouping of biological functions of ChIP SMAD3-bound target genes, ranked by statistical significance. Cell-to-Cell Signaling and Interaction and Cellular Movement are consistent with epithelial cells undergoing epithelial-to-mesenchymal transition (EMT). Connective Tissue Development and Function is consistent with cells producing and depositing extracellular matrix proteins. Organismal Development, Cell Death, Cellular Growth and Proliferation, and Cell Cycle are all functions consistent with the known functions of the growth factor/cytokine, $TGF\beta_1$. From Ingenuity Pathways Analysis [32].

8.4.3 INGENUITY PATHWAYS FUNCTIONAL GENE GROUPING, CHIP-ON-CHIP: PHYSIOLOGICAL FUNCTION

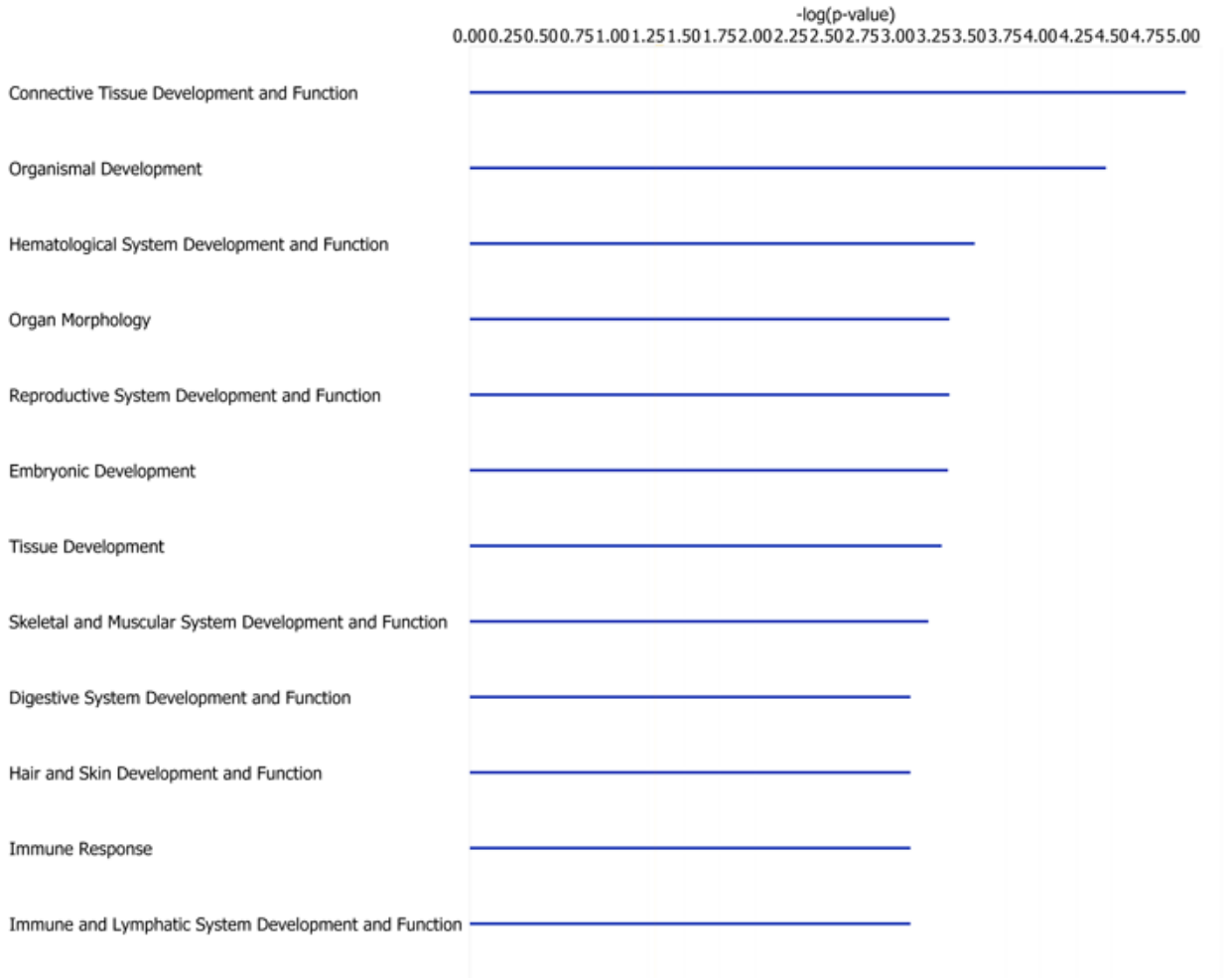


Figure 8-21. Functional grouping of ChIP SMAD3-bound target by physiological function, ranked by statistical significance. Connective Tissue Development and Function is consistent with cells producing and depositing extracelullar matrix proteins. From Ingenuity Pathways Analysis [32].

8.4.4 INGENUITY PATHWAYS FUNCTIONAL GENE GROUPING, CHIP-ON-CHIP: SIGNALING PATHWAYS

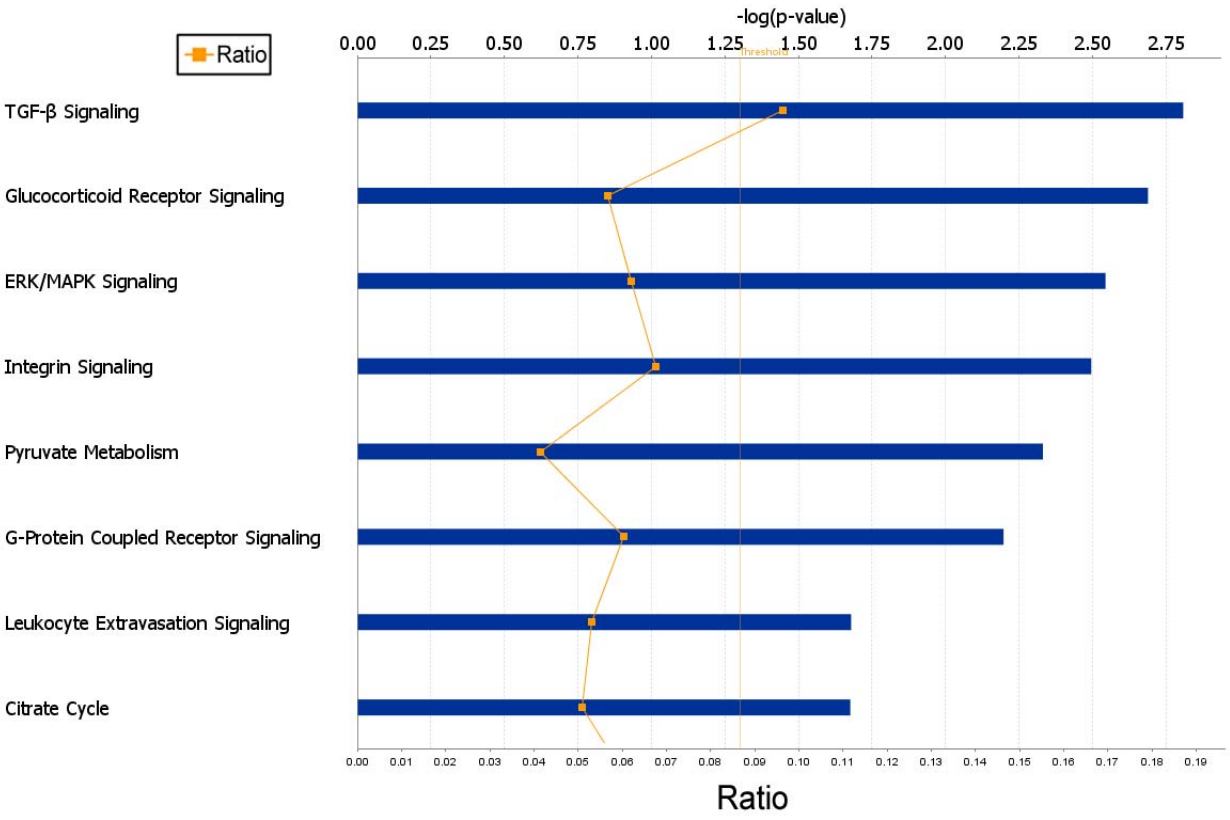


Figure 8-22. ChIP SMAD3-bound target genes grouped by signaling pathway and ranked in order of statistical significance. The ratio of genes (orange line) refers to number of genes involved in pathway divided by total genes; approximately 10% of bound genes are identified as belonging to the known $TGF\beta_1$ signaling pathway. Other prominent signaling pathways include ERK/MAPK and Integrin Signaling, which is consistent with known interactions of $TGF\beta_1$. From Ingenuity Pathways Analysis.

8.5 GENE EXPRESSION MICROARRAY RESULTS

Short Time Series Expression Mining (STEM) identified ten highly significant expression time course profiles (Figure 8-1). Expression profiles 41, 14, 23, 43, and 39 exhibit up-regulation activity while profiles 9, 44, 37, 24, and 19 exhibit down-regulation activity.

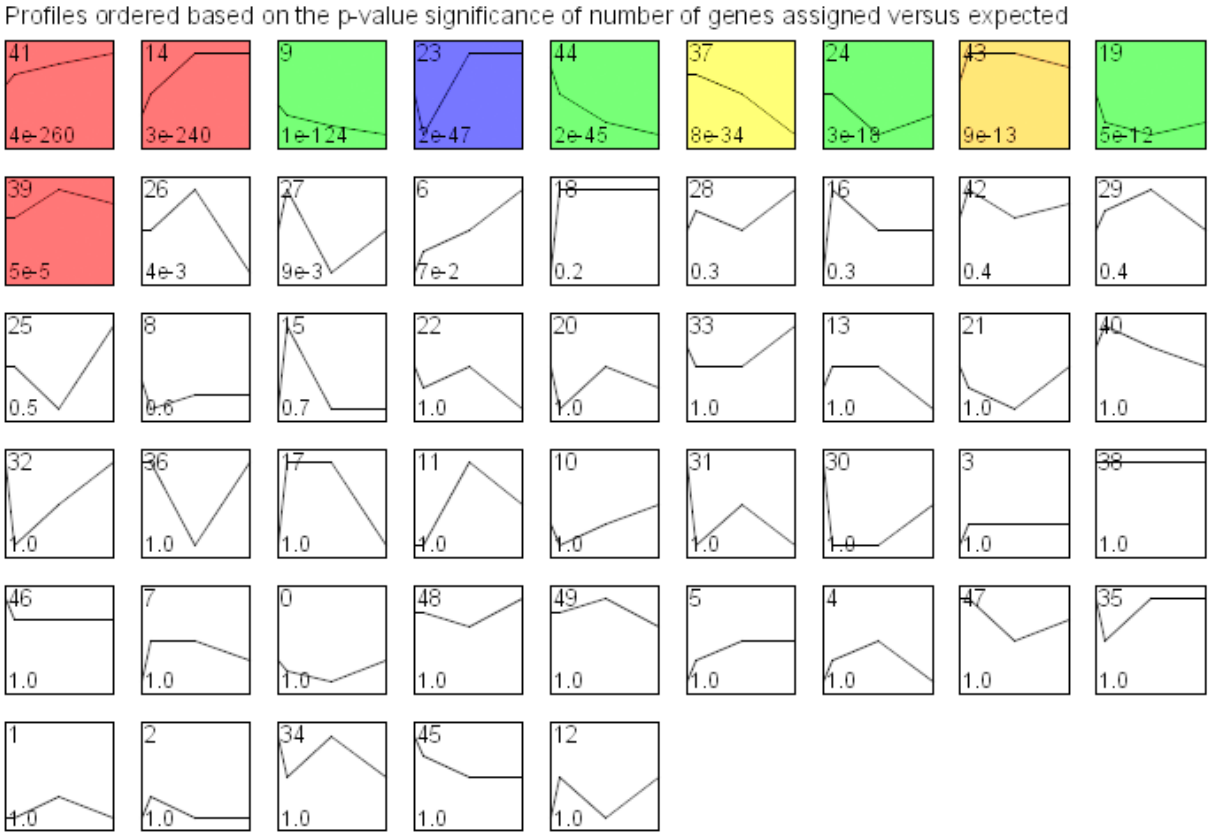


Figure 8-23. Screen output of STEM program showing identification of statistically signifant gene expression time course profiles.

Specific gene expression microarray results are consistent with many of the known TGFβ_1/SMAD3-responsive elements as discussed previously in Chapter 3; specific instances and functional clusters of genes will be illustrated below. Further, gene expression profiles of

$TGF\beta_1$-simulated cells previously treated with SIS3, a specific and potent SMAD3/ALK5 phosphorylation inhibitor, suggests that the changes we see in gene expression are indeed mediated specifically through the $TGF\beta_1$/SMAD3 regulatory pathway. In the following figure we see Serpine1 (PAI-1), SMAD6, SMAD7, $TGF\beta_1$, SMURF1, and Connective Tissue Growth Factor (CTGF) are highly upregulated after $TGF\beta_1$ simulation. Further, Jun, JunB, and the ERK/MAPK pathways are also affected by $TGF\beta_1$/SMAD3 as expected. These gene expression profiles were all identified as significantly up- or down-regulated ($p<0.00001$) by STEM [290, 291]. As added confirmation of the microarray results, Serpine1 (PAI-1) mRNA levels were measured by quantitative real-time PCR (qRT-PCR) in both the $TGF\beta_1$ stimulation time-series with and without SIS3 inhibition. The qRT-PCR results are statistically highly significant and are consistent with both the expected behavior of Serpine1 after $TGF\beta_1$ stimulation as well as the gene expression micorarray results (see Figure 8-11) below.

Figure 8-24. Heat map of average expression values for genes known to be affected by the TGFβ_1/SMAD3 pathway. Color intensity values correspond to $\log_2$ of absolute intensity and reach saturation on the heat map at value 4 to preserve dynamic range at lower values. The time series is in hours after TGFβ_1 stimulation and vehicle only (DMSO; left) and with TGFβ_1 stimulation and also inhibition of SMAD3/ALK5 phosphorylation by Specific Inhibitor of SMAD3 (SIS3) (right)[293]. The gene expression profiles on the left (non-SIS3-treated) were all identified as significantly up- or down-regulated ($p<0.00001$) by STEM [290, 291].

Similarly, in the following heat map we see significant up- and down-regulation of known key mediators of epithelial-mesenchymal transition (EMT). Again, most of the stimulatory effects of exogenous TGFβ_1 administration are largely abrogated by SIS3, suggesting specific transcriptional regulation of these genes is mediated through the SMAD3/ TGFβ_1 pathway.

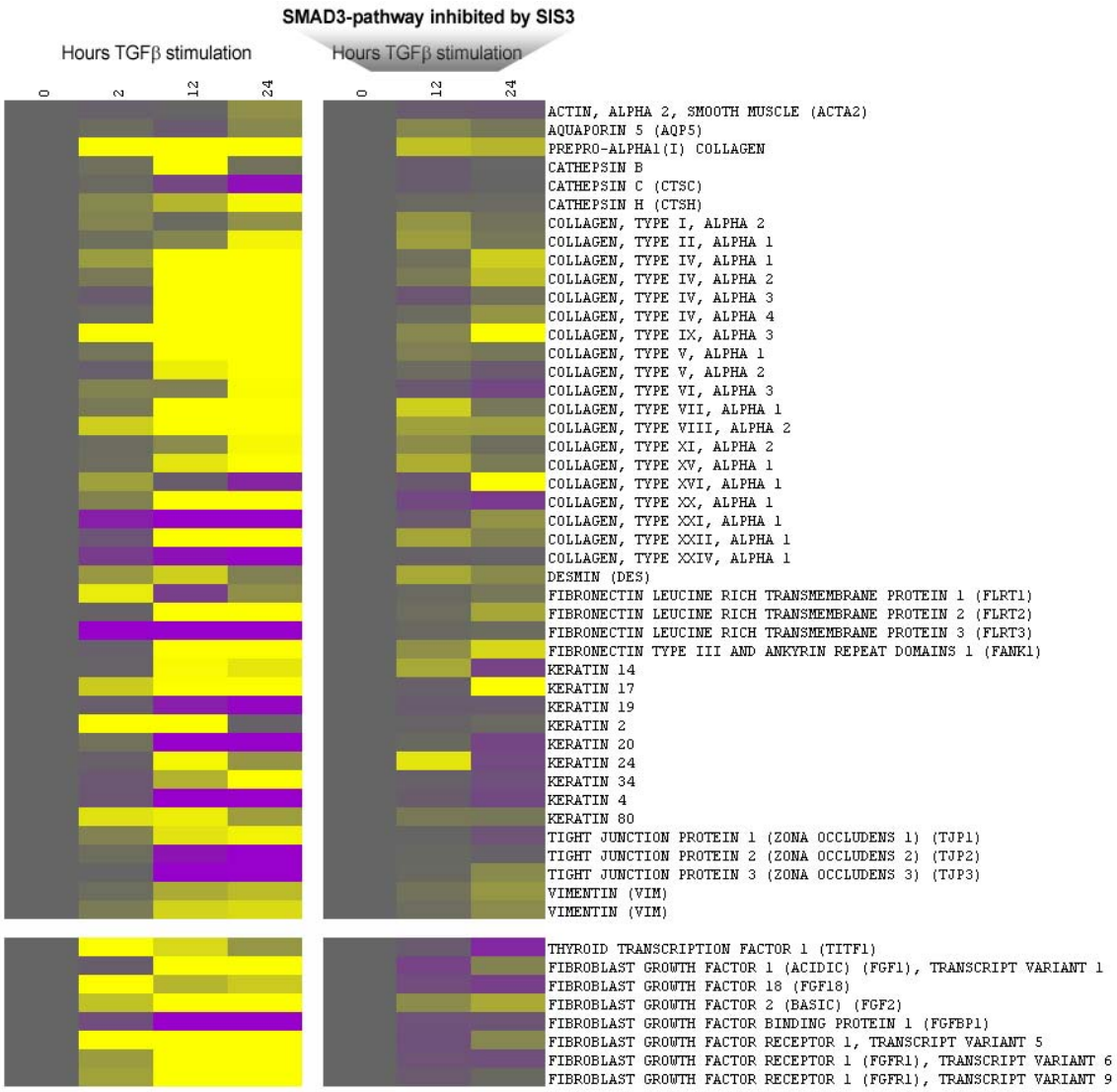


Figure 8-25. Heat map of average expression values for genes associated with epithelial-mesenchymal transition (EMT). Time series in hours after TGFβ_1 stimulation (left) and with TGFβ_1 stimulation and inhibition of SMAD3/ALK5 phosphorylation by Specific Inhibitor of SMAD3 (SIS3) (right) [293]. With the exception of α-smooth muscle actin on the first row (ACTA2) and collagen type I on the fourth row, the gene expression profiles on the left (non-SIS3-treated) were all identified as significantly up- or down-regulated ($p<0.00001$) by STEM [290, 291].

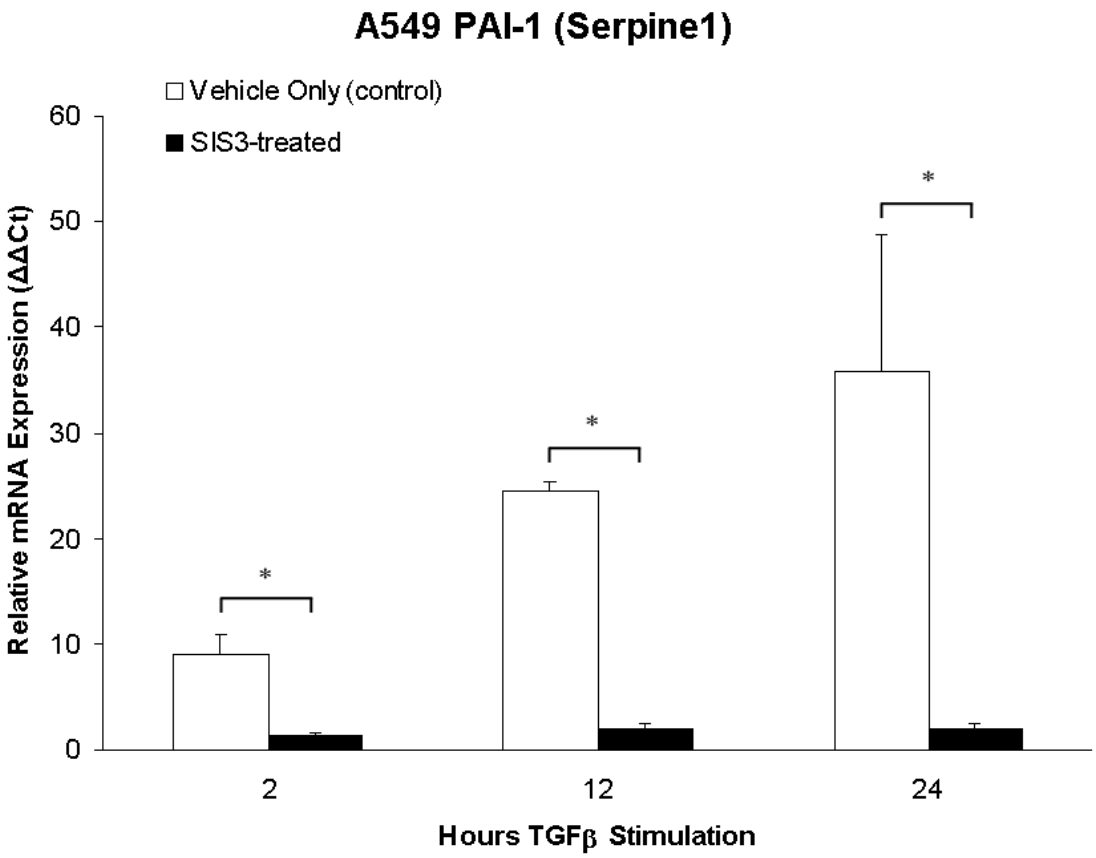


Figure 8-26. Quantitative real-time PCR results confirming induction of Serpine1 (PAI-1) in A549 cells after 2 ng/ml TGFβ_1 simulation at 2, 12, and 24 hours respectively. The up-regulation of Serpine1 was clearly suppressed by treatment with SIS3. The asterisk denotes a highly statistically significant ($p<0.001$; $n=3$) difference at each time point between SIS3-treated and vehicle-only controls after TGFβ_1 treatment.

Figure 8-27. Heat map of TGFβ_1 time series expression profiles of highest up-regulated genes.

Figure 8-28. Heat map of TGFβ_1 time series expression profiles of highest down-regulated genes.

8.6 FUNCTIONAL ANALYSIS OF COMBINED CHIP-CHIP AND GENE EXPRESSION DATA

The following diagrams illustrate the overlap between significant gene expression values and ChIP SMAD3 binding data:

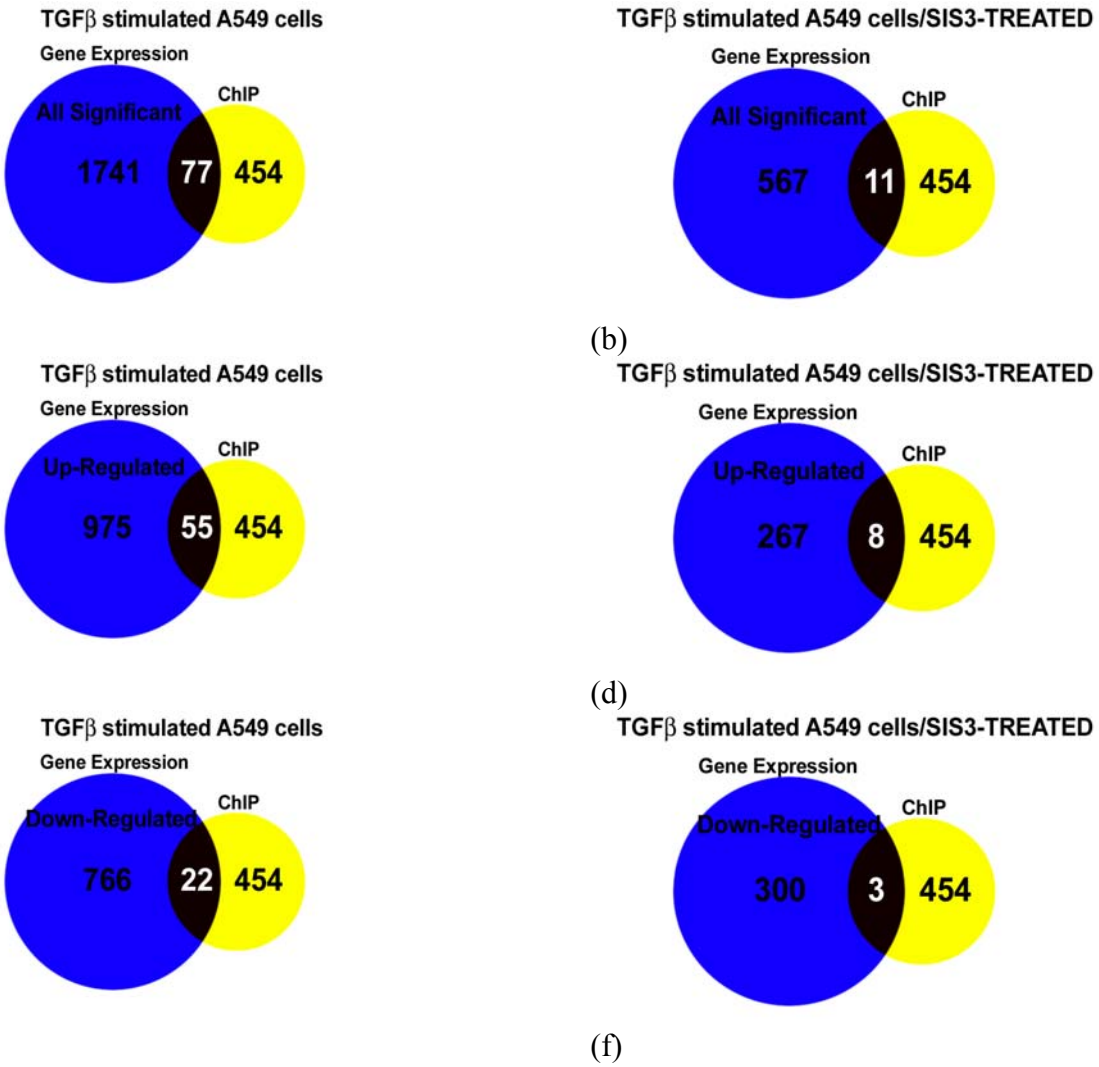


Figure 8-29. Venn diagrams of combined ChIP (yellow) and gene expression (blue) data. Numbers denote significant genes total (a,b) and both up- (c,d) and down-regulated (e,f)

in comparison to ChIP. The left column is TGFβ_1 simulated A549 cells. The right column is TGFβ_1 simulated A549 cells treated with SIS3 SMAD3-inhibitor.

8.6.1 COMBINED CHIP-ON-CHIP AND GENE EXPRESSION HEAT MAP

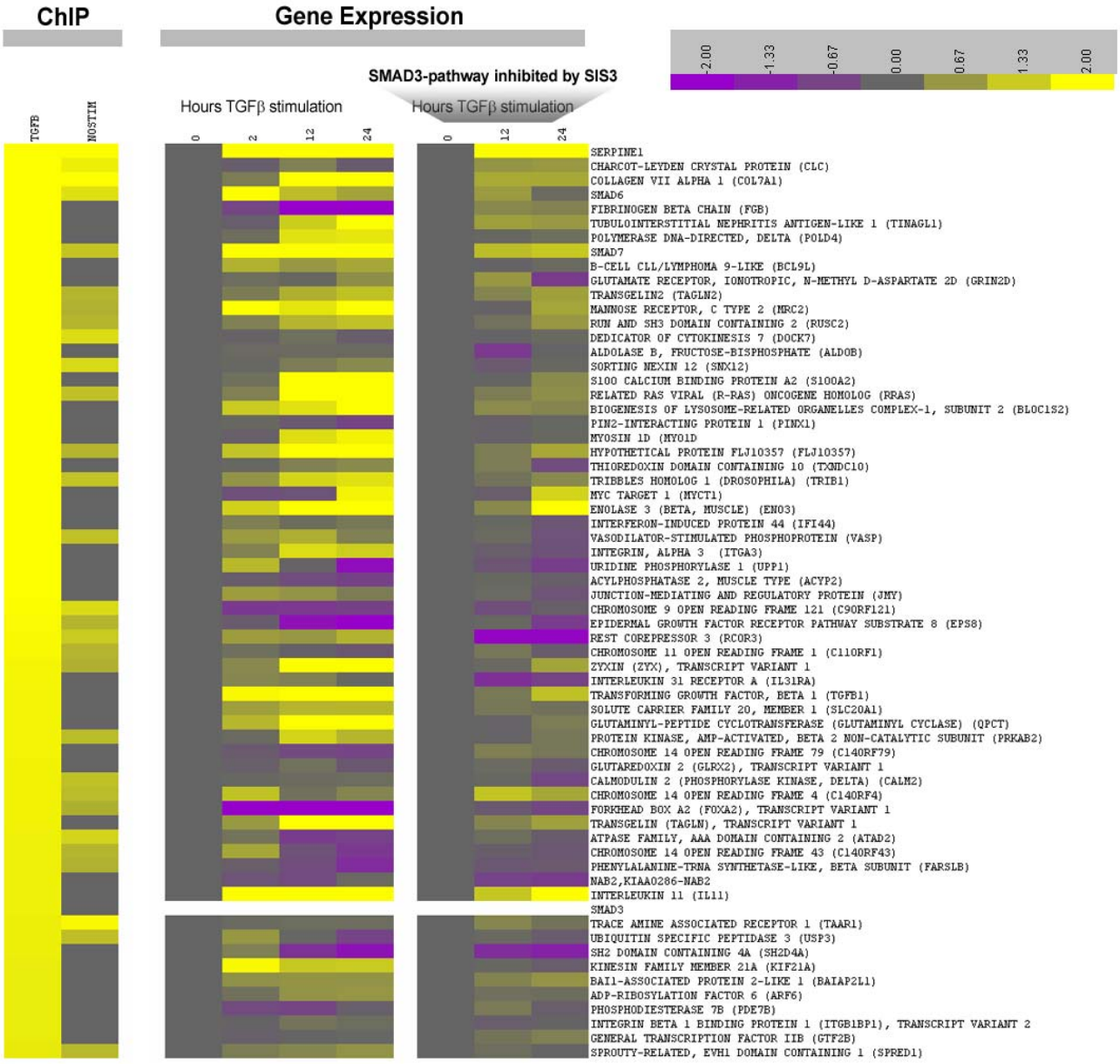


Figure 8-30. Heat map of SMAD3-target genes from ChIP sorted by peak height intensity for TGFβ_1-stimulated A549 cells, alongside respective gene expression microarray intensities for the same gene (vehicle, middle series) and SIS3-treated (right series).

Extensive functional analysis was performed on ChIP data, gene expression data, and combined data using several systems biology software tools. These include Ingenuity Pathways Analysis (IPA) [32], and Metacore GeneGo [33]. First, the list of significantly bound targets from ChIP was submitted for analysis through Ingenuity Pathways Analysis and MetaCore GeneGo and matched to their participation in known biological functions, physiological functions, and signal transduction pathways.

8.6.2 METACORE GENEGO FUNCTIONAL GROUPING

ChIP SMAD3-binding and gene expression data were also analyzed in combination with MetaCore GeneGo systems biology tools [33]. The results were consistent with TGFβ_1 signaling epithelial-mesenchymal transition (e.g., cytoskeletin remodeling, changes to cellular adhesion), and collagen deposition..

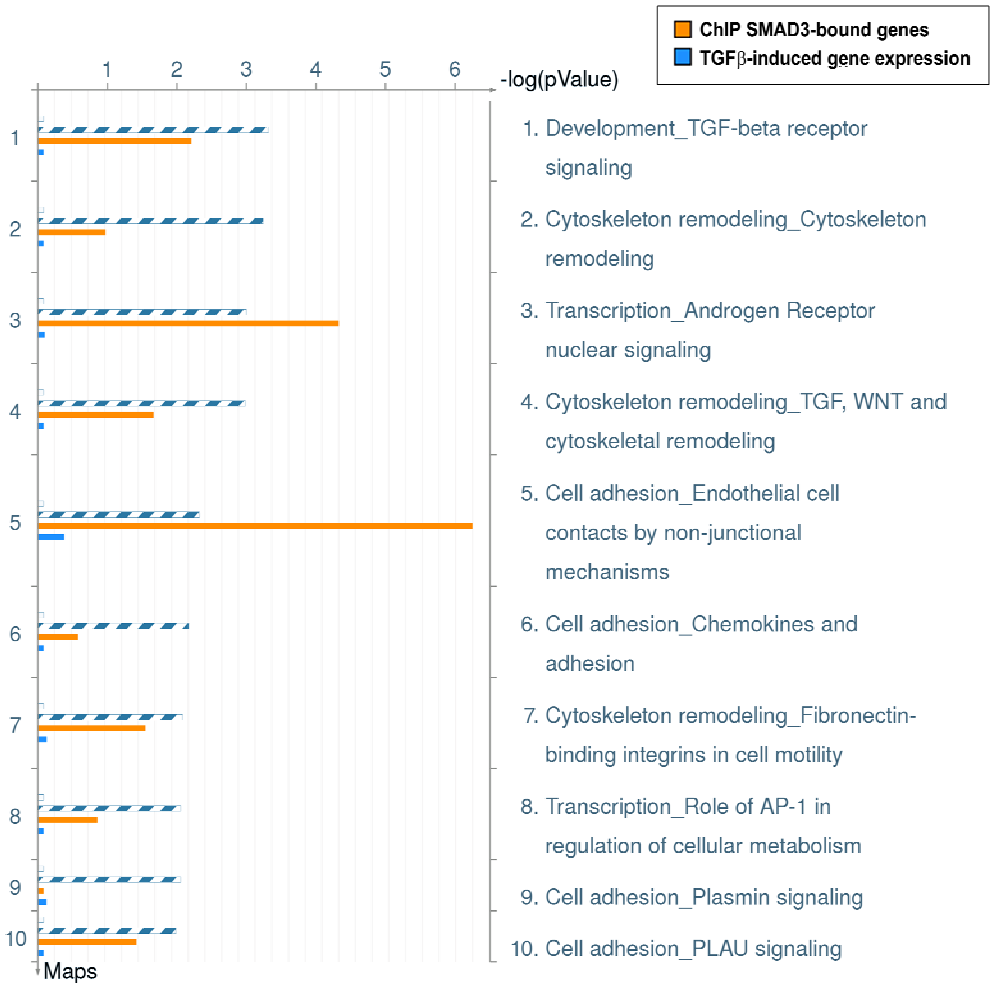


Figure 8-31. Combined ChIP SMAD3 and TGFβ_1-induced gene expression data grouped according to membership in known signaling pathways.

8.6.3 METACORE FUNCTIONAL GROUPING

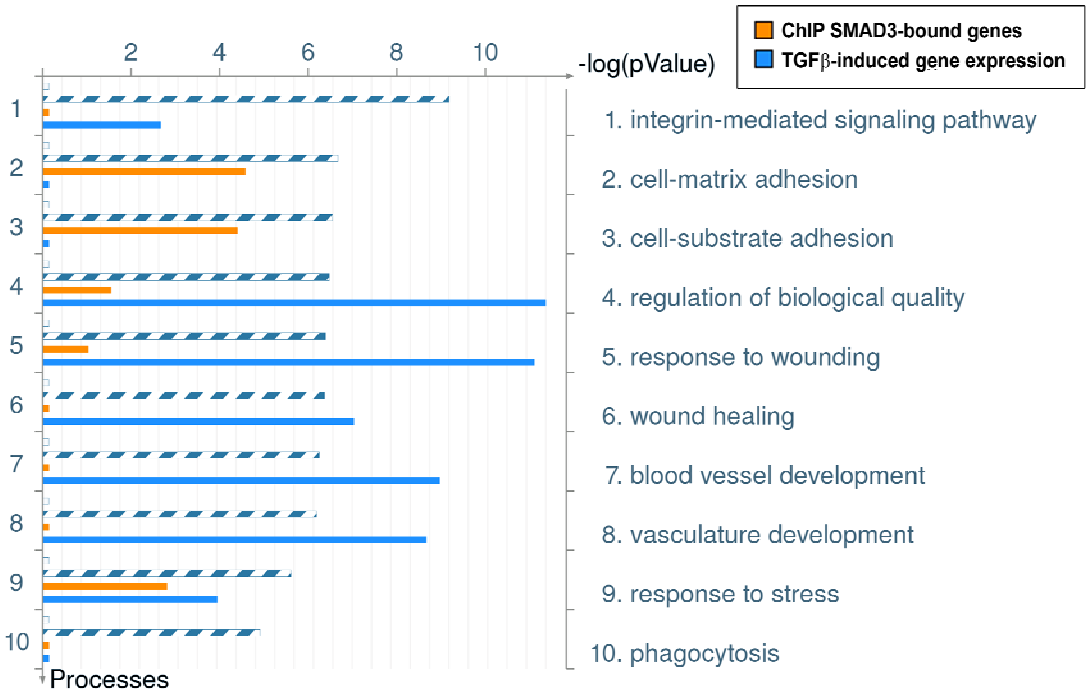


Figure 8-32. Combined ChIP SMAD3 and TGFβ_1-induced gene expression data grouped according to membership in known cellular processes.

8.6.4 METACORE FUNCTIONAL GROUPING

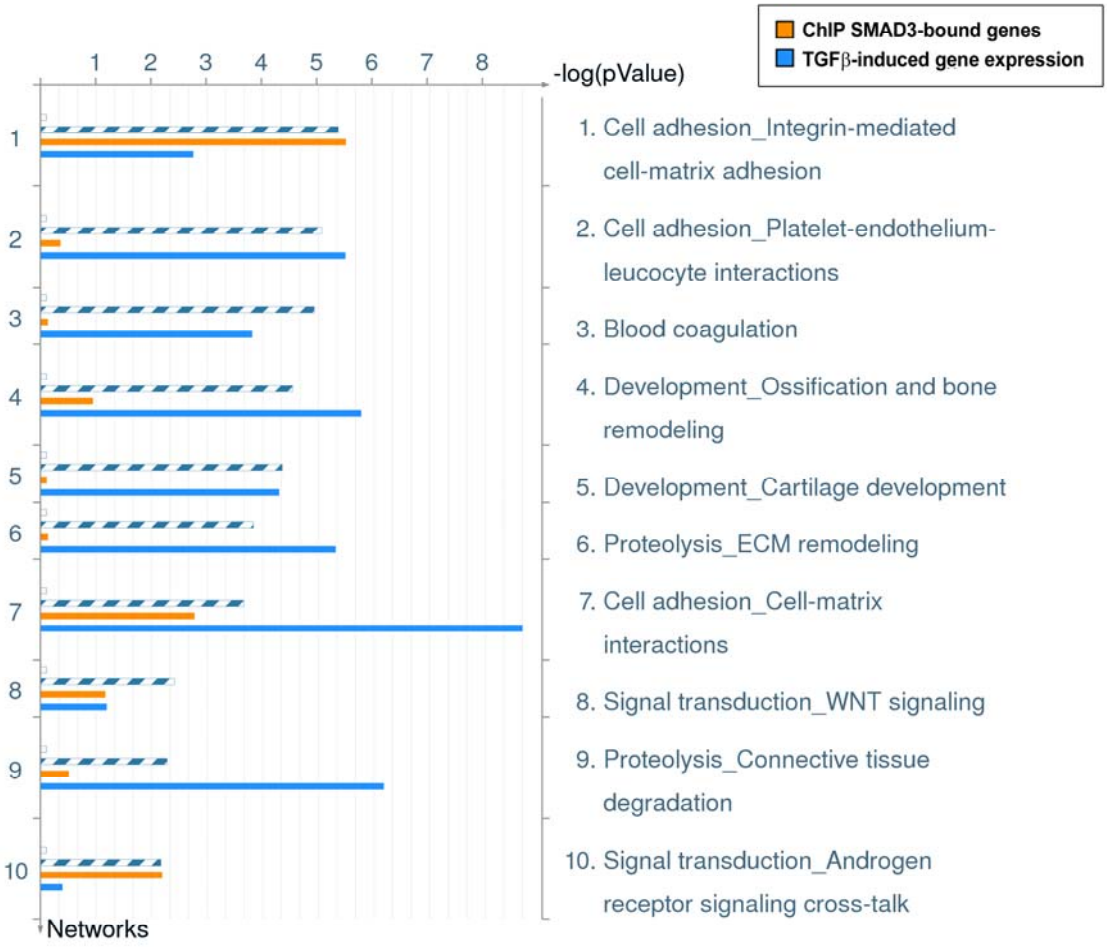


Figure 8-33. Combined ChIP SMAD3 and TGFβ_1-induced gene expression data grouped according to membership in known physiological responses.

8.6.5 TGFβ$_1$,/SMAD3 SIGNALING PATHWAY—CHIP

The following figures illustrate a systems-level visual integration of data from ChIP and gene expression data individually, and then combined in several canonical signaling pathways. The figures show involvement of SMAD3 binding and TGFβ$_1$-induced gene expression respectively and combined in TGFβ$_1$, ERK/MAPK, p38 MAPK, and NFκB signaling. It is known that the TGFβ$_1$/SMAD3 signaling pathways interact with ERK/MAPK, p38 MAPK, and NFκB. However, these visualization tools show the specific relationships between promoter binding and gene expression in the signaling pathway elements.

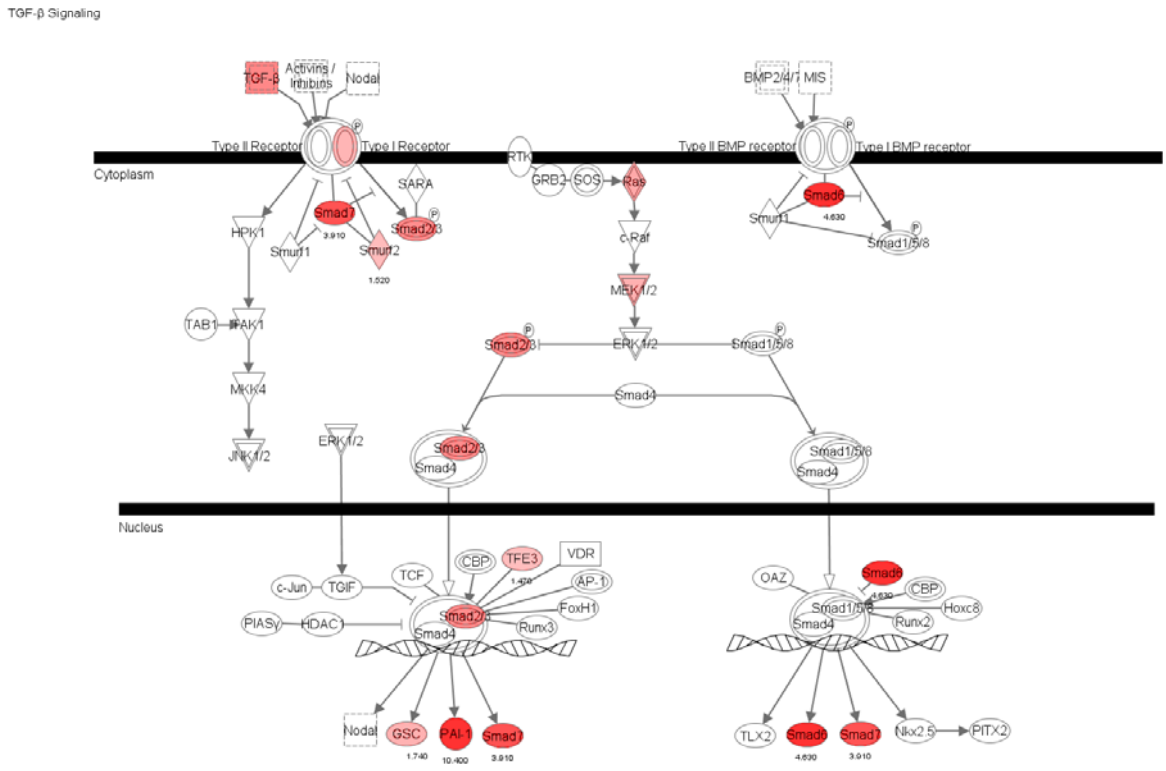


Figure 8-34. ChIP SMAD3-bound target genes illustrated in the TGFβ$_1$ signaling pathway. Red denotes bound target gene; Color intensity depicts ChIP SMAD3 binding peak height.

8.6.6 TGFβ$_1$,/SMAD3 SIGNALING PATHWAY—GENE EXPRESSION

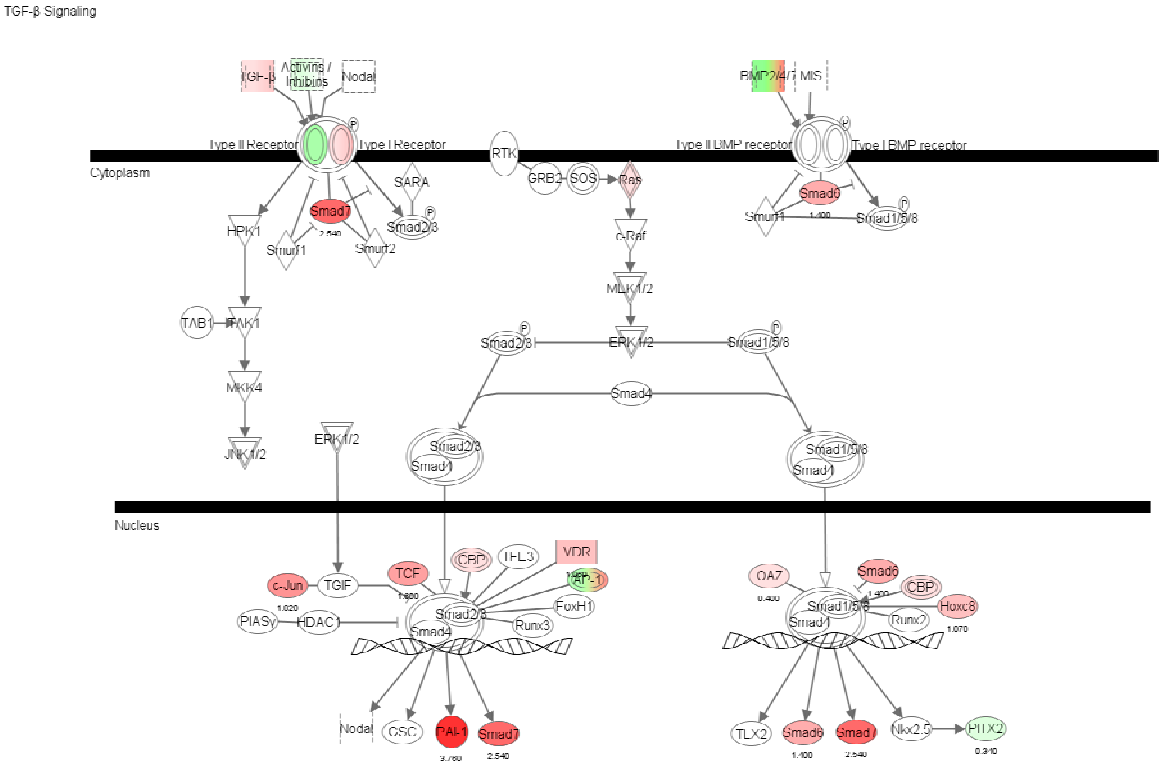


Figure 8-35. Gene expression values illustrated in the TGFβ$_1$ signaling pathway. Significant gene expression profiles after TGFβ$_1$ stimulation; red denotes up-regulation, green denotes down-regulation. Color intensity depicts relative gene expression levels.

8.6.7 TGFβ$_1$,/SMAD3 SIGNALING PATHWAY—COMBINED

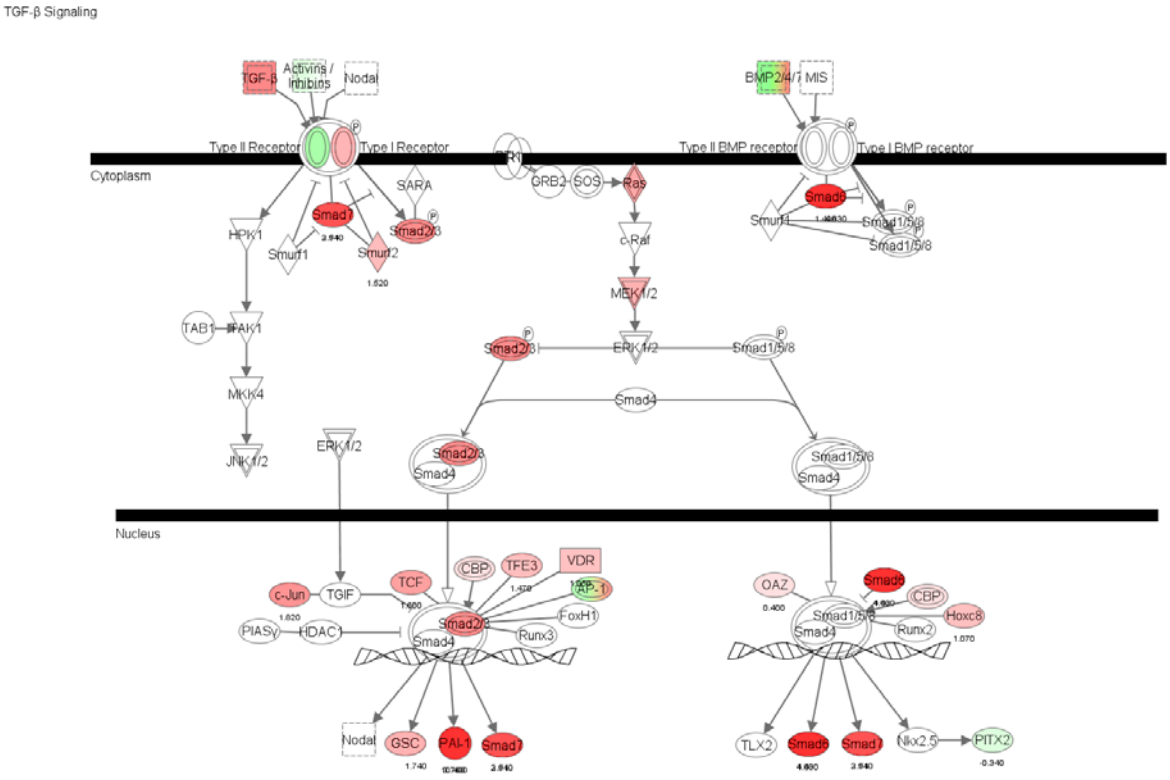


Figure 8-36. Combined ChIP target genes and gene expression data in the same TGFβ$_1$ signaling pathway illustration.

8.6.8 ERK/MAPK SIGNALING PATHWAY—CHIP

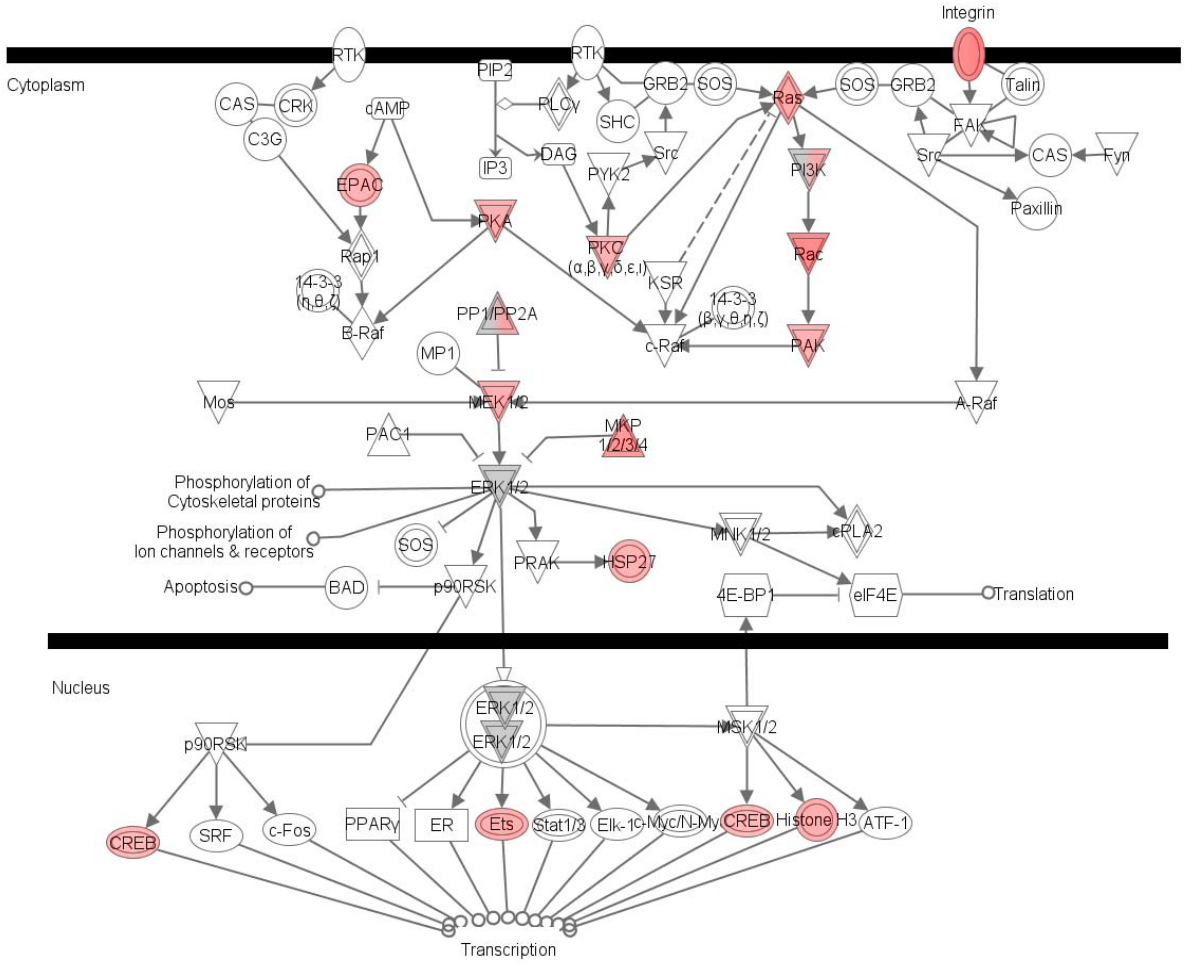


Figure 8-37. ChIP SMAD3-bound target genes illustrated in the ERK/MAPK signaling pathway. Red denotes bound target gene; Color intensity depicts ChIP SMAD3 binding peak height.

8.6.9 ERK/MAPK SIGNALING PATHWAY—GENE EXPRESSION

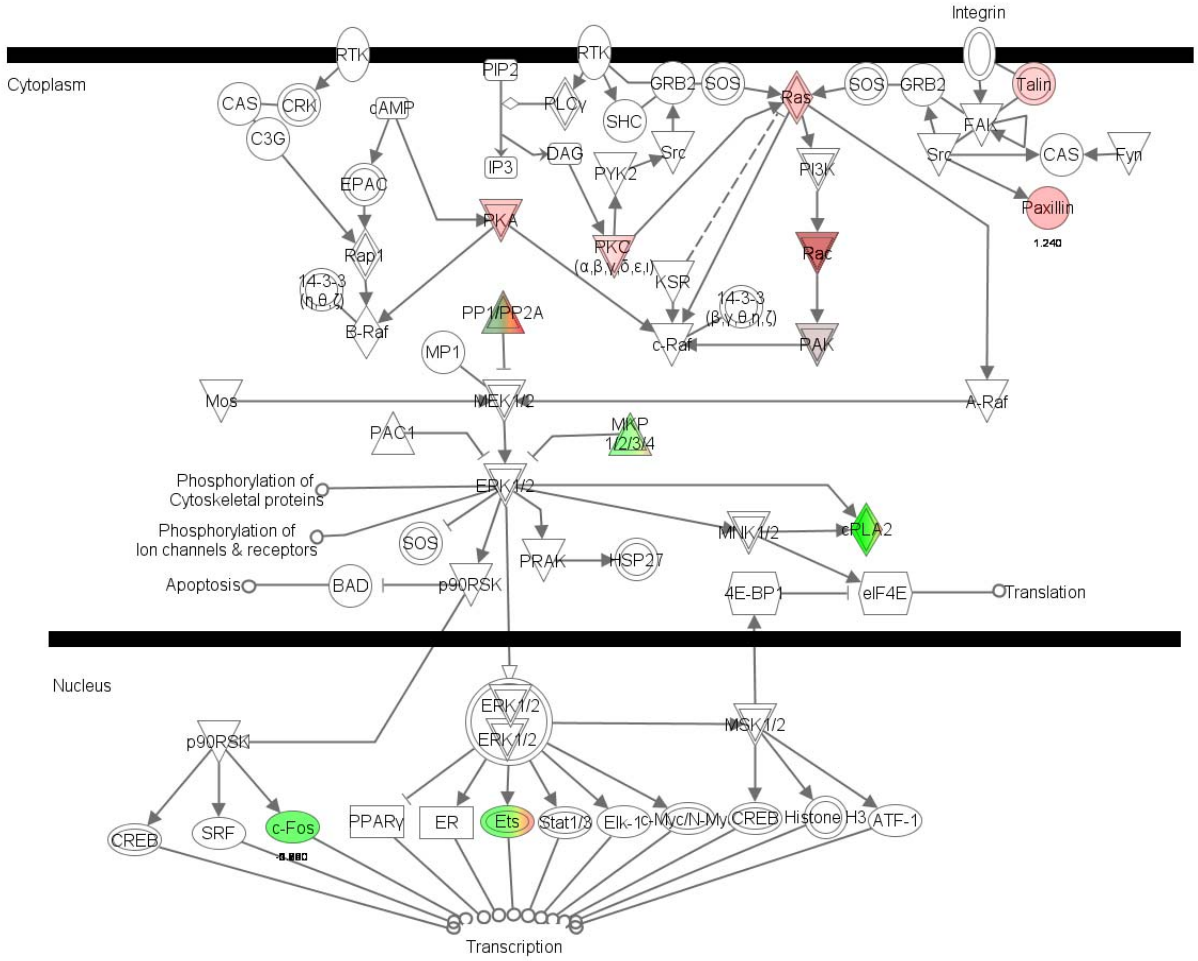


Figure 8-38. Gene expression values illustrated in the ERK/MAPK signaling pathway. Significant gene expression profiles after TGFβ_1 stimulation; red denotes up-regulation, green denotes down-regulation. Color intensity depicts relative gene expression levels.

8.6.10 ERK/MAPK SIGNALING PATHWAY—COMBINED

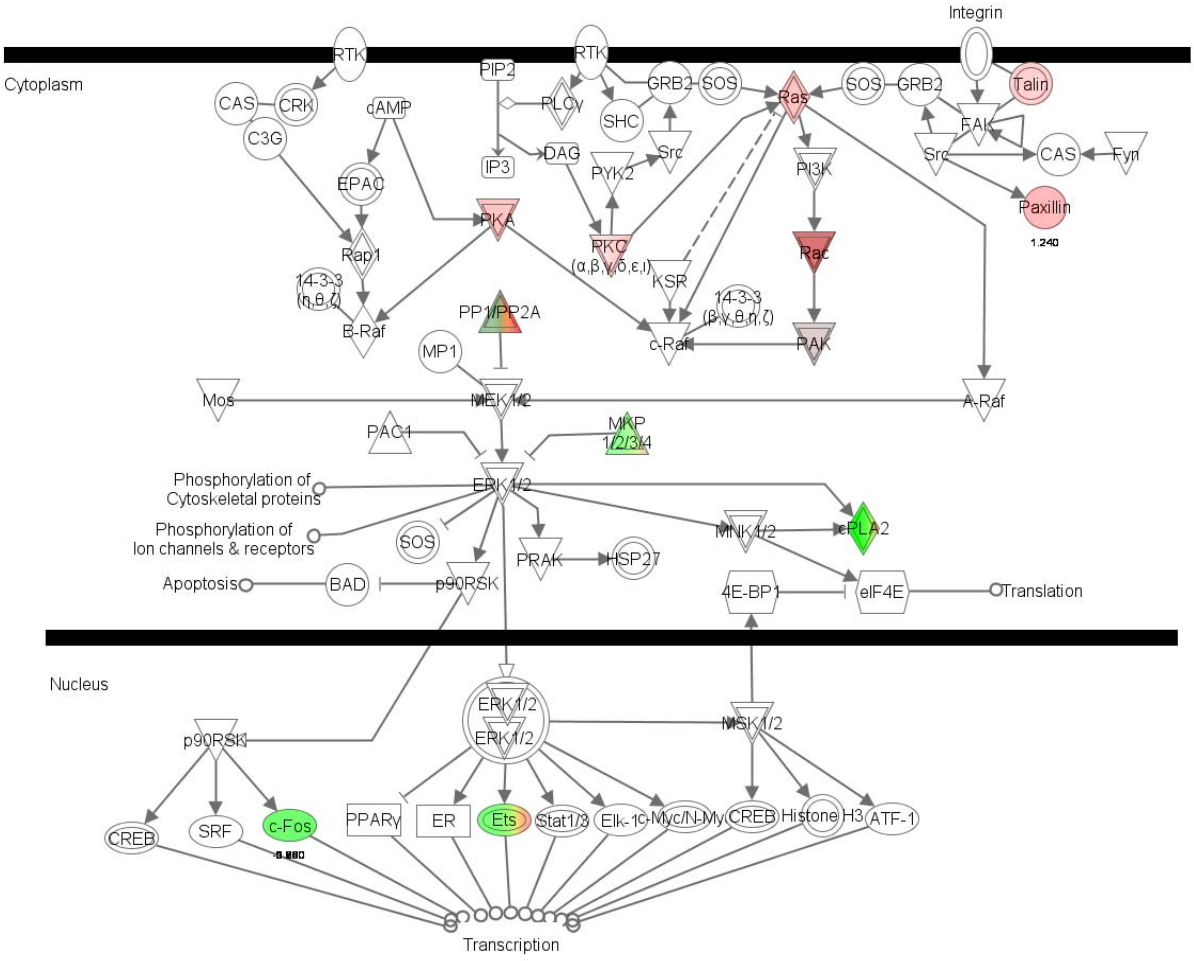


Figure 8-39. ERK/MAPK Signaling Pathway; Combined ChIP target genes and gene expression data in the same ERK/MAPK signaling pathway illustration.

8.6.11 P38/MAPK SIGNALING PATHWAY—CHIP

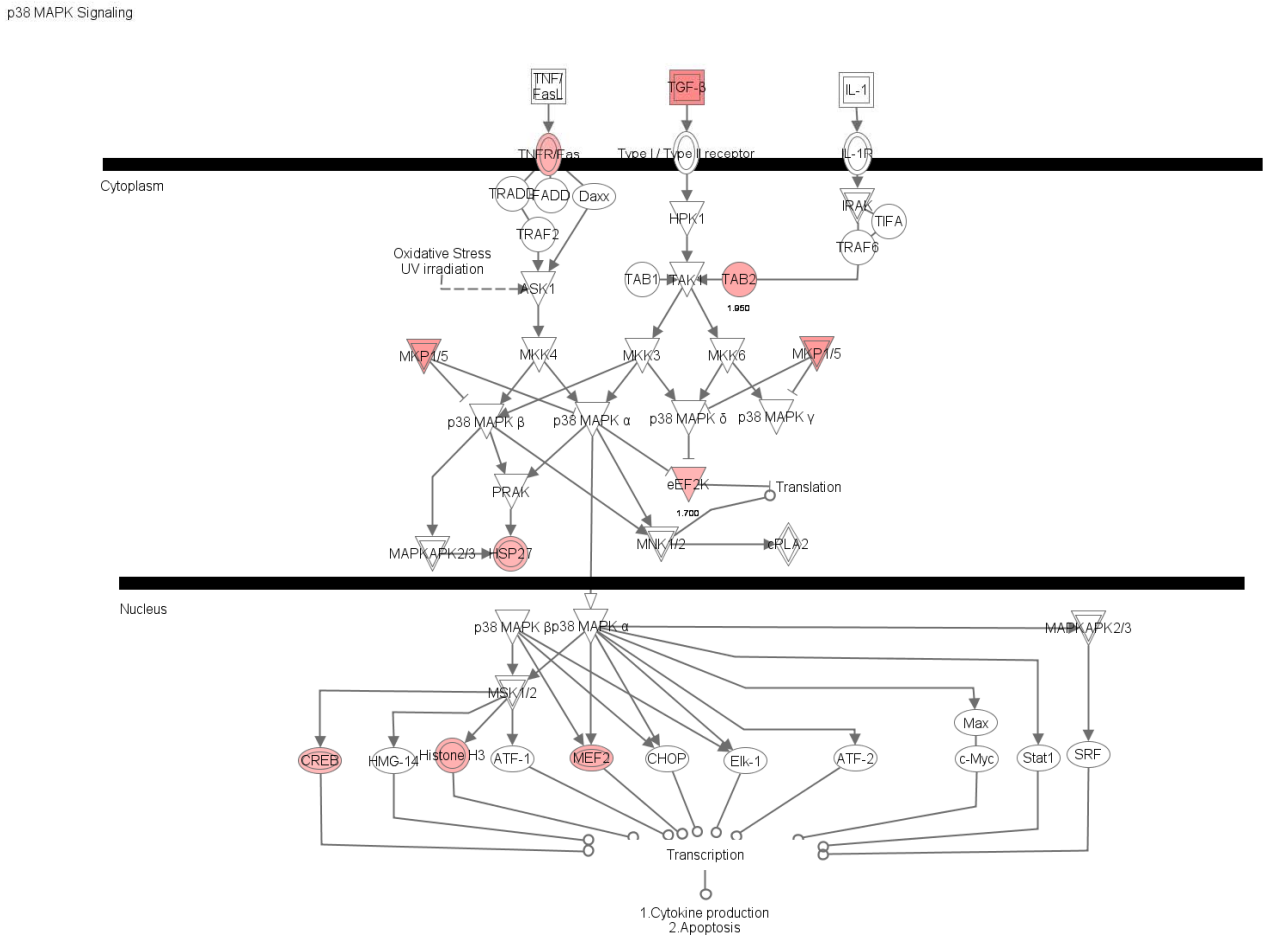


Figure 8-40. ChIP SMAD3-bound target genes illustrated in the p38 MAPK signaling pathway. Red denotes bound target gene; Color intensity depicts ChIP SMAD3 binding peak height.

8.6.12 P38/MAPK SIGNALING PATHWAY—GENE EXPRESSION

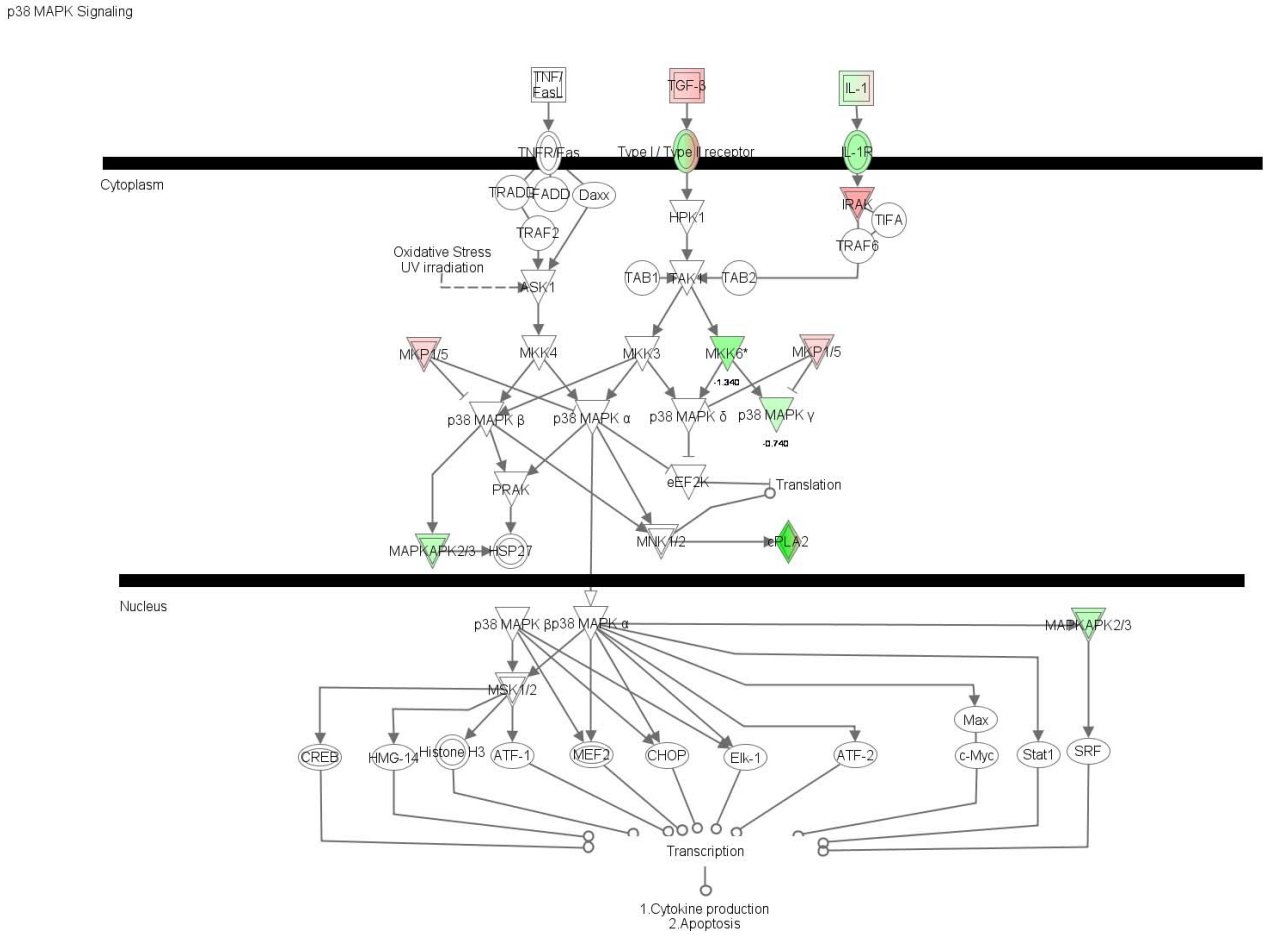


Figure 8-41. Gene expression values illustrated in the p38 MAPK signaling pathway. Significant gene expression profiles after TGFβ_1 stimulation; red denotes up-regulation, green denotes down-regulation. Color intensity depicts relative gene expression levels.

8.6.13 P38/MAPK SIGNALING PATHWAY—COMBINED

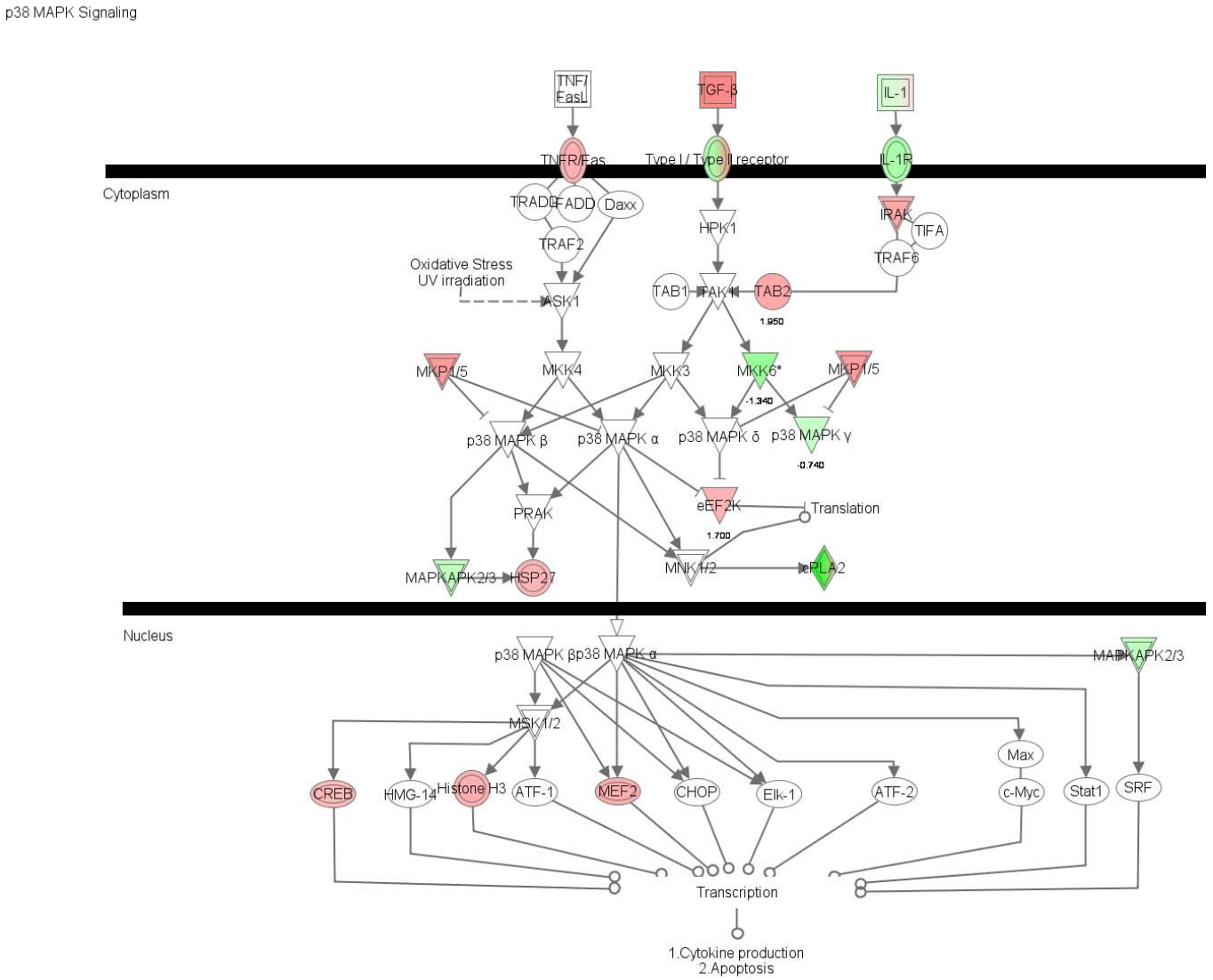


Figure 8-42. Combined ChIP target genes and gene expression data in the same p38 MAPK signaling pathway illustration.

8.6.14 NFκB SIGNALING PATHWAY—CHIP

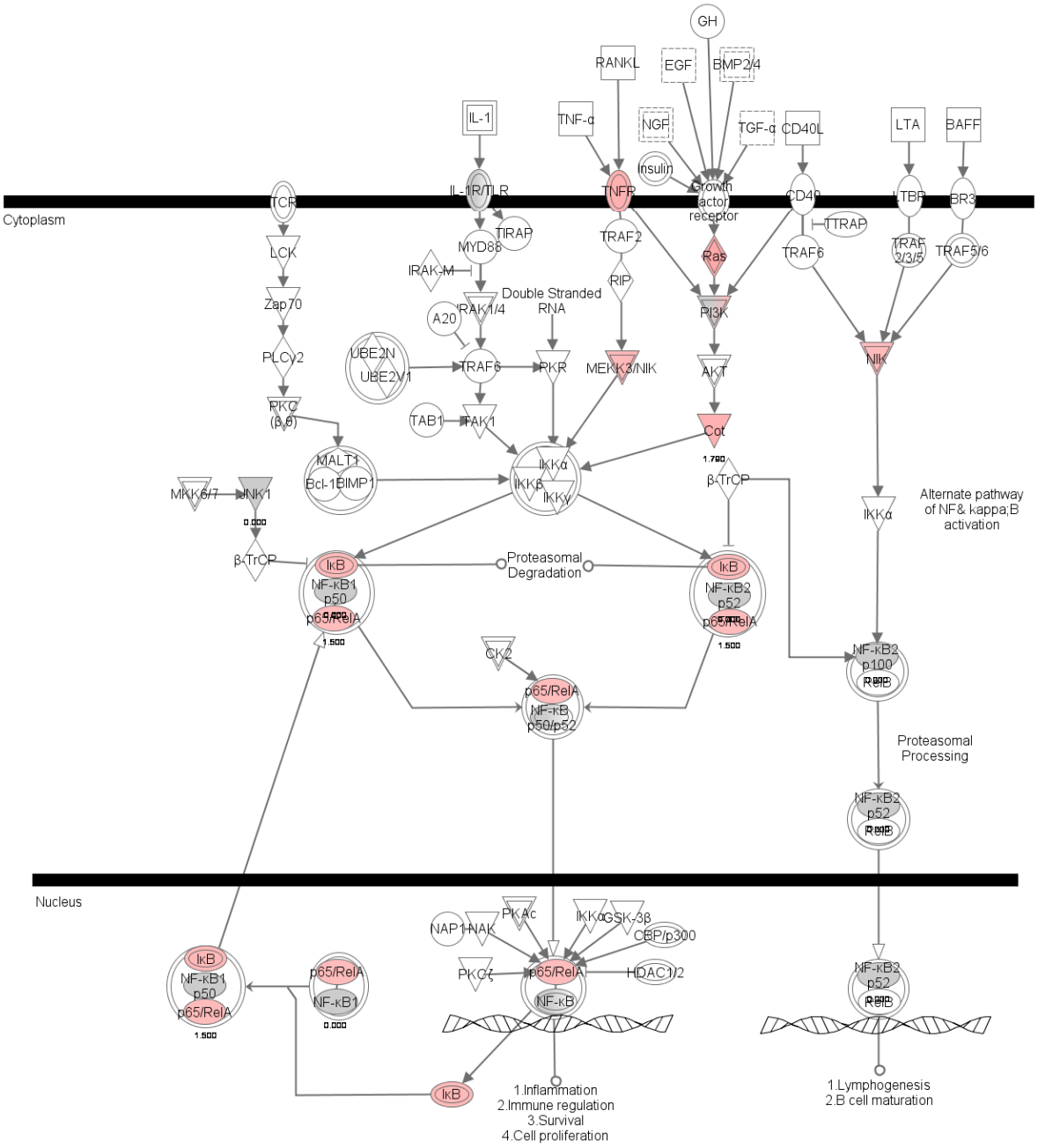


Figure 8-43. ChIP SMAD3-bound target genes illustrated in the NFκB signaling pathway. Red denotes bound target gene; Color intensity depicts ChIP SMAD3 binding peak height.

8.6.15 **NFκB SIGNALING PATHWAY—GENE EXPRESSION**

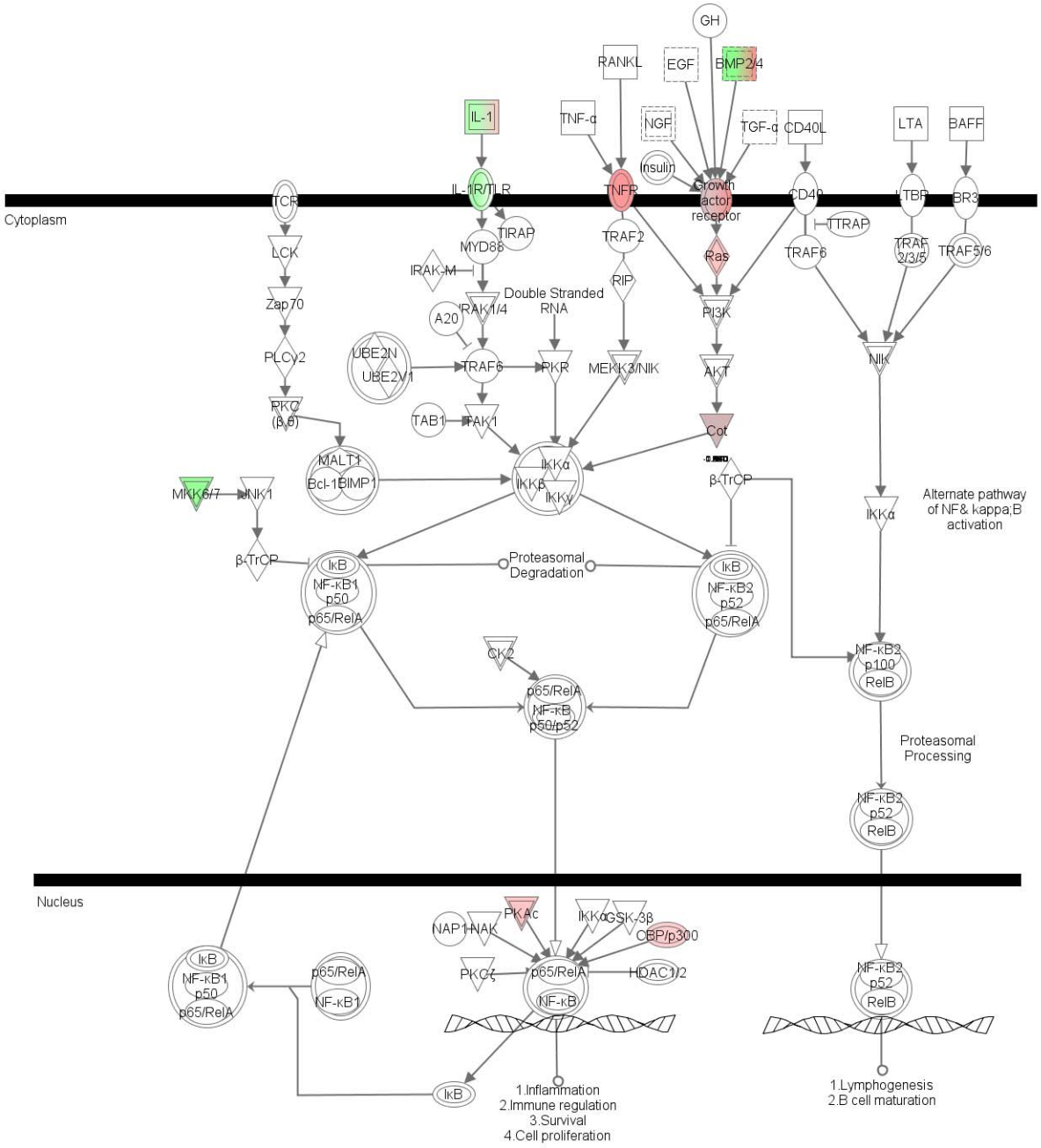


Figure 8-44. Gene expression values illustrated in the ChIP SMAD3-bound target genes illustrated in the NFκB signaling pathway. Significant gene expression profiles after TGFβ_1 stimulation; red denotes up-regulation, green denotes down-regulation. Color intensity depicts relative gene expression levels.

8.6.16 NFκB SIGNALING PATHWAY—COMBINED

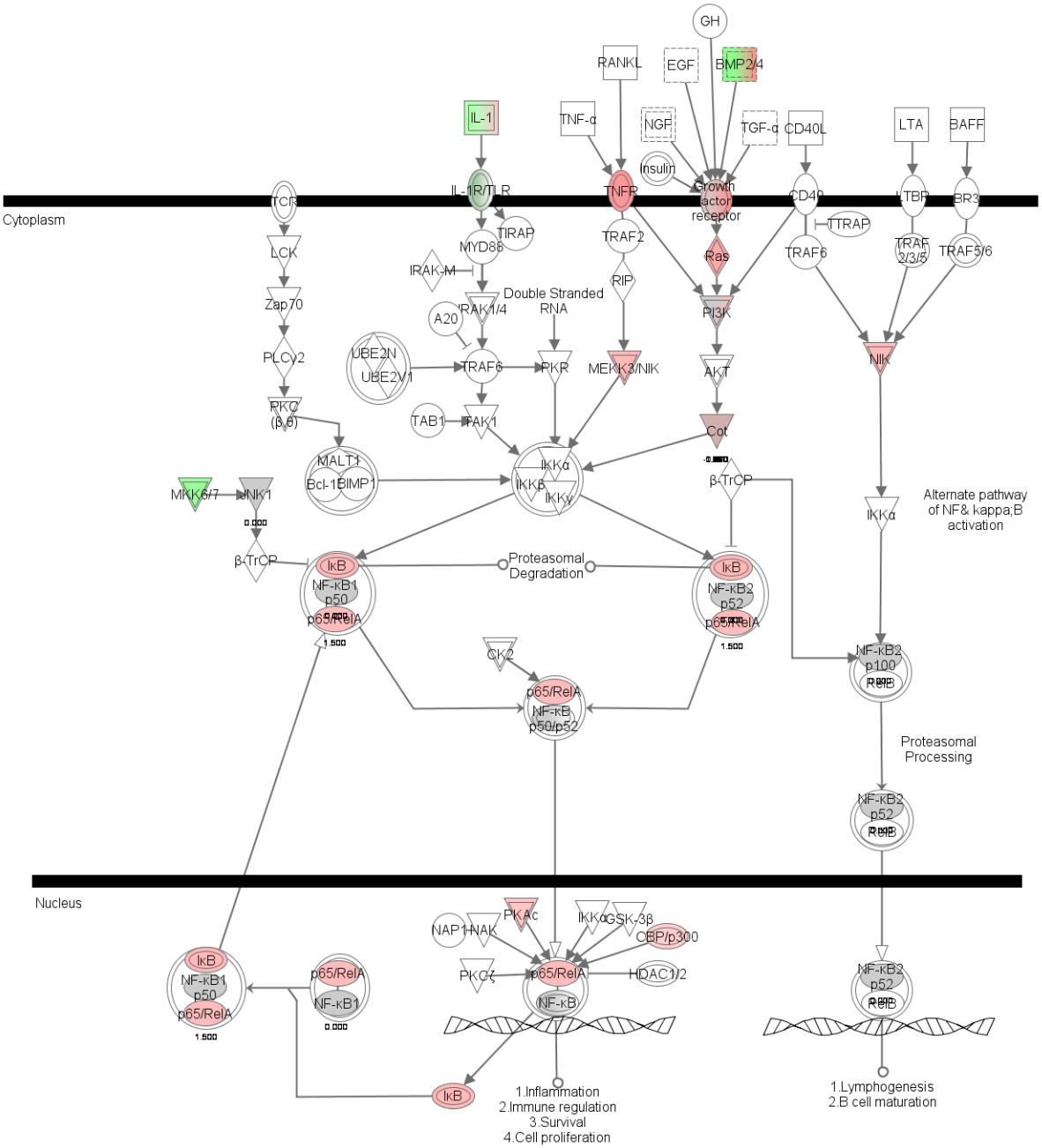


Figure 8-45. Combined ChIP target genes and gene expression data in the same NFκB signaling pathway illustration.

9 DISCUSSION AND CONCLUSIONS

The data presented here suggest three novel different transcriptional regulatory mechanisms by which the $TGF\beta_1$/SMAD3 pathway may directly contribute to the pathogenesis of IPF. As discussed in the introductory chapters, $TGF\beta_1$-induced transition of epithelial cells into myofibroblasts seems to play a major role in fibrotic deposition in the lung. ChIP-on-chip, a high-throughput method for simultaneous identification of transcription factor/promoter binding, led to the specific identification of transgelin (TAGLN) as a mediator of EMT in alveolar Type II epithelial cells. Similarly, ChIP-on-chip identified a novel connection between the TGF $TGF\beta_1$/SMAD3 transcriptional regulatory pathway and FOXA2, a transcription factor vitally necessary to proper lung development and function. Finally, a systems-level analysis of ChIP-on-chip and gene expression data led to the identification of a possible suppressor of telomerase (TERT) in pulmonary epithelial cells, which may be a third mechanism directly connecting the $TGF\beta_1$ cytokine with the pathogenesis of IPF.

9.1 FOXA2 AND SUPPRESSION OF PULMONARY SURFACTANTS

Pulmonary surfactants are vital lipoprotein complexes produced by type II alveolar epithelial cells [295]. Among their functions is to reduce liquid surface tension in the alveolar sacs. This allows the lung to expand fully for maximal gas exchange as well as prevent it from

collapsing at the end of expiration. In addition to these functions, surfactant protein A (Sp-A) and protein D (Sp-D) are involved in host defense again pathogens in the lung . Surfactant proteins B and C (Sp-B, Sp-C) are hydrophobic membrane proteins that enable the other surfactants to more easily spread over the alveolar surface [45, 295].

Sp-A, Sp-B, and Sp-C are each implicated separately in the pathogenesis of IPF through animal models and/or human studies. Genetic variants of Sp-A have been found to strongly increase risk of IPF [85]. Sp-A levels also are found to be increased in newly diagnosed IPF patients [84]. Administration of exogenous Sp-A in cell culture of fibroblasts doubled their collagen expression [86]. Sp-B (*SFTPB* -/-) null animals, and humans born with a congenital absence of Sp-B due to mutations, die shortly after birth from respiratory failure [45, 296-298]. As discussed previously, mutations in *SFTPC* and consequent defective Sp-C proteins have been clearly identified in clusters of familial IPF [41, 42, 44]. Further, in a bleomycin animal model of IPF, Sp-C null mice (*SFTPC* -/-) exhibited a much more severe and longer time course of pulmonary fibrosis than wild type controls (*SFTPC* +/+) [299].

TGFβ_1 is a known regulator of all these surfactant levels and is known to suppress levels of Sp-B and Sp-C specifically through Thyroid Transcription Factor (TTF-1), which is in turn regulated by FOXA2 [300-303]. Previously it was argued that FOXA2 regulates TTF-1 levels and Sp-B/C through protein-protein interactions [300, 301]. However, the current data strongly suggests that SMAD3 binds the promoter of FOXA2 and directly regulates it at the transcriptional level. No statistically significant changes in TTF-1 levels were seen in the current data; however, probes for only one of the two transcript variants are present on the array, and the differences in levels of TTF-1 may not have been significant enough to be detected by the array

measurements. The exact nature of the transcriptional regulatory mechanisms of surfactants through FOXA2 remain to be elucidated and warrant further study.

9.2 PINX1 AND SUPPRESSION OF HTERT

As discussed previously, telomere reverse transcriptase (*hTERT*) is another gene which has a strong association with IPF. Linkage analysis of 46 families with two or more individuals affected with IIPs identified *hTERT* as a candidate gene [51]. Mutation analysis revealed a missense mutation and a frameshift mutation in *hTERT* that co-segregated with IIP in two families [50]. Sequence comparisons of *hTERT* between 44 sporadic IPF cases and probands from 44 unrelated families revealed five other mutations. One family had a mutation in *hTERT* as well. Those individuals with heterozygous mutations in *hTERT* or *hTERC* had shorter telomere lengths compared to controls [50, 51]. Thus, telomerase is another gene with a distinct association with IPF. Since telomeres overall shorten with age, this connection may also provide a clue to the profoundly increased risk of IPF with age [56, 304, 305]. Finally, dyskeratosis congenita is a rare disorder caused mutations in dyskerin (*DSK1*) that results in loss of telomerase activity. Although not the primary effect of the genetic disease, approximately twenty percent of patients develop a form of pulmonary fibrosis that radiographically and clinically resembles that seen in IPF [306, 307].

In bleomycin-induced animal models of pulmonary fibrosis, the activities of *TERT* are restricted to cell type. In a bleomycin rat model, telomerase activity increased specifically in fibroblasts and not α-smooth muscle positive cells, indicating the effect is restricted to proliferating fibroblasts and does not occur in myofibroblasts [308, 309]. In a bleomycin mouse model, *mTERT* mRNA was upregulated after bleomycin-induced injury, and at 72 hours dropped

below controls. Decreased telomerase activity, either as result of bleomycin injury or directly induced, produced increased epithelial cell apoptosis [310]. In most normal cells, significant telomerase activity appears to be present only in injured cells, stem cells, and neoplastic cells [311-313]. Since the bleomycin animal models involve an injury phase followed by a repair phase during which the induced pulmonary fibrosis resolves, it is unclear how well these results can be generalized to human IPF (in which case pulmonary fibrosis is irreversible). However the information does suggest that there is a delicate balance of cell proliferation and cell apoptosis in the various pulmonary cell populations, and this balance in part depends on the activities of *TERT* and its various regulators.

In the present study, network analysis using IPA identified numerous members of the *hTERT* regulatory pathway populated by ChIP-identified promoters and significantly up- or down-regulated genes (see Figures 8-41 through 8-43). Most prominently, SMAD3 ChIP-on-chip identified the promoter for the protein PIN2-interacting protein 1, or PINX1. The binding of SMAD3 to the promoter of PINX1 was still significant without exogenous $TGF\beta_1$ stimulation, but reduced. The expression of genes in the *hTERT* regulatory was all but abolished with SIS3 treatment, strongly suggesting activation of that network is mediated through the $TGF\beta_1$/SMAD3 pathway (see Figure 8-44).

PINX1 is well-established as a potent inhibitor of hTERT activity, and this inhibition is a result of direct interaction with the PINX1 protein with the hTERT enzyme [46-49]. Further, PINX1 upregulation and hTERT inactivation shortened telomeres and promoted apoptosis in an HT1080 fibrosarcoma cell line [49]. This suggests the possibility that the $TGF\beta_1$/SMAD3 pathway may directly inhibit hTERT activity in epithelial cells, contributing to their apoptosis as well as reducing their ability to proliferate in response to injury. Thus, this may be one

mechanism that chronic lung exposure to TGFβ_1 may promote fibroblast proliferation while reducing epithelial cell populations. This is another pathway connection that warrants further research.

9.3 FUTURE DIRECTIONS

The systems biology methods used here demonstrated the utility of the approach in discovering new mechanistic connections, overall patterns, and suggesting new hypotheses. The results presented here are not an endpoint, but rather a starting point for follow-up by traditional reductionist techniques. In particular, the transcriptional regulatory targets of FOXA2 should be explored, particularly in the context of surfactant regulation. The transcriptional regulation of *hTERT* and *PINX1*, and their activities in normal alveolar epithelial cells and fibroblasts, with and with out TGFβ_1 stimulation, is another promising area that warrants a more detailed investigation.

APPENDIX A: MICROARRAY NORMALIZATION R CODE

```
This code was written by Thomas Richards, Ph.D.

###############################################################################
require(tcltk)
require(Hmisc)
require(affy)
assign("DIR",".",.GlobalEnv)
assign("ANNOTfn","./Agilent_HumanGenome.RData",.GlobalEnv)
setwd(DIR)
DIR <- getwd()
source("./MAfns.r")
CTRLlist <- NULL
###############################################################################
DATESTART <- Sys.time()
cDATESTART <- format(DATESTART,"%d%b%YAT%H%M%S")
LISTfn <- paste("List_",cDATESTART,".RData",sep="")
DFfn <- paste("gPS_",cDATESTART,".RData",sep="")
RDfn <- paste("RData","_",cDATESTART,sep="")
ROUT <- file(description=paste("Rout","_",cDATESTART,".log",sep=""),open="w")
sink(ROUT,split=TRUE)
###############################################################################
  ##  Read in data as list:
cat("\n\n")
print("ReadAgilent.r:  Reading Agilent data files...")
.LIST <- AgilentDataFolder2List(FOLDER=DIR)
###############################################################################
  ##  Make data frame from this list:
cat("\n\n")
print("ReadAgilent.r:  Making data frame from list of Agilent data...")
AgilentDF <- AgilentDataList2DF(.LIST)
###############################################################################
##  Load annotation file, from Agilent:
cat("\n\n")
print("ReadAgilent.r:  Loading Agilent annotation file...")
load(ANNOTfn)
###############################################################################
  ##  Merge data and annotations:
AgilentDF <-
merge(x=Agilent_HumanGenome,y=AgilentDF,by.x="Probe_ID",by.y="ProbeName",all.
x=FALSE,all.y=FALSE)
###############################################################################
###################################################################
  ##  Save list and data frame data, in RDfn, if possible.
  ##  RECALL:  DIR was made earlier.
setwd(DIR)
..TRY <- try(system(paste("MKDIR ",RDfn,sep="")))
if (!inherits(..TRY,"try-error")) setwd(RDfn)
save(list=".LIST",file=LISTfn)
```

```
rm(.LIST);gc()
save(list="AgilentDF",file=DFfn)
###########################################################################
  ##  Crucial variable names follow:
nms <- names(AgilentDF)
ANNOTnms <-
c("Probe_ID","Clone_Name","Genbank_Acc","UniGene_ID","EntrezGene_ID","Gene_Sy
mbol___Name","Symbol_",

"Gene_Synonyms","Human_TC","Human_GC","RefSeq_Acc","TC_PubMed_Ref","GO","TGI_
Annotation","Phy_Map",

"Genetic_Marker","Mouse_ortholog","Rat_ortholog","Zebrafish_ortholog","Xenopu
s_ortholog","Cattle_ortholog",

"Elegans_ortholog","Yeast_ortholog","Dog_ortholog","Chicken_ortholog","Featur
eNum","ProbeUID","ControlType",
  "GeneName","SystematicName","Description")
nmsDATA <- setdiff(nms,ANNOTnms)
WHCHcols <- which(nms %in% nmsDATA)
WHCHann <- which(nms %nin% nmsDATA)
###########################################################################
  ##  Limit to probes with Entrez Gene IDs:
AgilentDF <- AgilentDF[which(!is.na(AgilentDF$EntrezGene_ID)),]
###########################################################################
  ##  Convert to log2:
..MTX <- log2(as.matrix(AgilentDF[,WHCHcols]))
###########################################################################
  ##  Normalize by cyclic loess:
cat("\n\n")
print("ReadAgilent.r:  Running cyclic loess...")
####..MTXc <- normalize.loess(mat=..MTX, maxit = 4, epsilon=0.01,log.it =
FALSE, verbose = TRUE, span = 0.4, family.loess = "symmetric")
..MTXc <- myNL(mat=..MTX, maxit = 10, epsilon=0.01,log.it = FALSE, verbose =
TRUE, span = 0.4, family.loess = "symmetric")
AgilentDF[,WHCHcols] <- ..MTXc
rm(..MTX,..MTXc);gc()
###########################################################################
  ##  Combine probes for each ProbeName, one obs per ProbeName:
cat("\n\n")
print("ReadAgilent.r:  Combining probes...")
nms <- sub("Probe_ID","ProbeName",names(AgilentDF))
names(AgilentDF) <- nms
AgilentDF <-
CombineProbes(DF=AgilentDF,VAR="ProbeName",AnnotColumns=WHCHann,method="mean"
)
save(list="AgilentDF",file=paste("gPSntrz2clmnPN_",cDATESTART,".RData",sep=""
))
write.table(x=AgilentDF[,c("ProbeName","Description",nmsDATA)],

file=paste("gPSntrz2clmnPN_",cDATESTART,".tab",sep=""),sep="\t",row=FALSE,col
=TRUE,na="",quote=FALSE)
AgilentDFntrz2clmnPN <- AgilentDF
###########################################################################
  ##  Geometric mean normalize, over rows:
cat("\n\n")
```

```
print("ReadAgilent.r:  Applying Geometric Mean normalization, over all
samples in experiment...")
.MN <- apply(as.matrix(AgilentDF[,WHCHcols]),1,function(W)
mean(W,na.rm=TRUE))
for(J in WHCHcols) AgilentDF[,J] <- AgilentDF[,J] - .MN
.MN <- apply(as.matrix(AgilentDF[,WHCHcols]),1,function(W)
mean(W,na.rm=TRUE))
cat("\n\n")
print("ReadAgilent.r:  Summary, after Geometric mean normalization, over all
samples in experiment:")
print(summary(.MN))
###############################################################################
save(list="AgilentDF",file=paste("gPSntrz2clmnPNGM_",cDATESTART,".RData",sep=
""))
write.table(x=AgilentDF[,c("ProbeName","Description",nmsDATA)],

file=paste("gPSntrz2clmnPNGM_",cDATESTART,".tab",sep=""),sep="\t",row=FALSE,c
ol=TRUE,na="",quote=FALSE)
###############################################################################
setwd(DIR)
CTRLin <- try(file("Controls.tab","r"))
if (!inherits(CTRLin,"try-error")) CTRLlist <-
gsub("\\.txt","",gsub('\"',"",readLines(CTRLin)))
if (!is.null(CTRLlist)){
  cat("\n\n")
  print("ReadAgilent.r:  Applying Geometric mean normalization, over
controls...")
  WHCHctrl <- which(names(AgilentDFntrz2clmnPN) %in% CTRLlist)
  .MN <- apply(as.matrix(AgilentDFntrz2clmnPN[,WHCHctrl]),1,function(W)
mean(W,na.rm=TRUE))
  for(J in WHCHcols) AgilentDFntrz2clmnPN[,J] <- AgilentDFntrz2clmnPN[,J] -
.MN
  .MN <- apply(as.matrix(AgilentDFntrz2clmnPN[,WHCHctrl]),1,function(W)
mean(W,na.rm=TRUE))
  cat("\n\n")
  print("ReadAgilent.r:  Summary, after Geometric mean normalization, over
controls listed in Controls.tab:")
  print(summary(.MN))
  if (!inherits(..TRY,"try-error")) setwd(RDfn)

save(list="AgilentDFntrz2clmnPN",file=paste("gPSntrz2clmnPNGMctrl_",cDATESTAR
T,".RData",sep=""))
  write.table(x=AgilentDFntrz2clmnPN[,c("ProbeName","Description",nmsDATA)],

file=paste("gPSntrz2clmnPNGMCtrl_",cDATESTART,".tab",sep=""),sep="\t",row=FAL
SE,col=TRUE,na="",quote=FALSE)
  setwd(DIR)
}
###############################################################################
sink()
rm(AgilentDF,AgilentDFntrz2clmnPN,.MN,nms,ANNOTnms,nmsDATA,DIR,ANNOTfn,DATEST
ART,cDATESTART,
  RDfn,DFfn,ROUT,CTRLin,J,LISTfn,WHCHann,WHCHcols,WHCHctrl);gc()
```

APPENDIX B: CHIP-ON-CHIP RESULTS

ChIP-on-chip Significant Bound Genes, TGFβ-stimulated cells

GENE ID	PEAK HEIGHT	UNIGENE ID	ENTREZ ID	GENE DESCRIPTION
SERPINE1	10.40	Hs.414795	5054	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
FLJ45248	7.67	Hs.224506	401472	FLJ45248 protein
CCDC129	7.50	Hs.224269	223075	Coiled-coil domain containing 129
CLC	6.62	Hs.889	1178	Charcot-Leyden crystal protein
PPP1R13L	6.08	Hs.466937	10848	Protein phosphatase 1, regulatory (inhibitor) subunit 13 like
COL7A1	4.63	Hs.476218	1294	Collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive)
SMAD6	4.63	Hs.153863	4091	SMAD family member 6
STXBP1	4.20	Hs.288229	6812	Syntaxin binding protein 1
FAM129B	4.20	Hs.522401	64855	Family with sequence similarity 129, member B
FGB	4.10	Hs.300774	2244	Fibrinogen beta chain
TINAGL1	4.00	Hs.199368	64129	Tubulointerstitial nephritis antigen-like 1
POLD4	3.98	Hs.523829	57804	Polymerase (DNA-directed), delta 4
SMAD7	3.91	Hs.465087	4092	SMAD family member 7
BCL9L	3.84	Hs.414740	283149	B-cell CLL/lymphoma 9-like
GRIN2D	3.76	Hs.445015	2906	Glutamate receptor, ionotropic, N-methyl D-aspartate 2D
KDELR1	3.76	Hs.515515	10945	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 1
TAGLN2	3.72	Hs.517168	8407	Transgelin 2
MRC2	3.72	Hs.7835	9902	Mannose receptor, C type 2
RUSC2	3.70	Hs.493796	9853	RUN and SH3 domain containing 2
MUC1	3.68	Hs.89603	4582	Mucin 1, cell surface associated
DOCK7	3.60	Hs.538059	85440	Dedicator of cytokinesis 7
SRXN1	3.51	Hs.355284	140809	Sulfiredoxin 1 homolog (S. cerevisiae)
SNX12	3.34	Hs.260750	29934	Sorting nexin 12
S100A2	3.31	Hs.516484	6273	S100 calcium binding protein A2
RRAS	3.30	Hs.515536	6237	Related RAS viral (r-ras) oncogene homolog
SNCG	3.30	Hs.349470	6623	Synuclein, gamma (breast cancer-specific protein 1)
BLOC1S2	3.30	Hs.702055	282991	Biogenesis of lysosome-related organelles complex-1, subunit 2
PINX1	3.27	Hs.490991	54984	PIN2-interacting protein 1
MYO1D	3.23	Hs.658000	4642	Myosin ID
FLJ10357	3.20	Hs.35125	55701	Hypothetical protein FLJ10357
TRIB1	3.17	Hs.444947	10221	Tribbles homolog 1 (Drosophila)
C21orf84	3.04	Hs.592161	114038	Chromosome 21 open reading frame 84
GIPC1	3.03	Hs.655012	10755	GIPC PDZ domain containing family, member 1

ACLY	3.01	Hs.387567	47	ATP citrate lyase
ENO3	2.99	Hs.224171	2027	Enolase 3 (beta, muscle)
PFN1	2.99	Hs.494691	5216	Profilin 1
IFI44	2.97	Hs.82316	10561	Interferon-induced protein 44
VASP	2.95	Hs.702197	7408	Vasodilator-stimulated phosphoprotein
DC2	2.95	Hs.445803	58505	DC2 protein
ITGA3	2.93	Hs.265829	3675	Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)
PALLD	2.92	Hs.151220	23022	Palladin, cytoskeletal associated protein
UPP1	2.90	Hs.488240	7378	Uridine phosphorylase 1
HNRPUL1	2.90	Hs.699274	11100	Heterogeneous nuclear ribonucleoprotein U-like 1
ACYP2	2.89	Hs.516173	98	Acylphosphatase 2, muscle type
NXNL2	2.83	Hs.668937	158046	Nucleoredoxin-like 2
EPS8	2.78	Hs.591160	2059	Epidermal growth factor receptor pathway substrate 8
RCOR3	2.77	Hs.696152	55758	REST corepressor 3
CS	2.75	Hs.430606	1431	Citrate synthase
C11orf1	2.74	Hs.17546	64776	Chromosome 11 open reading frame 1
ALG9	2.74	Hs.503850	79796	Asparagine-linked glycosylation 9 homolog (S. cerevisiae, alpha- 1,2-mannosyltransferase)
ZYX	2.71	Hs.490415	7791	Zyxin
HERPUD2	2.70	Hs.599851	64224	HERPUD family member 2
IL31RA	2.70	Hs.55378	133396	Interleukin 31 receptor A
TGFB1	2.69	Hs.645227	7040	Transforming growth factor, beta 1
SLC20A1	2.68	Hs.187946	6574	Solute carrier family 20 (phosphate transporter), member 1
QPCT	2.67	Hs.79033	25797	Glutaminyl-peptide cyclotransferase (glutaminyl cyclase)
PRKAB2	2.66	Hs.50732	5565	Protein kinase, AMP-activated, beta 2 non-catalytic subunit
C14orf79	2.66	Hs.27183	122616	Chromosome 14 open reading frame 79
CALM2	2.62	Hs.643483	805	Calmodulin 2 (phosphorylase kinase, delta)
FOXA2	2.62	Hs.155651	3170	Forkhead box A2
TAGLN	2.62	Hs.632099	6876	Transgelin
C14orf4	2.62	Hs.179260	64207	Chromosome 14 open reading frame 4
NAB2	2.60	Hs.159223	4665	NGFI-A binding protein 2 (EGR1 binding protein 2)
TMEM194	2.60	Hs.591040	23306	Transmembrane protein 194
ATAD2	2.60	Hs.370834	29028	ATPase family, AAA domain containing 2
C14orf43	2.60	Hs.656506	91748	Chromosome 14 open reading frame 43
IL11	2.59	Hs.467304	3589	Interleukin 11
TMEM190	2.59	Hs.590943	147744	Transmembrane protein 190
SMAD3	2.58	Hs.618504	4088	SMAD family member 3
USP3	2.56	Hs.458499	9960	Ubiquitin specific peptidase 3
WDR55	2.55	Hs.286261	54853	WD repeat domain 55
SH2D4A	2.54	Hs.303208	63898	SH2 domain containing 4A
KIF21A	2.53	Hs.374201	55605	Kinesin family member 21A
ARF6	2.52	Hs.525330	382	ADP-ribosylation factor 6
BAIAP2L1	2.52	Hs.656063	55971	BAI1-associated protein 2-like 1

PDE7B	2.51	Hs.652367	27115	Phosphodiesterase 7B
GTF2B	2.50	Hs.481852	2959	General transcription factor IIB
ITGB1	2.50	Hs.695946	3688	Integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)
GCC2	2.48	Hs.705434	9648	GRIP and coiled-coil domain containing 2
ZCCHC11	2.48	Hs.655407	23318	Zinc finger, CCHC domain containing 11
SPRED1	2.48	Hs.525781	161742	Sprouty-related, EVH1 domain containing 1
ANXA2	2.47	Hs.511605	302	Annexin A2
PLAUR	2.47	Hs.466871	5329	Plasminogen activator, urokinase receptor
PLS3	2.47	Hs.496622	5358	Plastin 3 (T isoform)
SEMA4B	2.47	Hs.474935	10509	Sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4B
ZNF219	2.46	Hs.250493	51222	Zinc finger protein 219
C14orf104	2.44	Hs.231761	55172	Chromosome 14 open reading frame 104
C14orf24	2.44	Hs.446357	283635	Chromosome 14 open reading frame 24
KPNA1	2.41	Hs.161008	3836	Karyopherin alpha 1 (importin alpha 5)
SOX1	2.40	Hs.202526	6656	SRY (sex determining region Y)-box 1
SULT2B1	2.40	Hs.369331	6820	Sulfotransferase family, cytosolic, 2B, member 1
TRIM16	2.36	Hs.123534	10626	Tripartite motif-containing 16
CA3	2.35	Hs.82129	761	Carbonic anhydrase III, muscle specific
CSN2	2.35	Hs.2242	1447	Casein beta
DDX5	2.35	Hs.279806	1655	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5
CCDC45	2.35	Hs.569713	90799	Coiled-coil domain containing 45
PVR	2.33	Hs.171844	5817	Poliovirus receptor
XRCC2	2.33	Hs.647093	7516	X-ray repair complementing defective repair in Chinese hamster cells 2
MEX3B	2.33	Hs.104744	84206	Mex-3 homolog B (C. elegans)
UTRN	2.32	Hs.133135	7402	Utrophin
DHRS12	2.32	Hs.266728	79758	Dehydrogenase/reductase (SDR family) member 12
JMJD2A	2.31	Hs.155983	9682	Jumonji domain containing 2A
ZBTB45	2.31	Hs.515662	84878	Zinc finger and BTB domain containing 45
ART1	2.29	Hs.382188	417	ADP-ribosyltransferase 1
PC	2.29	Hs.89890	5091	Pyruvate carboxylase
LRCH1	2.29	Hs.656722	23143	Leucine-rich repeats and calponin homology (CH) domain containing 1
ART5	2.29	Hs.125680	116969	ADP-ribosyltransferase 5
ACTN1	2.28	Hs.509765	87	Actinin, alpha 1
DHRS7B	2.28	Hs.386989	25979	Dehydrogenase/reductase (SDR family) member 7B
CBX8	2.28	Hs.387258	57332	Chromobox homolog 8 (Pc class homolog, Drosophila)
ATF3	2.26	Hs.460	467	Activating transcription factor 3
DUSP1	2.26	Hs.171695	1843	Dual specificity phosphatase 1
SOCS5	2.26	Hs.468426	9655	Suppressor of cytokine signaling 5
POLE3	2.26	Hs.108112	54107	Polymerase (DNA directed), epsilon 3 (p17 subunit)
KCTD15	2.26	Hs.221873	79047	Potassium channel tetramerisation domain containing 15
C9orf43	2.26	Hs.632691	257169	Chromosome 9 open reading frame 43
PRCC	2.25	Hs.516948	5546	Papillary renal cell carcinoma (translocation-associated)

DYNLT1	2.25	Hs.445999	6993	Dynein, light chain, Tctex-type 1
ANKRD25	2.25	Hs.284208	25959	Ankyrin repeat domain 25
CSDC2	2.25	Hs.310893	27254	Cold shock domain containing C2, RNA binding
RHPN2	2.25	Hs.466435	85415	Rhophilin, Rho GTPase binding protein 2
SYTL3	2.25	Hs.436977	94120	Synaptotagmin-like 3
LDHA	2.24	Hs.2795	3939	Lactate dehydrogenase A
R3HDML	2.24	Hs.580807	140902	R3H domain containing-like
RTP3	2.23	Hs.196584	83597	Receptor (chemosensory) transporter protein 3
NANP	2.23	Hs.666255	140838	N-acetylneuraminic acid phosphatase
PRPS1	2.22	Hs.56	5631	Phosphoribosyl pyrophosphate synthetase 1
DNAJB2	2.21	Hs.77768	3300	DnaJ (Hsp40) homolog, subfamily B, member 2
RGS13	2.21	Hs.497220	6003	Regulator of G-protein signaling 13
FILIP1	2.21	Hs.696158	27145	Filamin A interacting protein 1
RBM25	2.21	Hs.531106	58517	RNA binding motif protein 25
BEST3	2.21	Hs.280782	144453	Bestrophin 3
PLK3	2.20	Hs.632415	1263	Polo-like kinase 3 (Drosophila)
UVRAG	2.20	Hs.202470	7405	UV radiation resistance associated gene
FADS2	2.20	Hs.502745	9415	Fatty acid desaturase 2
NUTF2	2.20	Hs.696342	10204	Nuclear transport factor 2
GTF3C2	2.19	Hs.75782	2976	General transcription factor IIIC, polypeptide 2, beta 110kDa
METTL2A	2.19	Hs.381204	339175	Methyltransferase like 2A
ZBTB25	2.18	Hs.654571	7597	Zinc finger and BTB domain containing 25
ZBTB1	2.18	Hs.655536	22890	Zinc finger and BTB domain containing 1
DNAJB11	2.18	Hs.317192	51726	DnaJ (Hsp40) homolog, subfamily B, member 11
DNAJB11	2.18	Hs.317192	51726	DnaJ (Hsp40) homolog, subfamily B, member 11
TBCCD1	2.18	Hs.518469	55171	TBCC domain containing 1
JMJD1A	2.18	Hs.557425	55818	Jumonji domain containing 1A
CD63	2.17	Hs.445570	967	CD63 molecule
TSKU	2.17	Hs.8361	25987	Tsukushin
CYHR1	2.17	Hs.459379	50626	Cysteine/histidine-rich 1
KIFC2	2.17	Hs.528713	90990	Kinesin family member C2
LENG8	2.17	Hs.502378	114823	Leukocyte receptor cluster (LRC) member 8
GPR92	2.16	Hs.155538	57121	G protein-coupled receptor 92
HTR1F	2.15	Hs.248136	3355	5-hydroxytryptamine (serotonin) receptor 1F
PHEX	2.15	Hs.495834	5251	Phosphate regulating endopeptidase homolog, X-linked (hypophosphatemia, vitamin D resistant rickets)
ODZ1	2.15	Hs.23796	10178	Odz, odd Oz/ten-m homolog 1(Drosophila)
KLF13	2.15	Hs.525752	51621	Kruppel-like factor 13
LTBP3	2.14	Hs.289019	4054	Latent transforming growth factor beta binding protein 3
P2RY2	2.14	Hs.339	5029	Purinergic receptor P2Y, G-protein coupled, 2
RDH13	2.14	Hs.327631	112724	Retinol dehydrogenase 13 (all-trans/9-cis)
CCND1	2.13	Hs.523852	595	Cyclin D1
NQO1	2.12	Hs.406515	1728	NAD(P)H dehydrogenase, quinone 1

LIF	2.12	Hs.2250	3976	Leukemia inhibitory factor (cholinergic differentiation factor)
NOTCH2	2.12	Hs.487360	4853	Notch homolog 2 (Drosophila)
MEP1A	2.11	Hs.179704	4224	Meprin A, alpha (PABA peptide hydrolase)
RSF1	2.11	Hs.420229	51773	Remodeling and spacing factor 1
FLJ45256	2.11	Hs.592028	400511	Hypothetical LOC400511
GSS	2.10	Hs.82327	2937	Glutathione synthetase
NME1	2.09	Hs.463456	4830	Non-metastatic cells 1, protein (NM23A) expressed in
NPY1R	2.09	Hs.519057	4886	Neuropeptide Y receptor Y1
ZNF384	2.09	Hs.103315	171017	Zinc finger protein 384
ALDH3A2	2.08	Hs.499886	224	Aldehyde dehydrogenase 3 family, member A2
CER1	2.08	Hs.248204	9350	Cerberus 1, cysteine knot superfamily, homolog (Xenopus laevis)
NUBP2	2.08	Hs.256549	10101	Nucleotide binding protein 2 (MinD homolog, E. coli)
SPSB3	2.08	Hs.592080	90864	SpIA/ryanodine receptor domain and SOCS box containing 3
SSX6	2.08	Hs.511998	280657	Synovial sarcoma, X breakpoint 6
CXADR	2.07	Hs.705503	1525	Coxsackie virus and adenovirus receptor
EDG6	2.07	Hs.662006	8698	Endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 6
SPECC1	2.07	Hs.431045	92521	Sperm antigen with calponin homology and coiled-coil domains 1
TIPARP	2.06	Hs.12813	25976	TCDD-inducible poly(ADP-ribose) polymerase
CCDC123	2.06	Hs.599703	84902	Coiled-coil domain containing 123
ITGB6	2.05	Hs.470399	3694	Integrin, beta 6
RNPS1	2.05	Hs.355643	10921	RNA binding protein S1, serine-rich domain
USP32	2.05	Hs.132868	84669	Ubiquitin specific peptidase 32
LOC338799	2.05	Hs.654994	338799	Hypothetical locus LOC338799
SEPT7	2.04	Hs.191346	989	Septin 7
ITGAV	2.04	Hs.436873	3685	Integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)
PYGL	2.03	Hs.282417	5836	Phosphorylase, glycogen; liver (Hers disease, glycogen storage disease type VI)
SSR3	2.03	Hs.518346	6747	Signal sequence receptor, gamma (translocon-associated protein gamma)
DYNLL1	2.03	Hs.5120	8655	Dynein, light chain, LC8-type 1
SPINT2	2.03	Hs.31439	10653	Serine peptidase inhibitor, Kunitz type, 2
C19orf33	2.03	Hs.631544	64073	Chromosome 19 open reading frame 33
PPP1R14A	2.03	Hs.631569	94274	Protein phosphatase 1, regulatory (inhibitor) subunit 14A
AVPI1	2.02	Hs.23918	60370	Arginine vasopressin-induced 1
TNMD	2.02	Hs.132957	64102	Tenomodulin
ADM2	2.02	Hs.449099	79924	Adrenomedullin 2
SHOC2	2.01	Hs.104315	8036	Soc-2 suppressor of clear homolog (C. elegans)
BRWD1	2.01	Hs.654740	54014	Bromodomain and WD repeat domain containing 1
BICD1	2.00	Hs.505202	636	Bicaudal D homolog 1 (Drosophila)
DBT	2.00	Hs.633217	1629	Dihydrolipoamide branched chain transacylase E2
DLX4	2.00	Hs.591167	1748	Distal-less homeobox 4
KLF10	2.00	Hs.435001	7071	Kruppel-like factor 10
ETHE1	2.00	Hs.7486	23474	Ethylmalonic encephalopathy 1
RNF19A	2.00	Hs.292882	25897	Ring finger protein 19A

GEMIN8	2.00	Hs.592237	54960	Gem (nuclear organelle) associated protein 8
B4GALNT2	2.00	Hs.374679	124872	Beta-1,4-N-acetyl-galactosaminyl transferase 2
ZNF575	2.00	Hs.213534	284346	Zinc finger protein 575
NOTCH2NL	2.00	Hs.655156	388677	Notch homolog 2 (Drosophila) N-terminal like
NR3C2	1.99	Hs.163924	4306	Nuclear receptor subfamily 3, group C, member 2
SLC16A1	1.99	Hs.75231	6566	Solute carrier family 16, member 1 (monocarboxylic acid transporter 1)
JPH2	1.99	Hs.441737	57158	Junctophilin 2
CDK7	1.98	Hs.184298	1022	Cyclin-dependent kinase 7
RARA	1.98	Hs.654583	5914	Retinoic acid receptor, alpha
GNL2	1.98	Hs.75528	29889	Guanine nucleotide binding protein-like 2 (nucleolar)
MLL5	1.98	Hs.592262	55904	Myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, Drosophila)
DAND5	1.98	Hs.331981	199699	DAN domain family, member 5
DAB2	1.97	Hs.481980	1601	Disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)
PHKA2	1.97	Hs.54941	5256	Phosphorylase kinase, alpha 2 (liver)
PNMA1	1.97	Hs.194709	9240	Paraneoplastic antigen MA1
LDOC1	1.97	Hs.45231	23641	Leucine zipper, down-regulated in cancer 1
BRF2	1.97	Hs.705411	55290	BRF2, subunit of RNA polymerase III transcription initiation factor, BRF1-like
KLC1	1.96	Hs.20107	3831	Kinesin light chain 1
RPH3AL	1.96	Hs.651925	9501	Rabphilin 3A-like (without C2 domains)
MTP18	1.96	Hs.656909	51537	Mitochondrial protein 18 kDa
MPP5	1.96	Hs.652312	64398	Membrane protein, palmitoylated 5 (MAGUK p55 subfamily member 5)
E2F4	1.95	Hs.108371	1874	E2F transcription factor 4, p107/p130-binding
IL6R	1.95	Hs.695954	3570	Interleukin 6 receptor
PFTK1	1.95	Hs.430742	5218	PFTAIRE protein kinase 1
HLTF	1.95	Hs.3068	6596	Helicase-like transcription factor
MAP3K7IP2	1.95	Hs.269775	23118	Mitogen-activated protein kinase kinase kinase 7 interacting protein 2
TPTE2	1.95	Hs.377488	93492	Transmembrane phosphoinositide 3-phosphatase and tensin homolog 2
FAM33A	1.95	Hs.463607	348235	Family with sequence similarity 33, member A
HIBCH	1.94	Hs.656685	26275	3-hydroxyisobutyryl-Coenzyme A hydrolase
AXL	1.93	Hs.590970	558	AXL receptor tyrosine kinase
ARL4C	1.93	Hs.699342	10123	ADP-ribosylation factor-like 4C
KIAA1161	1.93	Hs.522083	57462	KIAA1161
SLC25A28	1.93	Hs.403790	81894	Solute carrier family 25, member 28
KIAA1737	1.93	Hs.22452	85457	KIAA1737
CTPS	1.92	Hs.473087	1503	CTP synthase
DBP	1.92	Hs.414480	1628	D site of albumin promoter (albumin D-box) binding protein
C20orf29	1.92	Hs.104806	55317	Chromosome 20 open reading frame 29
KIAA1609	1.92	Hs.288274	57707	KIAA1609
EPB41L1	1.91	Hs.437422	2036	Erythrocyte membrane protein band 4.1-like 1
LBR	1.91	Hs.435166	3930	Lamin B receptor
MDFI	1.91	Hs.520119	4188	MyoD family inhibitor
SPIN1	1.91	Hs.146804	10927	Spindlin 1

TRIM31	1.91	Hs.493275	11074	Tripartite motif-containing 31
DDT	1.90	Hs.656723	1652	D-dopachrome tautomerase
GSTT2	1.90	Hs.654462	2953	Glutathione S-transferase theta 2
JARID1A	1.90	Hs.654806	5927	Jumonji, AT rich interactive domain 1A
PLA2G7	1.90	Hs.584823	7941	Phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma)
SLC7A5	1.90	Hs.513797	8140	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5
PGS1	1.90	Hs.654671	9489	Phosphatidylglycerophosphate synthase 1
ARHGAP11A	1.90	Hs.591130	9824	Rho GTPase activating protein 11A
SH3TC1	1.90	Hs.479116	54436	SH3 domain and tetratricopeptide repeats 1
MKI67IP	1.90	Hs.367842	84365	MKI67 (FHA domain) interacting nucleolar phosphoprotein
JUB	1.89	Hs.655832	84962	Jub, ajuba homolog (Xenopus laevis)
ARHGAP5	1.88	Hs.592313	394	Rho GTPase activating protein 5
CHRNB2	1.88	Hs.2306	1141	Cholinergic receptor, nicotinic, beta 2 (neuronal)
HNRPK	1.88	Hs.695973	3190	Heterogeneous nuclear ribonucleoprotein K
CCDC9	1.88	Hs.227782	26093	Coiled-coil domain containing 9
TMEM132A	1.88	Hs.118552	54972	Transmembrane protein 132A
UBE2Q1	1.88	Hs.607928	55585	Ubiquitin-conjugating enzyme E2Q (putative) 1
MUS81	1.88	Hs.288798	80198	MUS81 endonuclease homolog (S. cerevisiae)
MGC34761	1.88	Hs.556045	283971	Hypothetical protein MGC34761
SH3BGR	1.87	Hs.473847	6450	SH3 domain binding glutamic acid-rich protein
KIAA0174	1.87	Hs.232194	9798	KIAA0174
DSTN	1.87	Hs.304192	11034	Destrin (actin depolymerizing factor)
THAP8	1.87	Hs.350209	199745	THAP domain containing 8
WDR62	1.87	Hs.116244	284403	WD repeat domain 62
PNRC1	1.86	Hs.75969	10957	Proline-rich nuclear receptor coactivator 1
NHS	1.85	Hs.201623	4810	Nance-Horan syndrome (congenital cataracts and dental anomalies)
STAT3	1.85	Hs.463059	6774	Signal transducer and activator of transcription 3 (acute-phase response factor)
TXN	1.85	Hs.435136	7295	Thioredoxin
CATSPER2	1.85	Hs.662284	117155	Cation channel, sperm associated 2
HSPA2	1.84	Hs.432648	3306	Heat shock 70kDa protein 2
PHLDA2	1.84	Hs.154036	7262	Pleckstrin homology-like domain, family A, member 2
KIAA1576	1.84	Hs.461405	57687	KIAA1576 protein
KLC4	1.84	Hs.655123	89953	Kinesin light chain 4
ADPRHL1	1.84	Hs.98669	113622	ADP-ribosylhydrolase like 1
HTRA4	1.84	Hs.661014	203100	HtrA serine peptidase 4
SLC1A4	1.83	Hs.654352	6509	Solute carrier family 1 (glutamate/neutral amino acid transporter), member 4
RABGAP1	1.83	Hs.271341	23637	RAB GTPase activating protein 1
PGM2L1	1.83	Hs.26612	283209	Phosphoglucomutase 2-like 1
OR8H3	1.83	Hs.553745	390152	Olfactory receptor, family 8, subfamily H, member 3
NR6A1	1.82	Hs.586460	2649	Nuclear receptor subfamily 6, group A, member 1
TPSG1	1.81	Hs.592076	25823	Tryptase gamma 1
ZNF34	1.81	Hs.631854	80778	Zinc finger protein 34

C19orf36	1.81	Hs.424049	113177	Chromosome 19 open reading frame 36
MOBKL2A	1.81	Hs.86912	126308	MOB1, Mps One Binder kinase activator-like 2A (yeast)
RAB3IL1	1.80	Hs.13759	5866	RAB3A interacting protein (rabin3)-like 1
DMTF1	1.80	Hs.654981	9988	Cyclin D binding myb-like transcription factor 1
ACAA2	1.80	Hs.200136	10449	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)
ANKRD10	1.80	Hs.525163	55608	Ankyrin repeat domain 10
MTUS1	1.80	Hs.7946	57509	Mitochondrial tumor suppressor 1
C14orf80	1.80	Hs.72363	283643	Chromosome 14 open reading frame 80
MAP3K8	1.79	Hs.432453	1326	Mitogen-activated protein kinase kinase kinase 8
MEF2A	1.79	Hs.268675	4205	Myocyte enhancer factor 2A
PDE3A	1.79	Hs.591150	5139	Phosphodiesterase 3A, cGMP-inhibited
SIRT1	1.79	Hs.369779	23411	Sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)
WDR34	1.79	Hs.495240	89891	WD repeat domain 34
SELM	1.79	Hs.55940	140606	Selenoprotein M
VWA2	1.79	Hs.197741	340706	Von Willebrand factor A domain containing 2
LOC388272	1.79	Hs.705603	388272	Similar to RIKEN cDNA 4921524J17
ACACA	1.78	Hs.160556	31	Acetyl-Coenzyme A carboxylase alpha
PRKAR1A	1.78	Hs.280342	5573	Protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)
PPP1R11	1.78	Hs.82887	6992	Protein phosphatase 1, regulatory (inhibitor) subunit 11
C2orf25	1.78	Hs.5324	27249	Chromosome 2 open reading frame 25
SLC25A23	1.78	Hs.356231	79085	Solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 23
CRB3	1.78	Hs.150319	92359	Crumbs homolog 3 (Drosophila)
GPR39	1.77	Hs.432395	2863	G protein-coupled receptor 39
PPM1B	1.77	Hs.416769	5495	Protein phosphatase 1B (formerly 2C), magnesium-dependent, beta isoform
TXK	1.77	Hs.479669	7294	TXK tyrosine kinase
CIB1	1.77	Hs.135471	10519	Calcium and integrin binding 1 (calmyrin)
OR6C2	1.77	Hs.524483	341416	Olfactory receptor, family 6, subfamily C, member 2
DLX3	1.76	Hs.134194	1747	Distal-less homeobox 3
EWSR1	1.76	Hs.374477	2130	Ewing sarcoma breakpoint region 1
H3F3B	1.76	Hs.180877	3021	H3 histone, family 3B (H3.3B)
PFKFB3	1.76	Hs.195471	5209	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
STYX	1.76	Hs.364980	6815	Serine/threonine/tyrosine interacting protein
TRDN	1.76	Hs.654601	10345	Triadin
KIAA0182	1.76	Hs.461647	23199	KIAA0182
RHBDD3	1.76	Hs.106730	25807	Rhomboid domain containing 3
PLEKHG3	1.76	Hs.509637	26030	Pleckstrin homology domain containing, family G (with RhoGef domain) member 3
AXUD1	1.76	Hs.370950	64651	AXIN1 up-regulated 1
GSTP1	1.75	Hs.523836	2950	Glutathione S-transferase pi
MAP2K1	1.75	Hs.145442	5604	Mitogen-activated protein kinase kinase 1
CTDSP2	1.75	Hs.524530	10106	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 2
DDX39	1.75	Hs.311609	10212	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39
RAB40C	1.75	Hs.459630	57799	RAB40C, member RAS oncogene family

INTS4	1.75	Hs.533723	92105	Integrator complex subunit 4
PDHX	1.74	Hs.502315	8050	Pyruvate dehydrogenase complex, component X
APIP	1.74	Hs.447794	51074	APAF1 interacting protein
LENG9	1.74	Hs.590976	94059	Leukocyte receptor cluster (LRC) member 9
KCNN4	1.73	Hs.10082	3783	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4
RFX1	1.73	Hs.655215	5989	Regulatory factor X, 1 (influences HLA class II expression)
TNIP1	1.73	Hs.543850	10318	TNFAIP3 interacting protein 1
ANK2	1.72	Hs.620557	287	Ankyrin 2, neuronal
DIAPH2	1.72	Hs.696382	1730	Diaphanous homolog 2 (Drosophila)
TSC22D3	1.72	Hs.522074	1831	TSC22 domain family, member 3
ZNF155	1.72	Hs.502127	7711	Zinc finger protein 155
PAK6	1.72	Hs.513645	56924	P21(CDKN1A)-activated kinase 6
RFT1	1.72	Hs.631910	91869	RFT1 homolog (S. cerevisiae)
ME1	1.71	Hs.21160	4199	Malic enzyme 1, NADP(+)-dependent, cytosolic
PSMD13	1.71	Hs.134688	5719	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 13
VAPB	1.71	Hs.182625	9217	VAMP (vesicle-associated membrane protein)-associated protein B and C
COTL1	1.71	Hs.289092	23406	Coactosin-like 1 (Dictyostelium)
SIRT3	1.71	Hs.592292	23410	Sirtuin (silent mating type information regulation 2 homolog) 3 (S. cerevisiae)
TRPM7	1.71	Hs.512894	54822	Transient receptor potential cation channel, subfamily M, member 7
PPP2R2D	1.71	Hs.380372	55844	Protein phosphatase 2, regulatory subunit B, delta isoform
CARD14	1.71	Hs.675480	79092	Caspase recruitment domain family, member 14
HAVCR2	1.71	Hs.616365	84868	Hepatitis A virus cellular receptor 2
REEP3	1.71	Hs.499833	221035	Receptor accessory protein 3
GPLD1	1.70	Hs.591810	2822	Glycosylphosphatidylinositol specific phospholipase D1
MAP3K3	1.70	Hs.29282	4215	Mitogen-activated protein kinase kinase kinase 3
OR5I1	1.70	Hs.533706	10798	Olfactory receptor, family 5, subfamily I, member 1
PLEK2	1.70	Hs.170473	26499	Pleckstrin 2
EEF2K	1.70	Hs.498892	29904	Eukaryotic elongation factor-2 kinase
COX18	1.70	Hs.356697	285521	COX18 cytochrome c oxidase assembly homolog (S. cerevisiae)
ELA2	1.69	Hs.99863	1991	Elastase 2, neutrophil
SFTPD	1.69	Hs.253495	6441	Surfactant, pulmonary-associated protein D
DGCR6	1.69	Hs.474185	8214	DiGeorge syndrome critical region gene 6
DENR	1.69	Hs.22393	8562	Density-regulated protein
WDR53	1.69	Hs.385865	348793	WD repeat domain 53
FSHB	1.68	Hs.36975	2488	Follicle stimulating hormone, beta polypeptide
MMP12	1.68	Hs.1695	4321	Matrix metallopeptidase 12 (macrophage elastase)
NRL	1.68	Hs.652297	4901	Neural retina leucine zipper
PON2	1.68	Hs.530077	5445	Paraoxonase 2
RBMS2	1.68	Hs.645521	5939	RNA binding motif, single stranded interacting protein 2
WDR19	1.68	Hs.438482	57728	WD repeat domain 19
MET	1.67	Hs.132966	4233	Met proto-oncogene (hepatocyte growth factor receptor)
VDAC1	1.67	Hs.519320	7416	Voltage-dependent anion channel 1

PPP2R3C	1.67	Hs.530712	55012	Protein phosphatase 2 (formerly 2A), regulatory subunit B", gamma
H2AFV	1.67	Hs.488189	94239	H2A histone family, member V
CTGF	1.66	Hs.591346	1490	Connective tissue growth factor
GIPR	1.66	Hs.658534	2696	Gastric inhibitory polypeptide receptor
MCM6	1.66	Hs.444118	4175	Minichromosome maintenance complex component 6
FZD1	1.66	Hs.94234	8321	Frizzled homolog 1 (Drosophila)
MAN1B1	1.66	Hs.591887	11253	Mannosidase, alpha, class 1B, member 1
MAF1	1.66	Hs.19673	84232	MAF1 homolog (S. cerevisiae)
FCHO2	1.66	Hs.165762	115548	FCH domain only 2
ANXA5	1.65	Hs.480653	308	Annexin A5
SF3B2	1.65	Hs.406423	10992	Splicing factor 3b, subunit 2, 145kDa
ATF7	1.65	Hs.12286	11016	Activating transcription factor 7
OSBPL1A	1.65	Hs.370725	114876	Oxysterol binding protein-like 1A
CYP2C19	1.64	Hs.282409	1557	Cytochrome P450, family 2, subfamily C, polypeptide 19
MAP3K4	1.64	Hs.390428	4216	Mitogen-activated protein kinase kinase kinase 4
MYOG	1.64	Hs.2830	4656	Myogenin (myogenic factor 4)
PCBP2	1.64	Hs.546271	5094	Poly(rC) binding protein 2
PIN4	1.64	Hs.655623	5303	Protein (peptidylprolyl cis/trans isomerase) NIMA-interacting, 4 (parvulin)
WNT9A	1.64	Hs.149504	7483	Wingless-type MMTV integration site family, member 9A
OR10W1	1.64	Hs.531507	81341	Olfactory receptor, family 10, subfamily W, member 1
GAS2L3	1.64	Hs.20575	283431	Growth arrest-specific 2 like 3
C5orf13	1.63	Hs.36053	9315	Chromosome 5 open reading frame 13
ARL4A	1.63	Hs.245540	10124	ADP-ribosylation factor-like 4A
MESDC1	1.63	Hs.513071	59274	Mesoderm development candidate 1
P4HA3	1.63	Hs.660541	283208	Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), alpha polypeptide III
PTGS2	1.62	Hs.196384	5743	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
TEC	1.62	Hs.479670	7006	Tec protein tyrosine kinase
SLCO1B3	1.62	Hs.504966	28234	Solute carrier organic anion transporter family, member 1B3
VCPIP1	1.62	Hs.632066	80124	Valosin containing protein (p97)/p47 complex interacting protein 1
FAH	1.61	Hs.73875	2184	Fumarylacetoacetate hydrolase (fumarylacetoacetase)
PICALM	1.61	Hs.163893	8301	Phosphatidylinositol binding clathrin assembly protein
CIT	1.61	Hs.119594	11113	Citron (rho-interacting, serine/threonine kinase 21)
ANKRD2	1.61	Hs.73708	26287	Ankyrin repeat domain 2 (stretch responsive muscle)
LUC7L2	1.61	Hs.370475	51631	LUC7-like 2 (S. cerevisiae)
C21orf29	1.61	Hs.660703	54084	Chromosome 21 open reading frame 29
FBXO17	1.61	Hs.531770	115290	F-box protein 17
NEBL	1.60	Hs.5025	10529	Nebulette
FASTKD2	1.60	Hs.84429	22868	FAST kinase domains 2
TPCN1	1.60	Hs.524763	53373	Two pore segment channel 1
MDH1B	1.60	Hs.147816	130752	Malate dehydrogenase 1B, NAD (soluble)
GATA6	1.59	Hs.514746	2627	GATA binding protein 6
CCPG1	1.59	Hs.612814	9236	Cell cycle progression 1

RBM7	1.59	Hs.533736	10179	RNA binding motif protein 7
PPIL2	1.59	Hs.438587	23759	Peptidylprolyl isomerase (cyclophilin)-like 2
MTERFD3	1.59	Hs.5009	80298	MTERF domain containing 3
RTN3	1.58	Hs.473761	10313	Reticulon 3
DKFZP564J0863	1.58	Hs.356719	25923	DKFZP564J0863 protein
CTDSPL2	1.58	Hs.646495	51496	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase like 2
GALNT10	1.58	Hs.651323	55568	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 10 (GalNAc-T10)
ADCK4	1.58	Hs.130712	79934	AarF domain containing kinase 4
ITPKC	1.58	Hs.515415	80271	Inositol 1,4,5-trisphosphate 3-kinase C
OR4K17	1.58	Hs.553765	390436	Olfactory receptor, family 4, subfamily K, member 17
PIM3	1.58	Hs.530381	415116	Pim-3 oncogene
GALNAC4S	1.58	Hs.287537	51363	B cell RAG associated protein
BCL3	1.57	Hs.31210	602	B-cell CLL/lymphoma 3
CALR	1.57	Hs.515162	811	Calreticulin
FARSA	1.57	Hs.23111	2193	Phenylalanyl-tRNA synthetase, alpha subunit
GCLM	1.57	Hs.315562	2730	Glutamate-cysteine ligase, modifier subunit
KPNA4	1.57	Hs.288193	3840	Karyopherin alpha 4 (importin alpha 3)
PSMA1	1.57	Hs.102798	5682	Proteasome (prosome, macropain) subunit, alpha type, 1
HIP1R	1.57	Hs.524815	9026	Huntingtin interacting protein 1 related
GALNT7	1.57	Hs.548088	51809	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 7 (GalNAc-T7)
DAP3	1.56	Hs.516746	7818	Death associated protein 3
RGS20	1.56	Hs.368733	8601	Regulator of G-protein signaling 20
CAMTA1	1.56	Hs.397705	23261	Calmodulin binding transcription activator 1
MYEF2	1.56	Hs.6638	50804	Myelin expression factor 2
YY1AP1	1.56	Hs.584927	55249	YY1 associated protein 1
CENPN	1.56	Hs.55028	55839	Centromere protein N
ZNF223	1.55	Hs.279840	7766	Zinc finger protein 223
CCNK	1.55	Hs.705475	8812	Cyclin K
NPC2	1.55	Hs.433222	10577	Niemann-Pick disease, type C2
SNX5	1.55	Hs.316890	27131	Sorting nexin 5
SETD3	1.55	Hs.510407	84193	SET domain containing 3
ISCA2	1.55	Hs.702169	122961	Iron-sulfur cluster assembly 2 homolog (S. cerevisiae)
ADH1B	1.54	Hs.4	125	Alcohol dehydrogenase 1B (class I), beta polypeptide
PTK2B	1.54	Hs.491322	2185	PTK2B protein tyrosine kinase 2 beta
PAFAH1B2	1.54	Hs.696131	5049	Platelet-activating factor acetylhydrolase, isoform Ib, beta subunit 30kDa
PRCP	1.54	Hs.523936	5547	Prolylcarboxypeptidase (angiotensinase C)
SEPP1	1.54	Hs.275775	6414	Selenoprotein P, plasma, 1
SLC12A4	1.54	Hs.10094	6560	Solute carrier family 12 (potassium/chloride transporters), member 4
IQCE	1.54	Hs.520627	23288	IQ motif containing E
RTDR1	1.54	Hs.526920	27156	Rhabdoid tumor deletion region gene 1
C12orf5	1.54	Hs.504545	57103	Chromosome 12 open reading frame 5
TFB2M	1.54	Hs.7395	64216	Transcription factor B2, mitochondrial

C7orf27	1.54	Hs.520623	221927	Chromosome 7 open reading frame 27
CTNNA1	1.53	Hs.534797	1495	Catenin (cadherin-associated protein), alpha 1, 102kDa
EIF6	1.53	Hs.654848	3692	Eukaryotic translation initiation factor 6
POLG	1.53	Hs.702153	5428	Polymerase (DNA directed), gamma
CSAD	1.53	Hs.279815	51380	Cysteine sulfinic acid decarboxylase
SMPD4	1.53	Hs.516450	55627	Sphingomyelin phosphodiesterase 4, neutral membrane (neutral sphingomyelinase-3)
POLR3B	1.53	Hs.62696	55703	Polymerase (RNA) III (DNA directed) polypeptide B
FAM128B	1.53	Hs.469925	80097	Family with sequence similarity 128, member B
ZNF740	1.53	Hs.524458	283337	Zinc finger protein 740
RAB11FIP3	1.52	Hs.531642	9727	RAB11 family interacting protein 3 (class II)
SMURF2	1.52	Hs.705442	64750	SMAD specific E3 ubiquitin protein ligase 2
DHRS4L2	1.52	Hs.647569	317749	Dehydrogenase/reductase (SDR family) member 4 like 2
CTCF	1.51	Hs.368367	10664	CCCTC-binding factor (zinc finger protein)
MRPL39	1.51	Hs.420696	54148	Mitochondrial ribosomal protein L39
PARP16	1.51	Hs.30634	54956	Poly (ADP-ribose) polymerase family, member 16
ARPC5L	1.51	Hs.132499	81873	Actin related protein 2/3 complex, subunit 5-like
RELA	1.50	Hs.502875	5970	V-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian)
DHX38	1.50	Hs.570079	9785	DEAH (Asp-Glu-Ala-His) box polypeptide 38
POLR1D	1.50	Hs.507584	51082	Polymerase (RNA) I polypeptide D, 16kDa
PEX26	1.50	Hs.517400	55670	Peroxisome biogenesis factor 26
LNX2	1.50	Hs.132359	222484	Ligand of numb-protein X 2

ChIP-chip Significant Bound Genes, Non-Stimulated Cells

GENE ID	PEAK HEIGHT	UNIGENE ID	ENTREZ ID	GENE DESCRIPTION
NDUFA3	4.77	Hs.198269	4696	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3, 9kDa
OSCAR	4.77	Hs.347655	126014	Osteoclast associated, immunoglobulin-like receptor
SLC16A9	3.52	Hs.499709	220963	Solute carrier family 16, member 9 (monocarboxylic acid transporter 9)
SERPINE1	3.47	Hs.414795	5054	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
SRPX2	3.31	Hs.306339	27286	Sushi-repeat-containing protein, X-linked 2
COL7A1	3.08	Hs.476218	1294	Collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive)
C1orf43	2.96	Hs.287471	25912	Chromosome 1 open reading frame 43
UBAP2L	2.96	Hs.490551	9898	Ubiquitin associated protein 2-like
C14orf104	2.79	Hs.231761	55172	Chromosome 14 open reading frame 104
TAAR1	2.75	Hs.375030	134864	Trace amine associated receptor 1
TAP2	2.73	Hs.502	6891	Transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)
MTMR7	2.73	Hs.625674	9108	Myotubularin related protein 7
UGCGL1	2.69	Hs.34180	56886	UDP-glucose ceramide glucosyltransferase-like 1
SLITRK2	2.6	Hs.320368	84631	SLIT and NTRK-like family, member 2
CLC	2.57	Hs.889	1178	Charcot-Leyden crystal protein
CD44	2.5	Hs.502328	960	CD44 molecule (Indian blood group)
DDX53	2.46	Hs.434416	168400	DEAD (Asp-Glu-Ala-Asp) box polypeptide 53
FLJ45248	2.45	Hs.224506	401472	FLJ45248 protein

HOXB3	2.45	Hs.654560	3213	Homeobox B3
LOXHD1	2.44	Hs.345877	125336	Lipoxygenase homology domains 1
KIAA1576	2.44	Hs.461405	57687	KIAA1576 protein
PSPC1	2.43	Hs.213198	55269	Paraspeckle component 1
SMPX	2.42	Hs.86492	23676	Small muscle protein, X-linked
HEXB	2.4	Hs.69293	3074	Hexosaminidase B (beta polypeptide)
CYP7A1	2.38	Hs.1644	1581	Cytochrome P450, family 7, subfamily A, polypeptide 1
ZNF45	2.38	Hs.381285	7596	Zinc finger protein 45
TEX15	2.34	Hs.458316	56154	Testis expressed 15
CYP2C8	2.32	Hs.282871	1558	Cytochrome P450, family 2, subfamily C, polypeptide 8
OR1L3	2.32	Hs.626839	26735	Olfactory receptor, family 1, subfamily L, member 3
SMAD6	2.31	Hs.153863	4091	SMAD family member 6
UBAP2	2.3	Hs.493739	55833	Ubiquitin associated protein 2
OR6S1	2.25	Hs.513132	341799	Olfactory receptor, family 6, subfamily S, member 1
KIAA1377	2.24	Hs.156352	57562	KIAA1377
ANKRD32	2.24	Hs.657315	84250	Ankyrin repeat domain 32
DOCK7	2.23	Hs.538059	85440	Dedicator of cytokinesis 7
IPO11	2.21	Hs.482269	51194	Importin 11
C21orf84	2.21	Hs.592161	114038	Chromosome 21 open reading frame 84
MRC1	2.21	Hs.75182	4360	Mannose receptor, C type 1
NR0B1	2.2	Hs.268490	190	Nuclear receptor subfamily 0, group B, member 1
TCF20	2.2	Hs.475018	6942	Transcription factor 20 (AR1)
CHD1	2.2	Hs.643465	1105	Chromodomain helicase DNA binding protein 1
FAM12B	2.19	Hs.525202	64184	Family with sequence similarity 12, member B (epididymal)
SNX12	2.17	Hs.260750	29934	Sorting nexin 12
ZNF384	2.16	Hs.103315	171017	Zinc finger protein 384
MITF	2.14	Hs.166017	4286	Microphthalmia-associated transcription factor
KPTN	2.13	Hs.25441	11133	Kaptin (actin binding protein)
ITIH1	2.13	Hs.420257	3697	Inter-alpha (globulin) inhibitor H1
NEK4	2.13	Hs.631921	6787	NIMA (never in mitosis gene a)-related kinase 4
MAGEH1	2.12	Hs.279819	28986	Melanoma antigen family H, 1
ADPRHL1	2.12	Hs.98669	113622	ADP-ribosylhydrolase like 1
C14orf119	2.1	Hs.525238	55017	Chromosome 14 open reading frame 119
PTH	2.09	Hs.37045	5741	Parathyroid hormone
POU1F1	2.09	Hs.591654	5449	POU class 1 homeobox 1
MNDA	2.08	Hs.153837	4332	Myeloid cell nuclear differentiation antigen
ATAD2	2.07	Hs.370834	29028	ATPase family, AAA domain containing 2
USP9X	2.07	Hs.77578	8239	Ubiquitin specific peptidase 9, X-linked
DHX33	2.06	Hs.250456	56919	DEAH (Asp-Glu-Ala-His) box polypeptide 33
WDR47	2.06	Hs.654760	22911	WD repeat domain 47
IFNA8	2.06	Hs.73890	3445	Interferon, alpha 8
PLSCR2	2.05	Hs.147305	57047	Phospholipid scramblase 2
RANBP2	2.05	Hs.199561	5903	RAN binding protein 2
C8B	2.05	Hs.391835	732	Complement component 8, beta polypeptide

OR4D10	2.04	Hs.553756	390197	Olfactory receptor, family 4, subfamily D, member 10
VNN1	2.03	Hs.12114	8876	Vanin 1
PRDM2	2.02	Hs.371823	7799	PR domain containing 2, with ZNF domain
TMPO	2.01	Hs.11355	7112	Thymopoietin
WNT16	2.01	Hs.272375	51384	Wingless-type MMTV integration site family, member 16
KERA	2	Hs.125750	11081	Keratocan
CDC2	2	Hs.334562	983	Cell division cycle 2, G1 to S and G2 to M
ZNF615	1.99	Hs.368355	284370	Zinc finger protein 615
ZNF519	1.97	Hs.352635	162655	Zinc finger protein 519
C6orf1	1.97	Hs.381300	221491	Chromosome 6 open reading frame 1
IVD	1.97	Hs.449599	3712	Isovaleryl Coenzyme A dehydrogenase
GPHN	1.95	Hs.208765	10243	Gephyrin
IL20	1.95	Hs.272373	50604	Interleukin 20
CYP39A1	1.94	Hs.387367	51302	Cytochrome P450, family 39, subfamily A, polypeptide 1
PLCB4	1.94	Hs.472101	5332	Phospholipase C, beta 4
CCDC7	1.93	Hs.585464	221016	Coiled-coil domain containing 7
OR7A5	1.92	Hs.137573	26659	Olfactory receptor, family 7, subfamily A, member 5
FBXL19	1.92	Hs.152149	54620	F-box and leucine-rich repeat protein 19
PPM1E	1.92	Hs.245044	22843	Protein phosphatase 1E (PP2C domain containing)
NOX3	1.92	Hs.247776	50508	NADPH oxidase 3
NQO1	1.92	Hs.406515	1728	NAD(P)H dehydrogenase, quinone 1
WDR22	1.92	Hs.509780	8816	WD repeat domain 22
SCN10A	1.91	Hs.250443	6336	Sodium channel, voltage-gated, type X, alpha subunit
ZNF146	1.91	Hs.643436	7705	Zinc finger protein 146
C10orf120	1.9	Hs.363649	399814	Chromosome 10 open reading frame 120
KDELR2	1.9	Hs.654552	11014	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2
SGCG	1.89	Hs.37167	6445	Sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein)
ABCG2	1.89	Hs.480218	9429	ATP-binding cassette, sub-family G (WHITE), member 2
OR4A16	1.89	Hs.554530	81327	Olfactory receptor, family 4, subfamily A, member 16
CNGB3	1.88	Hs.154433	54714	Cyclic nucleotide gated channel beta 3
SUPT6H	1.87	Hs.250429	6830	Suppressor of Ty 6 homolog (S. cerevisiae)
RECK	1.87	Hs.388918	8434	Reversion-inducing-cysteine-rich protein with kazal motifs
RCOR3	1.87	Hs.696152	55758	REST corepressor 3
GJA1	1.87	Hs.74471	2697	Gap junction protein, alpha 1, 43kDa
DLAT	1.86	Hs.335551	1737	Dihydrolipoamide S-acetyltransferase (E2 component of pyruvate dehydrogenase complex)
DBP	1.86	Hs.414480	1628	D site of albumin promoter (albumin D-box) binding protein
OCM	1.86	Hs.571315	654231	Oncomodulin
SUMF2	1.85	Hs.279696	25870	Sulfatase modifying factor 2
CCT6A	1.85	Hs.82916	908	Chaperonin containing TCP1, subunit 6A (zeta 1)
NME1	1.84	Hs.463456	4830	Non-metastatic cells 1, protein (NM23A) expressed in
MGAT5	1.84	Hs.651869	4249	Mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase
ARHGAP1	1.83	Hs.138860	392	Rho GTPase activating protein 1
GPNMB	1.83	Hs.190495	10457	Glycoprotein (transmembrane) nmb
PDZRN4	1.83	Hs.380044	29951	PDZ domain containing RING finger 4

ZNF408	1.83	Hs.656931	79797	Zinc finger protein 408
SLC31A2	1.82	Hs.24030	1318	Solute carrier family 31 (copper transporters), member 2
WFDC3	1.82	Hs.419126	140686	WAP four-disulfide core domain 3
SCEL	1.81	Hs.534699	8796	Sciellin
DMTF1	1.81	Hs.654981	9988	Cyclin D binding myb-like transcription factor 1
ADH4	1.8	Hs.1219	127	Alcohol dehydrogenase 4 (class II), pi polypeptide
DDX5	1.8	Hs.279806	1655	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5
P4HA2	1.8	Hs.519568	8974	Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), alpha polypeptide II
OR4D9	1.8	Hs.553757	390199	Olfactory receptor, family 4, subfamily D, member 9
TMEM20	1.8	Hs.632085	159371	Transmembrane protein 20
CDC14A	1.79	Hs.127411	8556	CDC14 cell division cycle 14 homolog A (S. cerevisiae)
COTL1	1.79	Hs.289092	23406	Coactosin-like 1 (Dictyostelium)
RAB11A	1.79	Hs.321541	8766	RAB11A, member RAS oncogene family
S100A8	1.79	Hs.416073	6279	S100 calcium binding protein A8
MRPL39	1.79	Hs.420696	54148	Mitochondrial ribosomal protein L39
MCM7	1.79	Hs.438720	4176	Minichromosome maintenance complex component 7
ULK1	1.79	Hs.47061	8408	Unc-51-like kinase 1 (C. elegans)
MORF4L2	1.78	Hs.326387	9643	Mortality factor 4 like 2
SMAD7	1.78	Hs.465087	4092	SMAD family member 7
CSNK1E	1.78	Hs.474833	1454	Casein kinase 1, epsilon
C6orf170	1.77	Hs.121396	221322	Chromosome 6 open reading frame 170
FOXA3	1.77	Hs.36137	3171	Forkhead box A3
LCP1	1.77	Hs.381099	3936	Lymphocyte cytosolic protein 1 (L-plastin)
TRIB1	1.77	Hs.444947	10221	Tribbles homolog 1 (Drosophila)
GRM8	1.77	Hs.449625	2918	Glutamate receptor, metabotropic 8
FILIP1	1.77	Hs.696158	27145	Filamin A interacting protein 1
RAB2B	1.76	Hs.22399	84932	RAB2B, member RAS oncogene family
SLC19A2	1.76	Hs.30246	10560	Solute carrier family 19 (thiamine transporter), member 2
LECT2	1.76	Hs.512580	3950	Leukocyte cell-derived chemotaxin 2
KCNJ5	1.76	Hs.632109	3762	Potassium inwardly-rectifying channel, subfamily J, member 5
SCN2B	1.75	Hs.129783	6327	Sodium channel, voltage-gated, type II, beta
MBIP	1.75	Hs.368647	51562	MAP3K12 binding inhibitory protein 1
C14orf124	1.75	Hs.643552	56948	Chromosome 14 open reading frame 124
RPH3AL	1.75	Hs.651925	9501	Rabphilin 3A-like (without C2 domains)
NDUFB2	1.75	Hs.655788	4708	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa
TRIM16	1.74	Hs.123534	10626	Tripartite motif-containing 16
XDH	1.74	Hs.250	7498	Xanthine dehydrogenase
PSMC3	1.74	Hs.250758	5702	Proteasome (prosome, macropain) 26S subunit, ATPase, 3
SFTPD	1.74	Hs.253495	6441	Surfactant, pulmonary-associated protein D
RAB1A	1.74	Hs.310645	5861	RAB1A, member RAS oncogene family
CLDN11	1.74	Hs.31595	5010	Claudin 11 (oligodendrocyte transmembrane protein)
HDHD2	1.74	Hs.465041	84064	Haloacid dehalogenase-like hydrolase domain containing 2
OR8H3	1.74	Hs.553745	390152	Olfactory receptor, family 8, subfamily H, member 3
MLL5	1.74	Hs.592262	55904	Myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, Drosophila)

TBL1X	1.74	Hs.699315	6907	Transducin (beta)-like 1X-linked
UMPS	1.73	Hs.2057	7372	Uridine monophosphate synthetase (orotate phosphoribosyl transferase and orotidine-5'-decarboxylase)
USP52	1.73	Hs.273397	9924	Ubiquitin specific peptidase 52
RPS4Y1	1.73	Hs.282376	6192	Ribosomal protein S4, Y-linked 1
ENTPD1	1.73	Hs.576612	953	Ectonucleoside triphosphate diphosphohydrolase 1
CALM2	1.73	Hs.643483	805	Calmodulin 2 (phosphorylase kinase, delta)
GNRHR	1.72	Hs.407587	2798	Gonadotropin-releasing hormone receptor
POLI	1.72	Hs.438533	11201	Polymerase (DNA directed) iota
OPA3	1.72	Hs.466945	80207	Optic atrophy 3 (autosomal recessive, with chorea and spastic paraplegia)
FBXL2	1.72	Hs.475872	25827	F-box and leucine-rich repeat protein 2
MAGEA9	1.72	Hs.512582	4108	Melanoma antigen family A, 9
PJA1	1.72	Hs.522679	64219	Praja 1
SUMF1	1.72	Hs.588682	285362	Sulfatase modifying factor 1
ZDHHC7	1.72	Hs.592065	55625	Zinc finger, DHHC-type containing 7
TMEM41A	1.72	Hs.634586	90407	Transmembrane protein 41A
AKR1C3	1.72	Hs.78183	8644	Aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II)
ARID4A	1.71	Hs.161000	5926	AT rich interactive domain 4A (RBP1-like)
NAP1L3	1.71	Hs.21365	4675	Nucleosome assembly protein 1-like 3
WWP2	1.71	Hs.408458	11060	WW domain containing E3 ubiquitin protein ligase 2
IPO4	1.71	Hs.411865	79711	Importin 4
RRAS	1.71	Hs.515536	6237	Related RAS viral (r-ras) oncogene homolog
HFE2	1.71	Hs.632436	148738	Hemochromatosis type 2 (juvenile)
POM121	1.71	Hs.655217	9883	POM121 membrane glycoprotein (rat)
VASP	1.71	Hs.702197	7408	Vasodilator-stimulated phosphoprotein
PSMB7	1.7	Hs.213470	5695	Proteasome (prosome, macropain) subunit, beta type, 7
ART3	1.7	Hs.24976	419	ADP-ribosyltransferase 3
IL13RA2	1.7	Hs.336046	3598	Interleukin 13 receptor, alpha 2
SLC22A9	1.7	Hs.502772	114571	Solute carrier family 22 (organic anion/cation transporter), member 9
ZNF410	1.69	Hs.270869	57862	Zinc finger protein 410
PIAS3	1.69	Hs.435761	10401	Protein inhibitor of activated STAT, 3
TPSG1	1.69	Hs.592076	25823	Tryptase gamma 1
PPM1A	1.69	Hs.592298	5494	Protein phosphatase 1A (formerly 2C), magnesium-dependent, alpha isoform
ZBTB1	1.69	Hs.655536	22890	Zinc finger and BTB domain containing 1
HNRPUL1	1.69	Hs.699274	11100	Heterogeneous nuclear ribonucleoprotein U-like 1
SLITRK4	1.68	Hs.272284	139065	SLIT and NTRK-like family, member 4
USP3	1.68	Hs.458499	9960	Ubiquitin specific peptidase 3
HSPA5	1.68	Hs.605502	3309	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
GABRA3	1.67	Hs.123024	2556	Gamma-aminobutyric acid (GABA) A receptor, alpha 3
USP32	1.67	Hs.132868	84669	Ubiquitin specific peptidase 32
MRCL3	1.67	Hs.190086	10627	Myosin regulatory light chain MRCL3
IL22	1.67	Hs.287369	50616	Interleukin 22
C5orf15	1.67	Hs.355177	56951	Chromosome 5 open reading frame 15
SSX6	1.67	Hs.511998	280657	Synovial sarcoma, X breakpoint 6
PON2	1.67	Hs.530077	5445	Paraoxonase 2

TOB1	1.67	Hs.531550	10140	Transducer of ERBB2, 1
CCDC9	1.66	Hs.227782	26093	Coiled-coil domain containing 9
WDR34	1.66	Hs.495240	89891	WD repeat domain 34
BRMS1L	1.66	Hs.525299	84312	Breast cancer metastasis-suppressor 1-like
PTPN22	1.66	Hs.535276	26191	Protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
NCOA5	1.66	Hs.654991	57727	Nuclear receptor coactivator 5
PIN4	1.66	Hs.655623	5303	Protein (peptidylprolyl cis/trans isomerase) NIMA-interacting, 4 (parvulin)
SMYD2	1.66	Hs.66170	56950	SET and MYND domain containing 2
EDG6	1.66	Hs.662006	8698	Endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 6
CSN1S1	1.65	Hs.3155	1446	Casein alpha s1
CPA2	1.65	Hs.490038	1358	Carboxypeptidase A2 (pancreatic)
SEMA3E	1.65	Hs.528721	9723	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3E
OR1S2	1.65	Hs.553644	219958	Olfactory receptor, family 1, subfamily S, member 2
OR1S1	1.65	Hs.553645	219959	Olfactory receptor, family 1, subfamily S, member 1
PRKAB1	1.65	Hs.6061	5564	Protein kinase, AMP-activated, beta 1 non-catalytic subunit
DHX34	1.64	Hs.151706	9704	DEAH (Asp-Glu-Ala-His) box polypeptide 34
DMC1	1.64	Hs.339396	11144	DMC1 dosage suppressor of mck1 homolog, meiosis-specific homologous recombination (yeast)
EXTL2	1.64	Hs.357637	2135	Exostoses (multiple)-like 2
ITGB6	1.64	Hs.470399	3694	Integrin, beta 6
SLC30A7	1.64	Hs.533903	148867	Solute carrier family 30 (zinc transporter), member 7
C5orf5	1.64	Hs.657919	51306	Chromosome 5 open reading frame 5
USP51	1.63	Hs.134289	158880	Ubiquitin specific peptidase 51
FMO2	1.63	Hs.144912	2327	Flavin containing monooxygenase 2 (non-functional)
HDC	1.63	Hs.1481	3067	Histidine decarboxylase
C14orf4	1.63	Hs.179260	64207	Chromosome 14 open reading frame 4
GSR	1.63	Hs.271510	2936	Glutathione reductase
PDGFD	1.63	Hs.352298	80310	Platelet derived growth factor D
BEX2	1.63	Hs.398989	84707	Brain expressed X-linked 2
CDC14B	1.63	Hs.40582	8555	CDC14 cell division cycle 14 homolog B (S. cerevisiae)
KIAA0831	1.63	Hs.414809	22863	KIAA0831
CHM	1.63	Hs.496449	1121	Choroideremia (Rab escort protein 1)
KCTD7	1.63	Hs.546627	154881	Potassium channel tetramerisation domain containing 7
OR2T12	1.63	Hs.553582	127064	Olfactory receptor, family 2, subfamily T, member 12
SLC23A1	1.63	Hs.643467	9963	Solute carrier family 23 (nucleobase transporters), member 1
GNAS	1.62	Hs.125898	2778	GNAS complex locus
MAN2C1	1.62	Hs.26232	4123	Mannosidase, alpha, class 2C, member 1
PRRG2	1.62	Hs.35101	5639	Proline rich Gla (G-carboxyglutamic acid) 2
PRKAB2	1.62	Hs.50732	5565	Protein kinase, AMP-activated, beta 2 non-catalytic subunit
OCIAD1	1.62	Hs.518750	54940	OCIA domain containing 1
FH	1.62	Hs.592490	2271	Fumarate hydratase
DNAH10	1.62	Hs.622654	196385	Dynein, axonemal, heavy chain 10
PDC	1.62	Hs.654381	5132	Phosducin
NOSIP	1.62	Hs.7236	51070	Nitric oxide synthase interacting protein
CHRNA6	1.61	Hs.103128	8973	Cholinergic receptor, nicotinic, alpha 6

UVRAG	1.61	Hs.202470	7405	UV radiation resistance associated gene
CCL14	1.61	Hs.272493	6358	Chemokine (C-C motif) ligand 14
NPHP1	1.61	Hs.280388	4867	Nephronophthisis 1 (juvenile)
LBR	1.61	Hs.435166	3930	Lamin B receptor
SUV39H1	1.61	Hs.522639	6839	Suppressor of variegation 3-9 homolog 1 (Drosophila)
OR4M1	1.61	Hs.553829	441670	Olfactory receptor, family 4, subfamily M, member 1
MYEF2	1.61	Hs.6638	50804	Myelin expression factor 2
NDUFS2	1.6	Hs.173611	4720	NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase)
PARP2	1.6	Hs.409412	10038	Poly (ADP-ribose) polymerase family, member 2
LOC51057	1.6	Hs.414952	51057	Hypothetical protein LOC51057
CYHR1	1.6	Hs.459379	50626	Cysteine/histidine-rich 1
MDH1	1.6	Hs.526521	4190	Malate dehydrogenase 1, NAD (soluble)
OR4K13	1.6	Hs.553573	390433	Olfactory receptor, family 4, subfamily K, member 13
FUCA2	1.6	Hs.591332	2519	Fucosidase, alpha-L- 2, plasma
SERPINA12	1.6	Hs.99476	145264	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 12
SMAD2	1.59	Hs.12253	4087	SMAD family member 2
NEGR1	1.59	Hs.146542	257194	Neuronal growth regulator 1
H3F3B	1.59	Hs.180877	3021	H3 histone, family 3B (H3.3B)
ZNF302	1.59	Hs.436350	55900	Zinc finger protein 302
TWSG1	1.59	Hs.514685	57045	Twisted gastrulation homolog 1 (Drosophila)
SERPINA1	1.59	Hs.525557	5265	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
CASP10	1.59	Hs.5353	843	Caspase 10, apoptosis-related cysteine peptidase
ASB5	1.59	Hs.591712	140458	Ankyrin repeat and SOCS box-containing 5
B2M	1.59	Hs.626605	567	Beta-2-microglobulin
KIF23	1.58	Hs.270845	9493	Kinesin family member 23
DCTN4	1.58	Hs.328865	51164	Dynactin 4 (p62)
PPIG	1.58	Hs.470544	9360	Peptidylprolyl isomerase G (cyclophilin G)
AQR	1.58	Hs.510958	9716	Aquarius homolog (mouse)
AP2B1	1.58	Hs.514819	163	Adaptor-related protein complex 2, beta 1 subunit
PEX12	1.58	Hs.591190	5193	Peroxisomal biogenesis factor 12
LRRC6	1.58	Hs.591865	23639	Leucine rich repeat containing 6
LPPR2	1.58	Hs.6846	64748	Lipid phosphate phosphatase-related protein type 2
FAH	1.58	Hs.73875	2184	Fumarylacetoacetate hydrolase (fumarylacetoacetase)
ZBTB3	1.57	Hs.147554	79842	Zinc finger and BTB domain containing 3
RHOBTB2	1.57	Hs.372688	23221	Rho-related BTB domain containing 2
PEBP4	1.57	Hs.491242	157310	Phosphatidylethanolamine-binding protein 4
GJB2	1.57	Hs.524894	2706	Gap junction protein, beta 2, 26kDa
JMJD1A	1.57	Hs.557425	55818	Jumonji domain containing 1A
NCOA1	1.57	Hs.699183	8648	Nuclear receptor coactivator 1
SLC39A6	1.57	Hs.79136	25800	Solute carrier family 39 (zinc transporter), member 6
HTR1E	1.56	Hs.1611	3354	5-hydroxytryptamine (serotonin) receptor 1E
PPP1CA	1.56	Hs.183994	5499	Protein phosphatase 1, catalytic subunit, alpha isoform
KIAA0528	1.56	Hs.271014	9847	KIAA0528
TSGA14	1.56	Hs.368315	95681	Testis specific, 14

FLJ34503	1.56	Hs.376634	285759	Hypothetical FLJ34503
EFHA1	1.56	Hs.412103	221154	EF-hand domain family, member A1
NUDCD3	1.56	Hs.488171	23386	NudC domain containing 3
OR2T29	1.56	Hs.553707	343563	Olfactory receptor, family 2, subfamily T, member 29
PCDH9	1.56	Hs.654709	5101	Protocadherin 9
BNC2	1.56	Hs.656581	54796	Basonuclin 2
AKAP10	1.56	Hs.694769	11216	A kinase (PRKA) anchor protein 10
FOXP4	1.55	Hs.131436	116113	Forkhead box P4
SENP1	1.55	Hs.371957	29843	SUMO1/sentrin specific peptidase 1
MAP2K6	1.55	Hs.463978	5608	Mitogen-activated protein kinase kinase 6
SLC9A11	1.55	Hs.494981	284525	Solute carrier family 9, member 11
ZNF41	1.55	Hs.496074	7592	Zinc finger protein 41
SPRED1	1.55	Hs.525781	161742	Sprouty-related, EVH1 domain containing 1
GSPT1	1.55	Hs.528780	2935	G1 to S phase transition 1
COPS7A	1.55	Hs.530823	50813	COP9 constitutive photomorphogenic homolog subunit 7A (Arabidopsis)
SCAP	1.55	Hs.531789	22937	SREBF chaperone
RIN1	1.54	Hs.1030	9610	Ras and Rab interactor 1
C14orf126	1.54	Hs.116014	112487	Chromosome 14 open reading frame 126
EPHX2	1.54	Hs.212088	2053	Epoxide hydrolase 2, cytoplasmic
NFYC	1.54	Hs.233458	4802	Nuclear transcription factor Y, gamma
C21orf81	1.54	Hs.364456	114035	Chromosome 21 open reading frame 81
RBJ	1.54	Hs.434993	51277	Rab and DnaJ domain containing
FAM33A	1.54	Hs.463607	348235	Family with sequence similarity 33, member A
MTF1	1.54	Hs.471991	4520	Metal-regulatory transcription factor 1
GNL3L	1.54	Hs.654677	54552	Guanine nucleotide binding protein-like 3 (nucleolar)-like
SNX11	1.53	Hs.15827	29916	Sorting nexin 11
PRPF18	1.53	Hs.161181	8559	PRP18 pre-mRNA processing factor 18 homolog (S. cerevisiae)
ARIH1	1.53	Hs.268787	25820	Ariadne homolog, ubiquitin-conjugating enzyme E2 binding protein, 1 (Drosophila)
TXN	1.53	Hs.435136	7295	Thioredoxin
NEUROD1	1.53	Hs.440955	4760	Neurogenic differentiation 1
GLT6D1	1.53	Hs.522491	360203	Glycosyltransferase 6 domain containing 1
NOTCH2NL	1.53	Hs.655156	388677	Notch homolog 2 (Drosophila) N-terminal like
C14orf43	1.53	Hs.656506	91748	Chromosome 14 open reading frame 43
CBX1	1.53	Hs.77254	10951	Chromobox homolog 1 (HP1 beta homolog Drosophila)
HSPE1	1.52	Hs.1197	3336	Heat shock 10kDa protein 1 (chaperonin 10)
PARP15	1.52	Hs.120250	165631	Poly (ADP-ribose) polymerase family, member 15
HAP1	1.52	Hs.158300	9001	Huntingtin-associated protein 1 (neuroan 1)
TAF1	1.52	Hs.158560	6872	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa
CASP9	1.52	Hs.329502	842	Caspase 9, apoptosis-related cysteine peptidase
HIST4H4	1.52	Hs.352191	121504	Histone cluster 4, H4
RPS17	1.52	Hs.433427	6218	Ribosomal protein S17
DAB2	1.52	Hs.481980	1601	Disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)
RUSC2	1.52	Hs.493796	9853	RUN and SH3 domain containing 2
TAGLN2	1.52	Hs.517168	8407	Transgelin 2

EPS8	1.52	Hs.591160	2059	Epidermal growth factor receptor pathway substrate 8
HSPD1	1.52	Hs.595053	3329	Heat shock 60kDa protein 1 (chaperonin)
LOX	1.51	Hs.102267	4015	Lysyl oxidase
UQCRB	1.51	Hs.131255	7381	Ubiquinol-cytochrome c reductase binding protein
ACAA2	1.51	Hs.200136	10449	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)
STMN4	1.51	Hs.201058	81551	Stathmin-like 4
FLJ11184	1.51	Hs.267446	55319	Hypothetical protein FLJ11184
MAP3K3	1.51	Hs.29282	4215	Mitogen-activated protein kinase kinase kinase 3
FEN1	1.51	Hs.409065	2237	Flap structure-specific endonuclease 1
TARS	1.51	Hs.481860	6897	Threonyl-tRNA synthetase
ATXN3	1.51	Hs.532632	4287	Ataxin 3
DLG7	1.51	Hs.77695	9787	Discs, large homolog 7 (Drosophila)
SHOC2	1.5	Hs.104315	8036	Soc-2 suppressor of clear homolog (C. elegans)
PTDSS2	1.5	Hs.12851	81490	Phosphatidylserine synthase 2
AQP2	1.5	Hs.130730	359	Aquaporin 2 (collecting duct)
SNAP25	1.5	Hs.167317	6616	Synaptosomal-associated protein, 25kDa
EHBP1	1.5	Hs.271667	23301	EH domain binding protein 1
RBBP8	1.5	Hs.546282	5932	Retinoblastoma binding protein 8
SREBF1	1.5	Hs.592123	6720	Sterol regulatory element binding transcription factor 1
HEXA	1.5	Hs.604479	3073	Hexosaminidase A (alpha polypeptide)
MANBAL	1.5	Hs.6126	63905	Mannosidase, beta A, lysosomal-like
ZBTB20	1.5	Hs.655108	26137	Zinc finger and BTB domain containing 20
NEK11	1.5	Hs.657336	79858	NIMA (never in mitosis gene a)- related kinase 11
PPP1R11	1.5	Hs.82887	6992	Protein phosphatase 1, regulatory (inhibitor) subunit 11

BIBLIOGRAPHY

1. Bhatt, N., et al., *Promising pharmacologic innovations in treating pulmonary fibrosis.* Curr Opin Pharmacol, 2006. **6**(3): p. 284-92.
2. ATS, *American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS).* Am J Respir Crit Care Med, 2000. **161**(2 Pt 1): p. 646-64.
3. Walter, N., H.R. Collard, and T.E. King, Jr., *Current perspectives on the treatment of idiopathic pulmonary fibrosis.* Proc Am Thorac Soc, 2006. **3**(4): p. 330-8.
4. Martinez, F.J., et al., *The clinical course of patients with idiopathic pulmonary fibrosis.* Ann Intern Med, 2005. **142**(12 Pt 1): p. 963-7.
5. Kim, D.S., H.R. Collard, and T.E. King, Jr., *Classification and natural history of the idiopathic interstitial pneumonias.* Proc Am Thorac Soc, 2006. **3**(4): p. 285-92.
6. Selman, M., et al., *Idiopathic pulmonary fibrosis: pathogenesis and therapeutic approaches.* Drugs, 2004. **64**(4): p. 405-30.
7. Bouros, D. and K.M. Antoniou, *Current and future therapeutic approaches in idiopathic pulmonary fibrosis.* Eur Respir J, 2005. **26**(4): p. 693-702.
8. Selman, M., T.E. King, and A. Pardo, *Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy.* Ann Intern Med, 2001. **134**(2): p. 136-51.
9. Costabel, U. and T.E. King, *International consensus statement on idiopathic pulmonary fibrosis.* Eur Respir J, 2001. **17**(2): p. 163-7.
10. King, T.E., Jr., *Clinical advances in the diagnosis and therapy of the interstitial lung diseases.* Am J Respir Crit Care Med, 2005. **172**(3): p. 268-79.
11. King, T.E., Jr., et al., *Predicting survival in idiopathic pulmonary fibrosis: scoring system and survival model.* Am J Respir Crit Care Med, 2001. **164**(7): p. 1171-81.
12. Wynn, T., *Cellular and molecular mechanisms of fibrosis.* J Pathol, 2008. **214**(2): p. 199-210.
13. Kim, J.H., et al., *Transforming growth factor beta1 induces epithelial-to-mesenchymal transition of A549 cells.* J Korean Med Sci, 2007. **22**(5): p. 898-904.
14. Kim, K.K., et al., *Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix.* Proc Natl Acad Sci U S A, 2006. **103**(35): p. 13180-5.
15. Saika, S., et al., *Smad3 signaling is required for epithelial-mesenchymal transition of lens epithelium after injury.* Am J Pathol, 2004. **164**(2): p. 651-63.
16. Zavadil, J. and E.P. Bottinger, *TGF-beta and epithelial-to-mesenchymal transitions.* Oncogene, 2005. **24**(37): p. 5764-74.

17. Massague, J. and R.R. Gomis, *The logic of TGFbeta signaling.* FEBS Lett, 2006. **580**(12): p. 2811-20.
18. Shi, Y. and J. Massague, *Mechanisms of TGF-beta signaling from cell membrane to the nucleus.* Cell, 2003. **113**(6): p. 685-700.
19. Verrecchia, F. and A. Mauviel, *Transforming growth factor-beta signaling through the Smad pathway: role in extracellular matrix gene expression and regulation.* J Invest Dermatol, 2002. **118**(2): p. 211-5.
20. Feng, X.H. and R. Derynck, *Specificity and versatility in tgf-beta signaling through Smads.* Annu Rev Cell Dev Biol, 2005. **21**: p. 659-93.
21. Derynck, R., Y. Zhang, and X.H. Feng, *Smads: transcriptional activators of TGF-beta responses.* Cell, 1998. **95**(6): p. 737-40.
22. Derynck, R. and X.H. Feng, *TGF-beta receptor signaling.* Biochim Biophys Acta, 1997. **1333**(2): p. F105-50.
23. Derynck, R. and K. Miyazono, *The TGF - beta family*. Cold Spring Harbor monograph series. 2008, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. xiv, 1114.
24. Mehra, A. and J.L. Wrana, *TGF-beta and the Smad signal transduction pathway.* Biochem Cell Biol, 2002. **80**(5): p. 605-22.
25. Ewis, A.A., et al., *A history of microarrays in biomedicine.* Expert Rev Mol Diagn, 2005. **5**(3): p. 315-28.
26. Southern, E.M., *DNA microarrays. History and overview.* Methods Mol Biol, 2001. **170**: p. 1-15.
27. Schena, M., et al., *Microarrays: biotechnology's discovery platform for functional genomics.* Trends Biotechnol, 1998. **16**(7): p. 301-6.
28. Hardiman, G., *Microarray platforms--comparisons and contrasts.* Pharmacogenomics, 2004. **5**(5): p. 487-502.
29. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
30. Cheung, V.G., et al., *Making and reading microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 15-9.
31. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
32. Systems, I. *Ingenuity Pathways Analysis*. Ingenuity Pathways Analysis 2008 [cited 2008; Available from: http://www.ingenuity.com/.
33. GeneGo, M. *GeneGo is a leading provider of bioinformatics software solutions for data mining applications in systems biology.* 2008 [cited; Available from: http://www.genego.com/.
34. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
35. Yu, H., et al., *Transgelin is a direct target of TGF-{beta}/Smad3-dependent epithelial cell migration in lung fibrosis.* Faseb J, 2008.
36. Wan, H., et al., *Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis.* J Biol Chem, 2005. **280**(14): p. 13809-16.
37. Wan, H., et al., *Foxa2 is required for transition to air breathing at birth.* Proc Natl Acad Sci U S A, 2004. **101**(40): p. 14449-54.
38. Wan, H., et al., *Foxa2 regulates alveolarization and goblet cell hyperplasia.* Development, 2004. **131**(4): p. 953-64.

39. Maeda, Y., V. Dave, and J.A. Whitsett, *Transcriptional control of lung morphogenesis.* Physiol Rev, 2007. **87**(1): p. 219-44.
40. Besnard, V., et al., *Immunohistochemical localization of Foxa1 and Foxa2 in mouse embryos and adult tissues.* Gene Expr Patterns, 2004. **5**(2): p. 193-208.
41. Nogee, L.M., et al., *Mutations in the surfactant protein C gene associated with interstitial lung disease.* Chest, 2002. **121**(3 Suppl): p. 20S-21S.
42. Nogee, L.M., et al., *A mutation in the surfactant protein C gene associated with familial interstitial lung disease.* N Engl J Med, 2001. **344**(8): p. 573-9.
43. Lawson, W.E., et al., *Genetic mutations in surfactant protein C are a rare cause of sporadic cases of IPF.* Thorax, 2004. **59**(11): p. 977-80.
44. Thomas, A.Q., et al., *Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred.* Am J Respir Crit Care Med, 2002. **165**(9): p. 1322-8.
45. Whitsett, J.A., S.E. Wert, and Y. Xu, *Genetic disorders of surfactant homeostasis.* Biol Neonate, 2005. **87**(4): p. 283-7.
46. Lin, J., et al., *Characterization of a novel effect of hPinX1 on hTERT nucleolar localization.* Biochem Biophys Res Commun, 2007. **353**(4): p. 946-52.
47. Banik, S.S. and C.M. Counter, *Characterization of interactions between PinX1 and human telomerase subunits hTERT and hTR.* J Biol Chem, 2004. **279**(50): p. 51745-8.
48. Akiyama, Y., et al., *Human PinX1, a potent telomerase inhibitor, is not involved in human gastrointestinal tract carcinoma.* Oncol Rep, 2004. **11**(4): p. 871-4.
49. Zhou, X.Z. and K.P. Lu, *The Pin2/TRF1-interacting protein PinX1 is a potent telomerase inhibitor.* Cell, 2001. **107**(3): p. 347-59.
50. Tsakiri, K.D., et al., *Adult-onset pulmonary fibrosis caused by mutations in telomerase.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7552-7.
51. Armanios, M.Y., et al., *Telomerase mutations in families with idiopathic pulmonary fibrosis.* N Engl J Med, 2007. **356**(13): p. 1317-26.
52. Uthman, E., M.D. *End-stage interstitial lung disease*. 2008 [cited; Available from: http://web2.airmail.net/uthman/specimens/images/honycomb.html.
53. Khalil, N., et al., *Environmental, inhaled and ingested causes of pulmonary fibrosis.* Toxicol Pathol, 2007. **35**(1): p. 86-96.
54. Baumgartner, K.B., et al., *Occupational and environmental risk factors for idiopathic pulmonary fibrosis: a multicenter case-control study. Collaborating Centers.* Am J Epidemiol, 2000. **152**(4): p. 307-15.
55. NHLBI. *What Causes Idiopathic Pulmonary Fibrosis?* 2008 [cited.
56. Zisman, D.A., et al., *Pulmonary fibrosis.* Methods Mol Med, 2005. **117**: p. 3-44.
57. Coultas, D.B., Hubbard, R., *Epidemiology of idiopathic pulmonary fibrosis*, in *Idiopathic pulmonary fibrosis, lung biology in health and disease*. 2004, Marcel Dekker: New York.
58. PFF. *Pulmonary Fibrosis Foundation*. 2008 [cited 2008; Available from: http://www.pulmonaryfibrosis.org/ipf.htm.
59. Hunninghake, G.W. and M.I. Schwarz, *Does current knowledge explain the pathogenesis of idiopathic pulmonary fibrosis? A perspective.* Proc Am Thorac Soc, 2007. **4**(5): p. 449-52.
60. Baumgartner, K.B., et al., *Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis.* Am J Respir Crit Care Med, 1997. **155**(1): p. 242-8.

61. Raghu, G., et al., *High prevalence of abnormal acid gastro-oesophageal reflux in idiopathic pulmonary fibrosis.* Eur Respir J, 2006. **27**(1): p. 136-42.
62. Ryu, J.H., et al., *Diagnosis of interstitial lung diseases.* Mayo Clin Proc, 2007. **82**(8): p. 976-86.
63. Daniels, C.E., E.S. Yi, and J.H. Ryu, *Autopsy findings in 42 consecutive patients with idiopathic pulmonary fibrosis.* Eur Respir J, 2008.
64. Zisman, D.A., et al., *Prediction of pulmonary hypertension in idiopathic pulmonary fibrosis.* Respir Med, 2007. **101**(10): p. 2153-9.
65. du Bois, R.M., *Genetic factors in pulmonary fibrotic disorders.* Semin Respir Crit Care Med, 2006. **27**(6): p. 581-8.
66. Bonanni, P.P., J.W. Frymoyer, and R.F. Jacox, *A Family Study of Idiopathic Pulmonary Fibrosis. a Possible Dysproteinemic and Genetically Determined Disease.* Am J Med, 1965. **39**: p. 411-21.
67. Javaheri, S., et al., *Idiopathic pulmonary fibrosis in monozygotic twins. The importance of genetic predisposition.* Chest, 1980. **78**(4): p. 591-4.
68. Solliday, N.H., et al., *Familial chronic interstitial pneumonia.* Am Rev Respir Dis, 1973. **108**(2): p. 193-204.
69. Bitterman, P.B., et al., *Familial idiopathic pulmonary fibrosis. Evidence of lung inflammation in unaffected family members.* N Engl J Med, 1986. **314**(21): p. 1343-7.
70. Hughes, E.W., *Familial Interstitial Pulmonary Fibrosis.* Thorax, 1964. **19**: p. 515-25.
71. Swaye, P., et al., *Familial Hamman-Rich syndrome. Report of eight cases.* Dis Chest, 1969. **55**(1): p. 7-12.
72. Lee, H.L., et al., *Familial idiopathic pulmonary fibrosis: clinical features and outcome.* Chest, 2005. **127**(6): p. 2034-41.
73. Chou, Y.H., et al., *The gene responsible for familial hypocalciuric hypercalcemia maps to chromosome 3q in four unrelated families.* Nat Genet, 1992. **1**(4): p. 295-300.
74. Auwerx, J., et al., *Defective host defence mechanisms in a family with hypocalciuric hypercalcaemia and coexisting interstitial lung disease.* Clin Exp Immunol, 1985. **62**(1): p. 57-64.
75. DePinho, R.A. and K.L. Kaplan, *The Hermansky-Pudlak syndrome. Report of three cases and review of pathophysiology and management considerations.* Medicine (Baltimore), 1985. **64**(3): p. 192-202.
76. Terry, R.D., W.M. Sperry, and B. Brodoff, *Adult lipidosis resembling Niemann-Pick's disease.* Am J Pathol, 1954. **30**(2): p. 263-85.
77. Schneider, E.L., et al., *Severe pulmonary involvement in adult Gaucher's disease. Report of three cases and review of the literature.* Am J Med, 1977. **63**(3): p. 475-80.
78. Rockah, R., et al., *Linkage disequilibrium of common Gaucher disease mutations with a polymorphic site in the pyruvate kinase (PKLR) gene.* Am J Med Genet, 1998. **78**(3): p. 233-6.
79. Levran, O., R.J. Desnick, and E.H. Schuchman, *Niemann-Pick disease: a frequent missense mutation in the acid sphingomyelinase gene of Ashkenazi Jewish type A and B patients.* Proc Natl Acad Sci U S A, 1991. **88**(9): p. 3748-52.
80. Oh, J., et al., *Positional cloning of a gene for Hermansky-Pudlak syndrome, a disorder of cytoplasmic organelles.* Nat Genet, 1996. **14**(3): p. 300-6.
81. Steele, M.P., et al., *Clinical and pathologic features of familial interstitial pneumonia.* Am J Respir Crit Care Med, 2005. **172**(9): p. 1146-52.

82. Kipling, D., *Telomerase: immortality enzyme or oncogene?* Nat Genet, 1995. **9**(2): p. 104-6.
83. Hodgson, U., et al., *ELMOD2 is a candidate gene for familial idiopathic pulmonary fibrosis.* Am J Hum Genet, 2006. **79**(1): p. 149-54.
84. Phelps, D.S., et al., *Increased surfactant protein-A levels in patients with newly diagnosed idiopathic pulmonary fibrosis.* Chest, 2004. **125**(2): p. 617-25.
85. Selman, M., et al., *Surfactant protein A and B genetic variants predispose to idiopathic pulmonary fibrosis.* Hum Genet, 2003. **113**(6): p. 542-50.
86. Vazquez de Lara, L., et al., *Surfactant components modulate fibroblast apoptosis and type I collagen and collagenase-1 expression.* Am J Physiol Lung Cell Mol Physiol, 2000. **279**(5): p. L950-7.
87. Zorzetto, M., et al., *Complement receptor 1 gene polymorphisms are associated with idiopathic pulmonary fibrosis.* Am J Respir Crit Care Med, 2003. **168**(3): p. 330-4.
88. Whyte, M., et al., *Increased risk of fibrosing alveolitis associated with interleukin-1 receptor antagonist and tumor necrosis factor-alpha gene polymorphisms.* Am J Respir Crit Care Med, 2000. **162**(2 Pt 1): p. 755-8.
89. Pardo, A. and M. Selman, *Matrix metalloproteases in aberrant fibrotic tissue remodeling.* Proc Am Thorac Soc, 2006. **3**(4): p. 383-8.
90. Alberts, B., et. al., *Cell junctions, cell adhesion, and the extracellular matrix*, in *Molecular Biology of the Cell*. 2002, Garland Science: New York, NY. p. 1090-1100.
91. Phan, S.H., *The myofibroblast in pulmonary fibrosis.* Chest, 2002. **122**(6 Suppl): p. 286S-289S.
92. Nagase, H., R. Visse, and G. Murphy, *Structure and function of matrix metalloproteinases and TIMPs.* Cardiovasc Res, 2006. **69**(3): p. 562-73.
93. Visse, R. and H. Nagase, *Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry.* Circ Res, 2003. **92**(8): p. 827-39.
94. Wojtowicz-Praga, S.M., R.B. Dickson, and M.J. Hawkins, *Matrix metalloproteinase inhibitors.* Invest New Drugs, 1997. **15**(1): p. 61-75.
95. Page-McCaw, A., A.J. Ewald, and Z. Werb, *Matrix metalloproteinases and the regulation of tissue remodelling.* Nat Rev Mol Cell Biol, 2007. **8**(3): p. 221-33.
96. Garcia-Alvarez, J., et al., *Tissue inhibitor of metalloproteinase-3 is up-regulated by transforming growth factor-beta1 in vitro and expressed in fibroblastic foci in vivo in idiopathic pulmonary fibrosis.* Exp Lung Res, 2006. **32**(5): p. 201-14.
97. Garcia-Alvarez, J., et al., *Membrane type-matrix metalloproteinases in idiopathic pulmonary fibrosis.* Sarcoidosis Vasc Diffuse Lung Dis, 2006. **23**(1): p. 13-21.
98. Selman, M. and A. Pardo, *Idiopathic pulmonary fibrosis: an epithelial/fibroblastic cross-talk disorder.* Respir Res, 2002. **3**: p. 3.
99. Selman, M. and A. Pardo, *Idiopathic pulmonary fibrosis: misunderstandings between epithelial cells and fibroblasts?* Sarcoidosis Vasc Diffuse Lung Dis, 2004. **21**(3): p. 165-72.
100. Selman, M. and A. Pardo, *Role of epithelial cells in idiopathic pulmonary fibrosis: from innocent targets to serial killers.* Proc Am Thorac Soc, 2006. **3**(4): p. 364-72.
101. Hinz, B., et al., *The myofibroblast: one function, multiple origins.* Am J Pathol, 2007. **170**(6): p. 1807-16.

102. Kisseleva, T. and D.A. Brenner, *Mechanisms of fibrogenesis.* Exp Biol Med (Maywood), 2008. **233**(2): p. 109-22.
103. Horowitz, J.C. and V.J. Thannickal, *Epithelial-mesenchymal interactions in pulmonary fibrosis.* Semin Respir Crit Care Med, 2006. **27**(6): p. 600-12.
104. Hashimoto, N., et al., *Bone marrow-derived progenitor cells in pulmonary fibrosis.* J Clin Invest, 2004. **113**(2): p. 243-52.
105. Kalluri, R. and E.G. Neilson, *Epithelial-mesenchymal transition and its implications for fibrosis.* J Clin Invest, 2003. **112**(12): p. 1776-84.
106. Willis, B.C., R.M. duBois, and Z. Borok, *Epithelial origin of myofibroblasts during fibrosis in the lung.* Proc Am Thorac Soc, 2006. **3**(4): p. 377-82.
107. Willis, B.C., et al., *Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-beta1: potential role in idiopathic pulmonary fibrosis.* Am J Pathol, 2005. **166**(5): p. 1321-32.
108. Keane, M.P., et al., *Inflammation and angiogenesis in fibrotic lung disease.* Semin Respir Crit Care Med, 2006. **27**(6): p. 589-99.
109. Studer, S.M. and N. Kaminski, *Towards systems biology of human pulmonary fibrosis.* Proc Am Thorac Soc, 2007. **4**(1): p. 85-91.
110. Moeller, A., et al., *The bleomycin animal model: A useful tool to investigate treatment options for idiopathic pulmonary fibrosis?* Int J Biochem Cell Biol, 2008. **40**(3): p. 362-82.
111. Bonniaud, P., et al., *Smad3 null mice develop airspace enlargement and are resistant to TGF-beta-mediated pulmonary fibrosis.* J Immunol, 2004. **173**(3): p. 2099-108.
112. Sime, P.J., et al., *Adenovector-mediated gene transfer of active transforming growth factor-beta1 induces prolonged severe fibrosis in rat lung.* J Clin Invest, 1997. **100**(4): p. 768-76.
113. Bristol-Myers, S., *Package insert for BLENOXANE® (bleomycin sulfate for injection, USP)*. 2006.
114. Borzone, G., et al., *Bleomycin-induced chronic lung damage does not resemble human idiopathic pulmonary fibrosis.* Am J Respir Crit Care Med, 2001. **163**(7): p. 1648-53.
115. Gauldie, J. and M. Kolb, *Animal models of pulmonary fibrosis: how far from effective reality?* Am J Physiol Lung Cell Mol Physiol, 2008. **294**(2): p. L151.
116. Mayne, R. and R.E. Burgeson, *Structure and function of collagen types*. Biology of extracellular matrix. 1987, Orlando: Academic Press. x, 317.
117. Varga, J., D. Brenner, and S.H. Phan, *Methods for measuring hydroxyproline and estimating in vivo rates of collagen synthesis and degradation*, in *Fibrosis research : methods and protocols*. 2005, Humana Press: Totowa, N.J. p. 189-221.
118. Wells, A.U., et al., *Bronchoalveolar lavage cellularity: lone cryptogenic fibrosing alveolitis compared with the fibrosing alveolitis of systemic sclerosis.* Am J Respir Crit Care Med, 1998. **157**(5 Pt 1): p. 1474-82.
119. Goh, N.S., et al., *Bronchoalveolar lavage cellular profiles in patients with systemic sclerosis-associated interstitial lung disease are not predictive of disease progression.* Arthritis Rheum, 2007. **56**(6): p. 2005-12.
120. Alejandre-Alcazar, M.A., et al., *TGF-beta signaling is dynamically regulated during the alveolarization of rodent and human lungs.* Dev Dyn, 2008. **237**(1): p. 259-69.
121. Sheppard, D., *Transforming growth factor beta: a central modulator of pulmonary and airway inflammation and fibrosis.* Proc Am Thorac Soc, 2006. **3**(5): p. 413-7.

122. Flanders, K.C., *Smad3 as a mediator of the fibrotic response.* Int J Exp Pathol, 2004. **85**(2): p. 47-64.
123. Mishra, L., R. Derynck, and B. Mishra, *Transforming growth factor-beta signaling in stem cells and cancer.* Science, 2005. **310**(5745): p. 68-71.
124. Lee, C.G., et al., *Early growth response gene 1-mediated apoptosis is essential for transforming growth factor beta1-induced pulmonary fibrosis.* J Exp Med, 2004. **200**(3): p. 377-89.
125. Derynck, R. and R.J. Akhurst, *Differentiation plasticity regulated by TGF-beta family proteins in development and disease.* Nat Cell Biol, 2007. **9**(9): p. 1000-4.
126. Bonewald, L.F., *Regulation and regulatory activities of transforming growth factor beta.* Crit Rev Eukaryot Gene Expr, 1999. **9**(1): p. 33-44.
127. Lawrence, D.A., *Transforming growth factor-beta: a general review.* Eur Cytokine Netw, 1996. **7**(3): p. 363-74.
128. Assoian, R.K., *Purification of type-beta transforming growth factor from human platelets.* Methods Enzymol, 1987. **146**: p. 153-63.
129. Assoian, R.K., et al., *Transforming growth factor-beta in human platelets. Identification of a major storage site, purification, and characterization.* J Biol Chem, 1983. **258**(11): p. 7155-60.
130. Khalil, N., et al., *TGF-beta 1, but not TGF-beta 2 or TGF-beta 3, is differentially present in epithelial cells of advanced pulmonary fibrosis: an immunohistochemical study.* Am J Respir Cell Mol Biol, 1996. **14**(2): p. 131-8.
131. Moreland, J.L., et al., *The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications.* BMC Bioinformatics, 2005. **6**: p. 21.
132. Rifkin, B.D.a.D.B., *TGF-[beta] bioavailability : latency, targeting, and activation*, in *The TGF- beta family*, K.M. Rik Derynck, Editor. 2008, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y. p. xiv, 1114.
133. Annes, J.P., J.S. Munger, and D.B. Rifkin, *Making sense of latent TGFbeta activation.* J Cell Sci, 2003. **116**(Pt 2): p. 217-24.
134. Saharinen, J., et al., *Latent transforming growth factor-beta binding proteins (LTBPs)--structural extracellular matrix proteins for targeting TGF-beta action.* Cytokine Growth Factor Rev, 1999. **10**(2): p. 99-117.
135. Annes, J.P., et al., *Integrin alphaVbeta6-mediated activation of latent TGF-beta requires the latent TGF-beta binding protein-1.* J Cell Biol, 2004. **165**(5): p. 723-34.
136. Taipale, J., et al., *Latent transforming growth factor-beta 1 and its binding protein are components of extracellular matrix microfibrils.* J Histochem Cytochem, 1996. **44**(8): p. 875-89.
137. Koli, K., et al., *Sequential deposition of latent TGF-beta binding proteins (LTBPs) during formation of the extracellular matrix in human lung fibroblasts.* Exp Cell Res, 2005. **310**(2): p. 370-82.
138. Koli, K., et al., *Latency, activation, and binding proteins of TGF-beta.* Microsc Res Tech, 2001. **52**(4): p. 354-62.
139. Gauldie, J., et al., *Smad3 signaling involved in pulmonary fibrosis and emphysema.* Proc Am Thorac Soc, 2006. **3**(8): p. 696-702.
140. Brown, K.A., J.A. Pietenpol, and H.L. Moses, *A tale of two proteins: Differential roles and regulation of Smad2 and Smad3 in TGF-beta signaling.* J Cell Biochem, 2007. **101**(1): p. 9-33.

141. Gu, L., et al., *Effect of TGF-beta/Smad signaling pathway on lung myofibroblast differentiation.* Acta Pharmacol Sin, 2007. **28**(3): p. 382-91.
142. Josso, N. and N. di Clemente, *Serine/threonine kinase receptors and ligands.* Curr Opin Genet Dev, 1997. **7**(3): p. 371-7.
143. Derynck, R. and Y.E. Zhang, *Smad-dependent and Smad-independent pathways in TGF-beta family signalling.* Nature, 2003. **425**(6958): p. 577-84.
144. Miyazono, K., P. ten Dijke, and C.H. Heldin, *TGF-beta signaling by Smad proteins.* Adv Immunol, 2000. **75**: p. 115-57.
145. Itoh, S. and P. ten Dijke, *Negative regulation of TGF-beta receptor/Smad signal transduction.* Curr Opin Cell Biol, 2007. **19**(2): p. 176-84.
146. Afrakhte, M., et al., *Induction of inhibitory Smad6 and Smad7 mRNA by TGF-beta family members.* Biochem Biophys Res Commun, 1998. **249**(2): p. 505-11.
147. Nakao, A., et al., *Identification of Smad7, a TGFbeta-inducible antagonist of TGF-beta signalling.* Nature, 1997. **389**(6651): p. 631-5.
148. Zhang, Y., et al., *Regulation of Smad degradation and activity by Smurf2, an E3 ubiquitin ligase.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 974-9.
149. Wang, T., *The 26S proteasome system in the signaling pathways of TGF-beta superfamily.* Front Biosci, 2003. **8**: p. d1109-27.
150. Suzuki, C., et al., *Smurf1 regulates the inhibitory activity of Smad7 by targeting Smad7 to the plasma membrane.* J Biol Chem, 2002. **277**(42): p. 39919-25.
151. Massague, J., *How cells read TGF-beta signals.* Nat Rev Mol Cell Biol, 2000. **1**(3): p. 169-78.
152. Mizuide, M., et al., *Two short segments of Smad3 are important for specific interaction of Smad3 with c-Ski and SnoN.* J Biol Chem, 2003. **278**(1): p. 531-6.
153. Lin, X., Chen, Y-G., Feng, X-H., *Transcriptional control via SMADs*, in *The TGF - beta family*, R. Derynck and K. Miyazono, Editors. 2008, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y. p. xiv, 1114.
154. Leask, A. and D.J. Abraham, *TGF-beta signaling and the fibrotic response.* FASEB J, 2004. **18**(7): p. 816-27.
155. Xu, G.P., et al., *The Effect of TGF-beta1 and SMAD7 gene transfer on the phenotypic changes of rat alveolar epithelial cells.* Cell Mol Biol Lett, 2007.
156. Sato, M., et al., *Targeted disruption of TGF-beta1/Smad3 signaling protects against renal tubulointerstitial fibrosis induced by unilateral ureteral obstruction.* J Clin Invest, 2003. **112**(10): p. 1486-94.
157. Kapanci, Y., et al., *Cytoskeletal protein modulation in pulmonary alveolar myofibroblasts during idiopathic pulmonary fibrosis. Possible role of transforming growth factor beta and tumor necrosis factor alpha.* Am J Respir Crit Care Med, 1995. **152**(6 Pt 1): p. 2163-9.
158. Kapanci, Y., et al., *Phenotypic modulation of alveolar myofibroblasts in transplanted human lungs.* Mod Pathol, 1997. **10**(11): p. 1134-42.
159. Broekelmann, T.J., et al., *Transforming growth factor beta 1 is present at sites of extracellular matrix gene expression in human pulmonary fibrosis.* Proc Natl Acad Sci U S A, 1991. **88**(15): p. 6642-6.
160. Khalil, N., et al., *Biological effects of transforming growth factor-beta(1) in idiopathic pulmonary fibrosis may be regulated by the activation of latent transforming growth*

factor-beta(1) and the differential expression of transforming growth factor-beta receptors. Chest, 2001. **120**(1 Suppl): p. 48S.
161. Khalil, N., et al., *Increased production and immunohistochemical localization of transforming growth factor-beta in idiopathic pulmonary fibrosis.* Am J Respir Cell Mol Biol, 1991. **5**(2): p. 155-62.
162. Khalil, N., et al., *Regulation of alveolar macrophage transforming growth factor-beta secretion by corticosteroids in bleomycin-induced pulmonary inflammation in the rat.* J Clin Invest, 1993. **92**(4): p. 1812-8.
163. Zhao, J., et al., *Smad3 deficiency attenuates bleomycin-induced pulmonary fibrosis in mice.* Am J Physiol Lung Cell Mol Physiol, 2002. **282**(3): p. L585-93.
164. Zhao, Y. and D.A. Geverd, *Regulation of Smad3 expression in bleomycin-induced pulmonary fibrosis: a negative feedback loop of TGF-beta signaling.* Biochem Biophys Res Commun, 2002. **294**(2): p. 319-23.
165. Xavier, S., et al., *Amelioration of radiation-induced fibrosis: inhibition of transforming growth factor-beta signaling by halofuginone.* J Biol Chem, 2004. **279**(15): p. 15167-76.
166. Nakao, A., et al., *Transient gene transfer and expression of Smad7 prevents bleomycin-induced lung fibrosis in mice.* J Clin Invest, 1999. **104**(1): p. 5-11.
167. Venkatesan, N., L. Pini, and M.S. Ludwig, *Changes in Smad expression and subcellular localization in bleomycin-induced pulmonary fibrosis.* Am J Physiol Lung Cell Mol Physiol, 2004. **287**(6): p. L1342-7.
168. Kitano, H., *Foundations of systems biology*. 2001, Cambridge, Mass.: MIT Press. vii, 297.
169. Bringmann, P., *Systems biology : applications and perspectives*. Ernst Schering Research Foundation workshop, 61. 2007, Berlin ; New York: Springer. xiii, 172.
170. Konopka, A.K., *Systems biology : principles, methods, and concepts*. 2007, Boca Raton: CRC Press/Taylor & Francis. 244.
171. Davidson, E.H., *Genomic regulatory systems : development and evolution*. 2001, San Diego: Academic Press. xii, 261.
172. Kremling, A. and J. Saez-Rodriguez, *Systems biology--an engineering perspective.* J Biotechnol, 2007. **129**(2): p. 329-51.
173. Tuncay, K., et al., *Transcriptional regulatory networks via gene ontology and expression data.* In Silico Biol, 2007. **7**(1): p. 21-34.
174. Huang, Y., et al., *Systematic discovery of functional modules and context-specific functional annotation of human genome.* Bioinformatics, 2007. **23**(13): p. i222-9.
175. Tornow, S. and H.W. Mewes, *Functional modules by relating protein interaction networks and gene expression.* Nucleic Acids Res, 2003. **31**(21): p. 6283-9.
176. Bhalla, U.S. and R. Iyengar, *Functional modules in biological signalling networks.* Novartis Found Symp, 2001. **239**: p. 4-13; discussion 13-5, 45-51.
177. Wu, H., et al., *Prediction of functional modules based on comparative genome analysis and Gene Ontology application.* Nucleic Acids Res, 2005. **33**(9): p. 2822-37.
178. Luo, F., et al., *Application of random matrix theory to microarray data for discovering functional gene modules.* Phys Rev E Stat Nonlin Soft Matter Phys, 2006. **73**(3 Pt 1): p. 031924.
179. Lubovac, Z., J. Gamalielsson, and B. Olsson, *Combining functional and topological properties to identify core modules in protein interaction networks.* Proteins, 2006. **64**(4): p. 948-59.

180. Bower, J.M. and H. Bolouri, *Computational modeling of genetic and biochemical networks*. 2001, Cambridge, Mass.: MIT Press. xx, 336 p., [28] p. plates.
181. Yuh, C.H., H. Bolouri, and E.H. Davidson, *Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control.* Development, 2001. **128**(5): p. 617-29.
182. Liu, E.T., *Systems biology, integrative biology, predictive biology.* Cell, 2005. **121**(4): p. 505-6.
183. Aderem, A., *Systems biology: its practice and challenges.* Cell, 2005. **121**(4): p. 511-3.
184. Friboulet, A. and D. Thomas, *Systems Biology-an interdisciplinary approach.* Biosens Bioelectron, 2005. **20**(12): p. 2404-7.
185. Wendl, M.C., et al., *Automated sequence preprocessing in a large-scale sequencing environment.* Genome Res, 1998. **8**(9): p. 975-84.
186. Drexler, K.E., *Molecular nanomachines: physical principles and implementation strategies.* Annu Rev Biophys Biomol Struct, 1994. **23**: p. 377-405.
187. Freitas, R.A., Jr., *The future of nanofabrication and molecular scale devices in nanomedicine.* Stud Health Technol Inform, 2002. **80**: p. 45-59.
188. Martin, W.J. and R.M. Walmsley, *Vision assisted robotics and tape technology in the life-science laboratory: applications to genome analysis.* Biotechnology (N Y), 1990. **8**(12): p. 1258-62.
189. Putman, E., *Entering the small, small world of nanotechnology.* Biomed Instrum Technol, 2002. **36**(6): p. 375-81.
190. Stevenson, R., *Microrobotics in biotechnology.* Am Biotechnol Lab, 1990. **8**(1): p. 6.
191. Terstegge, S., et al., *Automated maintenance of embryonic stem cell cultures.* Biotechnol Bioeng, 2007. **96**(1): p. 195-201.
192. Whitesides, G.M., *The once and future nanomachine.* Sci Am, 2001. **285**(3): p. 78-83.
193. Lewin, R., *National Academy looks at human genome project, sees progress.* Science, 1987. **235**(4790): p. 747-8.
194. Collins, F.S. and M.K. Mansoura, *The Human Genome Project. Revealing the shared inheritance of all humankind.* Cancer, 2001. **91**(1 Suppl): p. 221-5.
195. Koski, C.A., *The Human Genome Project: an examination of its challenge to the technological imperative.* New Genet Soc, 2005. **24**(3): p. 265-81.
196. *International consortium completes human genome project.* Pharmacogenomics, 2003. **4**(3): p. 241.
197. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-90.
198. Dietrich, W.F., *The origin and implications of the Human Genome Project: scientific overview.* Natl Cathol Bioeth Q, 2001. **1**(4): p. 489-95.
199. Ensembl. *Ensembl Human (www.ensembl.org)*. 2008 [cited; Available from: http://www.ensembl.org/Homo_sapiens/index.html.
200. Cotton, R.G., et al., *Recommendations of the 2006 Human Variome Project meeting.* Nat Genet, 2007. **39**(4): p. 433-6.
201. Gerstein, M.B., et al., *What is a gene, post-ENCODE? History and updated definition.* Genome Res, 2007. **17**(6): p. 669-81.
202. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

203. Latchman, D.S., *Eukaryotic transcription factors*. 5th ed. 2008, Amsterdam ; Boston: Elsevier/Academic Press. xxviii, 471.
204. Watson, J.D., *Molecular biology of the gene*. 6th ed. 2008, San Francisco
Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ;
Cold Spring Harbor Laboratory Press. xxxii, 841.
205. White, R.J., *Gene transcription: mechanisms and control*. 2001, Oxford ; Malden, MA: Blackwell Science. xii, 273.
206. Alberts, B., et. al., *How Cells Read the Genome: From DNA to Protein*, in *Molecular Biology of the Cell*, B. Alberts, et. al., Editor. 2008, Garland Science: New York, NY.
207. Mahony, S., P.E. Auron, and P.V. Benos, *DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies.* PLoS Comput Biol, 2007. **3**(3): p. e61.
208. Qian, Z., Y.D. Cai, and Y. Li, *Automatic transcription factor classifier based on functional domain composition.* Biochem Biophys Res Commun, 2006. **347**(1): p. 141-4.
209. Mahony, S., P.E. Auron, and P.V. Benos, *Inferring protein-DNA dependencies using motif alignments and mutual information.* Bioinformatics, 2007. **23**(13): p. i297-304.
210. Macisaac, K.D., et al., *A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data.* Bioinformatics, 2006. **22**(4): p. 423-9.
211. Buck, M.J. and J.D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.* Genomics, 2004. **83**(3): p. 349-60.
212. Euskirchen, G.M., et al., *Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies.* Genome Res, 2007. **17**(6): p. 898-909.
213. Johnson, D.S., et al., *Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets.* Genome Res, 2008.
214. Weinmann, A.S. and P.J. Farnham, *Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation.* Methods, 2002. **26**(1): p. 37-47.
215. Ji, H., S.A. Vokes, and W.H. Wong, *A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors.* Nucleic Acids Res, 2006. **34**(21): p. e146.
216. Ng, P., C.L. Wei, and Y. Ruan, *Paired-end diTagging for transcriptome and genome analysis.* Curr Protoc Mol Biol, 2007. **Chapter 21**: p. Unit 21 12.
217. Kim, J. and V.R. Iyer, *Identifying chromosomal targets of DNA-binding proteins by Sequence Tag Analysis of Genomic Enrichment (STAGE).* Curr Protoc Mol Biol, 2005. **Chapter 21**: p. Unit 21 10.
218. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
219. NEB. *New England Biolabs, Inc.* 2008 [cited; Available from: http://www.neb.com.
220. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
221. Wells, J. and P.J. Farnham, *Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation.* Methods, 2002. **26**(1): p. 48-56.

222. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
223. Lee, T.I., S.E. Johnstone, and R.A. Young, *Chromatin immunoprecipitation and microarray-based analysis of protein location.* Nat Protoc, 2006. **1**(2): p. 729-48.
224. Metz, B., et al., *Identification of formaldehyde-induced modifications in proteins: reactions with model peptides.* J Biol Chem, 2004. **279**(8): p. 6235-43.
225. Orlando, V., *Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation.* Trends Biochem Sci, 2000. **25**(3): p. 99-104.
226. Orlando, V., H. Strutt, and R. Paro, *Analysis of chromatin structure by in vivo formaldehyde cross-linking.* Methods, 1997. **11**(2): p. 205-14.
227. Toth, J. and M.D. Biggin, *The specificity of protein-DNA crosslinking by formaldehyde: in vitro and in drosophila embryos.* Nucleic Acids Res, 2000. **28**(2): p. e4.
228. Suslick, K.S., *Ultrasound : its chemical, physical, and biological effects*. 1988, New York, N.Y.: VCH Publishers. xiii, 336 p.
229. Kim, T.H., L.O. Barrera, and B. Ren, *ChIP-chip for genome-wide analysis of protein binding in mammalian cells.* Curr Protoc Mol Biol, 2007. **Chapter 21**: p. Unit 21 13.
230. Harlow, E. and D. Lane, *Using antibodies : a laboratory manual*. 1999, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. xiv, 495.
231. Sjobring, U., L. Bjorck, and W. Kastern, *Streptococcal protein G. Gene structure and protein binding properties.* J Biol Chem, 1991. **266**(1): p. 399-405.
232. King, B.F. and B.J. Wilkinson, *Binding of human immunoglobulin G to protein A in encapsulated Staphylococcus aureus.* Infect Immun, 1981. **33**(3): p. 666-72.
233. Agilent, *Agilent Mammalian ChIP-on-chip Protocol Version 9.2, May 2007*. 2007.
234. Upstate, *ChIP Kit Protocol catalog number 17-295.* 2008.
235. Ren, B. and B.D. Dynlacht, *Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors.* Methods Enzymol, 2004. **376**: p. 304-15.
236. Pfeifer, G.P., et al., *Genomic sequencing and methylation analysis by ligation mediated PCR.* Science, 1989. **246**(4931): p. 810-3.
237. Ausubel, F.M. and Wiley InterScience (Online service), *Enzymatic Manipulation of DNA and RNA*, in *Current protocols in molecular biology*. 2001, J. Wiley: New York.
238. Bakel, H.V., et al., *Improved genome-wide localization by ChIP-chip using double-round T7 RNA polymerase-based amplification.* Nucleic Acids Res, 2008.
239. Keohavong, P. and W.G. Thilly, *Fidelity of DNA polymerases in DNA amplification.* Proc Natl Acad Sci U S A, 1989. **86**(23): p. 9253-7.
240. Bracho, M.A., A. Moya, and E. Barrio, *Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity.* J Gen Virol, 1998. **79 (Pt 12)**: p. 2921-8.
241. Dietrich, J., et al., *PCR performance of the highly thermostable proof-reading B-type DNA polymerase from Pyrococcus abyssi.* FEMS Microbiol Lett, 2002. **217**(1): p. 89-94.
242. Cline, J., J.C. Braman, and H.H. Hogrefe, *PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases.* Nucleic Acids Res, 1996. **24**(18): p. 3546-51.
243. Telenius, H., et al., *Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer.* Genomics, 1992. **13**(3): p. 718-25.
244. Hittelman, A., et al., *Evaluation of whole genome amplification protocols for array and oligonucleotide CGH.* Diagn Mol Pathol, 2007. **16**(4): p. 198-206.

245. Barker, D.L., et al., *Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel.* Genome Res, 2004. **14**(5): p. 901-7.
246. Park, J.W., et al., *Comparing whole-genome amplification methods and sources of biological samples for single-nucleotide polymorphism genotyping.* Clin Chem, 2005. **51**(8): p. 1520-3.
247. Liu, C.L., S.L. Schreiber, and B.E. Bernstein, *Development and validation of a T7 based linear amplification for genomic DNA.* BMC Genomics, 2003. **4**(1): p. 19.
248. Hughes, S., et al., *The use of whole genome amplification in the study of human disease.* Prog Biophys Mol Biol, 2005. **88**(1): p. 173-89.
249. Manduchi, E., et al., *Comparison of different labeling methods for two-channel high-density microarray experiments.* Physiol Genomics, 2002. **10**(3): p. 169-79.
250. Capaldi, S., R.C. Getts, and S.D. Jayasena, *Signal amplification through nucleotide extension and excision on a dendritic DNA platform.* Nucleic Acids Res, 2000. **28**(7): p. E21.
251. Leiske, D.L., et al., *A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray.* BMC Genomics, 2006. **7**: p. 72.
252. Boyer, L.A., et al., *Core transcriptional regulatory circuitry in human embryonic stem cells.* Cell, 2005. **122**(6): p. 947-56.
253. NimbleGen. *NimbleGen Chromatin Immunoprecipitation products*. 2008 2008 [cited; Available from: http://www.nimblegen.com/products/chip/index.html.
254. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**: p. 829-836.
255. Trayhurn, P., *Northern blotting.* Proc Nutr Soc, 1996. **55**(1B): p. 583-9.
256. Alwine, J.C., et al., *Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper.* Methods Enzymol, 1979. **68**: p. 220-42.
257. Alwine, J.C., D.J. Kemp, and G.R. Stark, *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5350-4.
258. Kroczek, R.A. and E. Siebert, *Optimization of northern analysis by vacuum-blotting, RNA-transfer visualization, and ultraviolet fixation.* Anal Biochem, 1990. **184**(1): p. 90-5.
259. Maskos, U., *A novel method of nucleic acid sequence analysis*. 1991, Oxford University.
260. Maskos, U. and E.M. Southern, *Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ.* Nucleic Acids Res, 1992. **20**(7): p. 1679-84.
261. Maskos, U. and E.M. Southern, *Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation.* Nucleic Acids Res, 1992. **20**(7): p. 1675-8.
262. Maskos, U. and E.M. Southern, *A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support.* Nucleic Acids Res, 1993. **21**(20): p. 4663-9.
263. Maskos, U. and E.M. Southern, *A novel method for the analysis of multiple sequence variants by hybridisation to oligonucleotides.* Nucleic Acids Res, 1993. **21**(9): p. 2267-8.

264. Southern, E.M. and U. Maskos, *Parallel synthesis and analysis of large numbers of related chemical compounds: applications to oligonucleotides.* J Biotechnol, 1994. **35**(2-3): p. 217-27.
265. Southern, E.M., U. Maskos, and J.K. Elder, *Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models.* Genomics, 1992. **13**(4): p. 1008-17.
266. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
267. Schena, M., et al., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.* Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10614-9.
268. Lennon, G., et al., *The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression.* Genomics, 1996. **33**(1): p. 151-2.
269. Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, *dbEST--database for "expressed sequence tags".* Nat Genet, 1993. **4**(4): p. 332-3.
270. Bowtell, D.D., *Options available--from start to finish--for obtaining expression data by microarray.* Nat Genet, 1999. **21**(1 Suppl): p. 25-32.
271. Handley, D., et al., *Evidence of systematic expressed sequence tag IMAGE clone cross-hybridization on cDNA microarrays.* Genomics, 2004. **83**(6): p. 1169-75.
272. Rouillard, J.M., M. Zuker, and E. Gulari, *OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.* Nucleic Acids Res, 2003. **31**(12): p. 3057-62.
273. Chomczynski, P. and N. Sacchi, *Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.* Anal Biochem, 1987. **162**(1): p. 156-9.
274. Van Gelder, R.N., et al., *Amplified RNA synthesized from limited quantities of heterogeneous cDNA.* Proc Natl Acad Sci U S A, 1990. **87**(5): p. 1663-7.
275. Stahlberg, A., M. Kubista, and M. Pfaffl, *Comparison of reverse transcriptases in gene expression analysis.* Clin Chem, 2004. **50**(9): p. 1678-80.
276. Wick, L.M., et al., *On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes.* Nucleic Acids Res, 2006. **34**(3): p. e26.
277. Shapiro, H.M., *Excitation and Emission Spectra of Common Dyes*, in *Current Protocols in Cytometry*. 2003, John Wiley & Sons, Inc.
278. Benjamini, Y., Hochberg, Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser. B., 1995. **57**: p. 289-300.
279. Rahnenfuhrer, J., *Clustering algorithms and other exploratory methods for microarray data analysis.* Methods Inf Med, 2005. **44**(3): p. 444-8.
280. Xu, R. and D. Wunsch, 2nd, *Survey of clustering algorithms.* IEEE Trans Neural Netw, 2005. **16**(3): p. 645-78.
281. Witten, I.H. and E. Frank, *Data mining : practical machine learning tools and techniques*. 2nd ed. Morgan Kaufmann series in data management systems. 2005, Amsterdam ; Boston, MA: Morgan Kaufman. xxxi, 525.
282. Mitchell, T.M., *Machine Learning*. 1997, New York: McGraw-Hill. xvii, 414.
283. Asyali, M.H., *Gene expression profile class prediction using linear Bayesian classifiers.* Comput Biol Med, 2007. **37**(12): p. 1690-9.

284. Michailidis, G. and K. Shedden, *The application of rule-based methods to class prediction problems in genomics.* J Comput Biol, 2003. **10**(5): p. 689-98.
285. Bura, E. and R.M. Pfeiffer, *Graphical methods for class prediction using dimension reduction techniques on DNA microarray data.* Bioinformatics, 2003. **19**(10): p. 1252-8.
286. Radmacher, M.D., L.M. McShane, and R. Simon, *A paradigm for class prediction using gene expression profiles.* J Comput Biol, 2002. **9**(3): p. 505-11.
287. Shi, L., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **24**(9): p. 1151-61.
288. Kaplan, T.F., N.,, *Model-Based Analysis of High-Resolution Chromatin Immunoprecipitation Data*, in *Technical Report*. 2006.
289. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
290. Ernst, J. and Z. Bar-Joseph, *STEM: a tool for the analysis of short time series gene expression data.* BMC Bioinformatics, 2006. **7**: p. 191.
291. Ernst, J., G.J. Nau, and Z. Bar-Joseph, *Clustering short time series gene expression data.* Bioinformatics, 2005. **21 Suppl 1**: p. i159-68.
292. Andrews, N.C. and D.V. Faller, *A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells.* Nucleic Acids Res, 1991. **19**(9): p. 2499.
293. Jinnin, M., H. Ihn, and K. Tamaki, *Characterization of SIS3, a novel specific inhibitor of Smad3, and its effect on transforming growth factor-beta1-induced extracellular matrix expression.* Mol Pharmacol, 2006. **69**(2): p. 597-607.
294. Monaghan, A.P., et al., *Postimplantation expression patterns indicate a role for the mouse forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm.* Development, 1993. **119**(3): p. 567-78.
295. Whitsett, J.A. and T.E. Weaver, *Hydrophobic surfactant proteins in lung function and disease.* N Engl J Med, 2002. **347**(26): p. 2141-8.
296. Nogee, L.M., *Alterations in SP-B and SP-C expression in neonatal lung disease.* Annu Rev Physiol, 2004. **66**: p. 601-23.
297. Nesslein, L.L., et al., *Partial SP-B deficiency perturbs lung function and causes air space abnormalities.* Am J Physiol Lung Cell Mol Physiol, 2005. **288**(6): p. L1154-61.
298. Melton, K.R., et al., *SP-B deficiency causes respiratory failure in adult mice.* Am J Physiol Lung Cell Mol Physiol, 2003. **285**(3): p. L543-9.
299. Lawson, W.E., et al., *Increased and prolonged pulmonary fibrosis in surfactant protein C-deficient mice following intratracheal bleomycin.* Am J Pathol, 2005. **167**(5): p. 1267-77.
300. Minoo, P., et al., *SMAD3 prevents binding of NKX2.1 and FOXA1 to the SpB promoter through its MH1 and MH2 domains.* Nucleic Acids Res, 2008. **36**(1): p. 179-88.
301. Li, C., et al., *Transforming growth factor-beta inhibits pulmonary surfactant protein B gene transcription through SMAD3 interactions with NKX2.1 and HNF-3 transcription factors.* J Biol Chem, 2002. **277**(41): p. 38399-408.
302. Zhou, L., et al., *Thyroid transcription factor-1, hepatocyte nuclear factor-3beta, surfactant protein B, C, and Clara cell secretory protein in developing mouse lung.* J Histochem Cytochem, 1996. **44**(10): p. 1183-93.

303. Bohinski, R.J., R. Di Lauro, and J.A. Whitsett, *The lung-specific surfactant protein B gene promoter is a target for thyroid transcription factor 1 and hepatocyte nuclear factor 3, indicating common factors for organ-specific gene expression along the foregut axis.* Mol Cell Biol, 1994. **14**(9): p. 5671-81.
304. King, T.E., Jr., et al., *Idiopathic pulmonary fibrosis: relationship between histopathologic features and mortality.* Am J Respir Crit Care Med, 2001. **164**(6): p. 1025-32.
305. Garcia, C.K., W.E. Wright, and J.W. Shay, *Human diseases of telomerase dysfunction: insights into tissue aging.* Nucleic Acids Res, 2007. **35**(22): p. 7406-16.
306. Armanios, M., et al., *Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 15960-4.
307. Utz, J.P., et al., *Usual interstitial pneumonia complicating dyskeratosis congenita.* Mayo Clin Proc, 2005. **80**(6): p. 817-21.
308. Liu, T., et al., *Telomerase activity is required for bleomycin-induced pulmonary fibrosis in mice.* J Clin Invest, 2007. **117**(12): p. 3800-9.
309. Schissel, S.L. and M.D. Layne, *Telomerase, myofibroblasts, and pulmonary fibrosis.* Am J Respir Cell Mol Biol, 2006. **34**(5): p. 520-2.
310. Fridlender, Z.G., et al., *Telomerase activity in bleomycin-induced epithelial cell apoptosis and lung fibrosis.* Eur Respir J, 2007. **30**(2): p. 205-13.
311. Ducrest, A.L., et al., *Regulation of the human telomerase reverse transcriptase gene.* Oncogene, 2002. **21**(4): p. 541-52.
312. Cong, Y.S., W.E. Wright, and J.W. Shay, *Human telomerase and its regulation.* Microbiol Mol Biol Rev, 2002. **66**(3): p. 407-25, table of contents.
313. Yim, H.W., et al., *Smoking is associated with increased telomerase activity in short-term cultures of human bronchial epithelial cells.* Cancer Lett, 2007. **246**(1-2): p. 24-33.