UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Reprogramming scleroderma-specific and normal fibroblast gene subsets using single versus multiple transcription factors

Permalink

https://escholarship.org/uc/item/0r08856j

Author

Van Buren, Tyler Martin

Publication Date

2010

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Reprogramming Scleroderma-specific and Normal Fibroblast Gene Subsets Using Single Versus Multiple Transcription Factors

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Biology

by

Tyler Martin Van Buren

Committee in charge:

Professor Benjamin D. Yu, Chair
Professor Colin C. Jamora, Co-Chair
Professor Steven P. Briggs

2010

Copyright©

Tyler Martin Van Buren, 2010

All rights reserved.

The Thesis of Tyler Martin Van Buren is approved and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST of ABBREVIATIONS

BJ	Dermal Fibroblast Cell Line Derived from Foreskin
cDNA	Complementary DNA
CRM	Cis Regulatory Module
cRNA	Complementary Ribonucleic Acid
cTFBS	Cluster of Transcription Factor Binding Sites
DNA	DeoxyriboNucleic Acid
EI	Enhancer Identification
ES	Embryonic Stem Cell
HaCaT	Keratinocyte Cell Line
HCT	Homotypic Cluster of Transcription Factor Binding Sites
hES	Human Embryonic Stem Cell
iPS	Induced Pluripotent State
mRNA	Messenger Ribonucleic Acid
PBS	Phosphate Buffered Saline
qPCR	Quantitative Polymerase Chain Reaction
RE	Regulatory Element
RNA	Ribonucleic Acid
SRE	Synonymous Regulatory Element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
UTR	Untranslated Region
ΔΔCt	Delta Delta Threshold Cycle

LIST OF FIGURES and TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

ACKNOWLEDGEMENTS

I would like to begin by acknowledging my faculty advisor, Dr. Ben Yu. I am grateful for having the opportunity to conduct research in his lab. The training I have received and the scientific methods that I have learned far surpass what I knew before I joined. Ben has allowed me to gain exposure to all of the projects in his lab and apply the scientific method to all the work that I perform. That is a skill that I will carry throughout my career. I truly appreciate Ben's unique perspective on scientific research and the thought provoking scientific questions that he asks. I have a great deal of respect for Ben and I thank him for all of the wonderful memories I have compiled over the past couple of years.

I would like to thank the members of my committee for their support and advice throughout the course of my project. They are individuals who I highly respect and their experience in scientific research is invaluable. My committee is a great representation of the professors that I have had a chance to learn from throughout my education at UCSD.

I would also like to thank the past and present members of the Yu Lab for their support, discussions, and entertainment. In particular, I would like to thank Sangyoon Han who brought me up to speed on all

of the scientific methods needed for my project. Also, I would like to thank Chris Cowing-Zitron for his endless support and help with anything bioinformatic. My project would not have been as successful without their help and willingness to teach me.

Above all, I would like to thank my parents, Mike and Kathy Burns, and my brother and sister, Justin and Kaitlyn Van Buren. Also, I would like to thank my fiancée and soon-to-be wife, Tljana Nargilic, for her positive attitude and patience with me during this program. She compliments me very well and helps me to perform at a higher level day-in and day-out. My family has had an instrumental role in my development as an individual and they have helped foster my interest in science since grade school. Nothing that I do would be as enjoyable or worthwhile without the help of my family.

VITA

2004-2005 Virology Research Intern, U. S. Department of Agriculture

2007 Oncology Research Assistant, Sanford | Burnham Institute

2009 Bachelor of Science, University of California, San Diego

2009-2010 Teaching Assistant, Department of Biology
University of California, San Diego
UCSD *TA Excellence in Teaching Award*

2010 Master of Science, University of California, San Diego

PUBLICATIONS

R. L. Jordan, M. A. Guaragna, and T. Van Buren, "First Report of a New Potyvirus, Tricyrtis virus Y, and *Lily virus X*, a Potexvirus, in *Tricyrtis formosana* in the United States," Plant Disease Abstract, Beltsville, MD (April 2008, Vol. 92, Num. 4, Pg. 648).

FIELDS OF STUDY

Major Field: Developmental Biology, Virology

Studies in Stem Cell Concepts
Professor Steven P. Briggs

Studies in Animal Virology
Professor Deborah H. Spector

ABSTRACT OF THE THESIS

Reprogramming Scleroderma-specific and Normal Fibroblast Gene Subsets Using Single Versus Multiple Transcription Factors

by

Tyler Martin Van Buren

Master of Science in Biology

University of California, San Diego, 2010

Professor Benjamin D. Yu, Chair

Professor Colin C. Jamora, Co-Chair

The recent successes of reprogramming iPS cells and other cell types offer a novel platform that can be used to treat numerous diseases. These studies reveal that transcription factors can be used to alter gene expression, thereby reprogramming a new cellular phenotype that is eventually maintained by endogenous transcription

factors. With this information, we believe it is possible to reprogram the altered phenotype of a diseased cell back to its normal state. Scleroderma in particular, is a non-curable scarring disease with a distinct phenotype that can be defined by its unique gene expression pattern.

Here we report the successful induction of a subset of scleroderma-specific genes using predicted and endogenous transcription factors. Microarray data of scleroderma patients was analyzed finding 165 differentially expressed genes. Both DiRE and SynoR were bioinformatic tools employed to determine upstream regulators and downstream gene targets, respectively. Gain of function experiments were carried out as transcription factors were transiently transfected using overexpression vectors in multiple cell types and gene expression was measured via qPCR.

22% of total scleroderma-specific downstream targets were differentially expressed using predicted transcription factors. Additionally, 33% of predicted downstream targets were differentially expressed using endogenously expressed transcription factors. Our data suggest that primary sequence analysis of clusters of transcription

factors binding sites in promoters is more predictive than evolutionary and conservation-based approaches alone. Our study represents a novel approach to treating and even curing disease by way of cellular reprogramming.

CHAPTER 1

Introduction to Reprogramming Disease

Transcription factors play a major role in defining cell identity. Even in complex cell types, the addition of the correct combination of transcription factors can in some cases induce a new cell identity. Recently, the transcription factors that define the embryonic stem cell have been discovered. Subsequently, investigators have found that introduction of four core transcription factors (*Oct4, Sox2, Myc, and Klf4*) can induce a pluripotent stem cell phenotype in fibroblasts (Takahashi and Yamanaka, 2006; Wernig, 2007; Yu, 2007). Surprisingly, in this case, an endogenous cellular program takes over and maintains the new identity, thus making iPS (induced pluripotent stem cells) cells independent of externally added transcription factors.

With the discovery of iPS cells, a serious following of scientists not only to understand reprogramming, but to attempt to reprogram other cell lineages, has ensued. If it is possible to take an otherwise "committed" cell type such as a fibroblast and revert it to a non-committed iPS cell, then it must be possible to force the differentiation of iPS cells into other cell tissue types. That is in fact true as Choi et al. generated CD34+CD43+ hematopoietic progenitors and CD31+CD43– endothelial cells in co-culture with the OP9 differentiation system. Around the same time, Zhang and colleagues reported a highly

efficient approach to inducing ES cells and iPS cells into pancreatic cells. Further down the road, this very method could lead to the curing of diabetes, a disease where pancreatic islet cells cease to produce insulin, which is necessary to maintain proper blood glucose levels in the body. Also, scientists have shown the ability to differentiate iPS cells into all three germ layers, which includes peripheral neurons (Lee et al., 2009). This discovery could be used towards future potential cures for brain disease.

These findings suggest that understanding the defining transcription factors for any given cell identity is critical for cell-based therapies. One avenue of research, which remains unexplored, is the possibility of reprogramming disease states. For example, by understanding the disease and normal transcriptional networks, applying the correct combination of transcription factors could in theory restore the normal cellular state.

Scleroderma is a systemic autoimmune and fibrotic disorder with no known cure. Scleroderma affects the skin, lung, kidney, and gastrointestinal tract. Women are four times more likely to contract scleroderma than men and one in every 1,000 people will be

diagnosed, typically between the age of 25 and 55 years of age. Defined in Greek as "hard skin," scleroderma has the appearance of hard, smooth, and ivory-colored skin, which appears to be immobile and may occur in both localized and systemic forms. Scleroderma is known to result in the overproduction of collagen in dermal cells (Scleroderma Foundation, 2009). Previous work has identified increase biosynthesis of collagen and alpha-actin in scleroderma fibroblasts (Uitto, 1979; Kirk, 1995). This is important as the scleroderma phenotype is characterized by apoptosis of the endothelial cells of the arterioles and smooth muscle cells. Once this occurs, collagen, fibrous material, and inflammatory cells infiltrate the space causing further damage (Gabrielli et al., 2009).

Microarray studies have also been performed between scleroderma and normal fibroblasts (Whitfield, 2003). Importantly, these microarray studies provide datasets to determine global patterns of gene expression in scleroderma versus normal fibroblasts and to identify transcription factors expressed specifically in scleroderma. In addition, altered gene expression in scleroderma reflects the activity of a network of transcription factors in scleroderma. These approaches should lead to the identification of disease-determining transcription

factors.

Our hypothesis is that by identifying unique transcription factors in disease, the disease-specific gene expression patterns can be induced in the lab. Depending on one's ability to induce scleroderma-specific gene expression; the diseased phenotype can then be reproduced in the lab and subsequently reverted to its normal state. This has tremendous therapeutic implications as current treatments for scarring only ameliorate the condition, not cure it.

SPECIFIC AIMS

The specific aims of this research project are: to identify the major transcription factors that drive scleroderma and specific gene expression, to discover the downstream targets for these transcription factors, and to identify the downstream targets of transcription factors that are highly expressed in scleroderma and fibrotic cells.

CHAPTER 2

Predicting the Major Transcription Factors that Drive Scleroderma-Specific Gene Expression

Transcription Factors Regulate Global Gene Expression

Gene transcription is responsible for tissue-specific gene expression and the regulation of gene activity in response to certain stimuli (Latchman, David S., 1997). Regulatory regions often have short recurring patterns of DNA sequence that act as binding sites for activating and inhibiting proteins called transcription factors. While genes can also be regulated post-translationally, the majority of gene regulation occurs at the level of transcription initiation, a process that is primarily determined by transcription factors (TFs). In addition, scientists' initial inspections of regulatory regions with the same pattern of gene expression revealed that TFs have been found to regulate the expression of a particular gene. TFs have the ability to either upregulate or downregulate the expression of a particular gene to affect transcription and ultimately the cellular phenotype.

TFs are generally classified on the basis of their DNA binding domains. These domains are often used to bind to specific DNA sequences. As TFs bind to a particular region of DNA, they recruit several other factors, including RNA Polymerase to begin the process of messenger RNA (mRNA) transcription, which is later translated into proteins. In addition to their initiating role in 1 transcript, TFs can also act

to inhibit transcription. Complex networks of transcription factors work in combination to define the overall gene expression of an organism and ultimately the organism's phenotype. The activity of a TF is regulated in two ways: 1) by controlling the synthesis or expression of a particular TF, and 2) by modulating their activity through post-translational modifications or expression. An example of the first method is the MyoD TF, which is specifically expressed in skeletal muscle cells. In addition, ectopic expression of MyoD induces muscle-like features in some non-muscle cell types. In terms of regulating the activity of TFs, IL6 serves as an example where it induces/enhances the synthesis of NF-IL6β TF, which in turn is complemented by the transcription of other factors, NF IL-6 and STAT-3 (Latchman, David S., 1997). In this chapter, we aim to explore the method of controlling the synthesis of TFs via overexpression, resulting in the alteration of global gene expression patterns.

Defining Promoters and Co-regulated Regions by Evolutionary Conservation

A major problem in defining the regulatory region of a gene is determining the sequences necessary to drive expression. Usually this is done by reporter assays or experiments involving transgenic mouse lines. Even with these approaches, important regulatory regions may go

undiscovered. Another approach to discover regulatory regions responsible for gene expression is to predict those regions using the concept of evolutionary conservation. Once those regions are predicted, scientists can use sequence information from known transcription factor binding sites to determine what TFs may regulate that gene.

The recent sequencing of the human and eukaryotic genomes provides a scaffold for understanding the genetic mechanisms that regulate biological function. As a result of sequencing the human genome, fewer genes than expected were found and the role of regulatory elements other than promoters are poorly understood. However, the recent advancements in understanding the human genome provides hope for our ability to decipher regulatory patterns and gene regulation. Early experiments that aimed to understand regulatory mechanisms focueds on simple organisms like yeast and drosophila, where *in silico* mechanisms are easier to validate experimentally (Gotea et al., 2008). While most current prediction methods look at local enrichment of transcription factor binding sites (TFBSs), DiRE aims to improve upon current prediction methods by looking at additional information, which includes sequence

conservation across taxa, nucleosome occupancy, and binding competition between factors.

It is known that the regulation of gene expression in the eukaryotic genome is achieved through a complex set of networks of regulatory elements. The creators of DiRE have created the Enhancer Identification (EI) method, which infers position and functional information on distant regulatory elements (RE) from the analysis of either microarray gene expression, or co-regulation data. DiRE is uniquely integrated with the Array2BIO server, which gives the user access to raw microarray expression data to aid in their analysis. In a study of 79 groups of tissue-specific genes, 23% of candidate regulatory elements were found in the promoter region, while over half of the remaining elements resided in intronic or intergenic regions (Gotea et al., 2008). This EI method combines gene co-expression data and gene microarray data with evolutionary conservation across genomes to accurately predict upstream regulators, or TFs of target genes. By an *in vivo* validation of transgenic mice, DiRE was shown to have 28% sensitivity and 50% precision. Our goal was to use this web-based server to predict upstream regulators, otherwise known as TFs, to regulate our

set of 165 differentially expressed genes thought to be responsible for the scleroderma gene expression pattern, and phenotype.

MATERIALS AND METHODS:

Microarray Data Analysis

Raw E-MEXP-32 cel files (Whitfield et al., 2003) containing skin gene expression data of normal and scleroderma patients were loaded into Bioconductor, a tool in the statistical program R (Bioconductor, 2004). A quantile normalization was then applied. Next, a probe match only summarization using the median polish algorithm was done, generating Log_2 signal intensities (Knowledge and Information Systems, 2003). Probe sets that had a change in expression of 25% from the signal intensities of normal were analyzed. Those probe sets were used to conduct multiple hypothesis testing using a Benjamini-Hochberg correction and genes were selected with differential expression of $p<0.05$. The final output of this analysis was 165 differentially expressed genes between Scleroderma and normal patients skin samples.

Prediction of Binding Sites in Scleroderma/Normal Skin Genes (http://dire.dcode.org/)

The 165 genes determined to be differentially expressed from our microarray data analysis were input as a set of co-regulated genes. A random set of 5000 background genes were used as control, and the target elements selected were the top 3 evolutionary conserved

regions (ECRs) and the top 3 promoter ECRs. These target elements of our co-regulated gene sets were tested. Analysis was done with the human genome build 18 (hg18). From the output, TFs and downstream genes were selected based upon the criteria mentioned in the upcoming results section.

Cloning Transcription Factors and Vectors Used

Cloning primers were designed and amplified from complimentary DNA (cDNA). Linkers specific to the overexpression vector multiple cloning site (MCS) were added along with the Kozak Consensus Sequence. The amplified gene of interest was extracted from 1% agarose gel using standard protocol for the Zymoclean Gel DNA Recovery and Cleanup Kits. The overexpression vector of interest, (**Figure 1**) pcDNA 3.0, was digested at its MCS and TF genes were ligated into it. The resulting overexpression vectors were grown up and checked with sequencing. ELF1, ETF1, YY1 and ZF5 were cloned using this above method. An additional 6 TFs, SP1, SMAD3, EGR2, POU3F2, ELK1, and NF-KB1, were obtained from an Open Biosystems Human TF Library. These 6 TFs were previously cloned into the mammalian overexpression vector, pCMV-Sport6 (**Figure 1**).

Cell Culture and Transfections

Dermal fibroblasts (BJ cell line) were thawed from cell stock at passaged #8 and cultured in D10 media (DMEM plus 10% Fetal Calf Serum (FCS)) with penicillin-streptomycin bacterial antibiotic. Cells were passaged using Trypsin and stored in a 5% CO_2, 37°C incubator. Dermal Fibroblasts were transfected using the standard electroporation protocol provided by Invitrogen for the Neon electroporation unit. A 24-well cell-line specific optimization was completed to determine the optimal parameters for the Neon, which was 1 pulse with a pulse width of 20ms, and a pulse voltage of 1650 volts (also optimized by Invitrogen).

Quantitative Polymerase Chain Reaction (qPCR)

Total RNA was extracted from dermal fibroblasts using the Zymo RNA Purification Kit. cDNA was synthesized using the Fermentas Maxima First Strand cDNA Synthesis Kit. Primers were designed using sequence information from Ensembl and Primer3 to optimize conditions. The qPCR primers were designed to span exon-exon gap junctions to eliminate non-specific binding to possible genomic DNA contamination. The 10 qPCR TF primer sets used in this chapter are:

"Species-Gene-Exon-Direction;Sequence; Melting Temp Product Size"

Hu-ZF5-Exon2-F;GATATGGGTCTGCAGGATGG;60C 208bp
Hu-ZF5-Exon4-R;CTTCCAGGCGTTGTTCATTT;60C 208bp

Hu-YY1-Exon2-F;ACCTGGCATTGACCTCTCAG;60C 193bp
Hu-YY1-Exon4-R;TTCTGCACAGACGTGGACTC;60C 193bp

Hu-ETF1-Exon4-F;CACTTTTTGGCACACTCCAA;60C 120bp
Hu-ETF1-Exon5-R;CCATTCTTAAACGGGCAAAA;60C 120bp

Hu-ELK1-Exon2-F;GGCTACGCAAGAACAAGACC;60C 200bp
HU-ELK1-Exon3-R;ATTTGGCATGGTGGAGGTAA;60C 200bp

Hu-SMAD3-Exon1-F;GAGGAGAAATGGTGCGAGAA;60C 192bp
Hu-SMAD3-Exon2-R;GCGGCAGTAGATGACATGAG;60C 192bp

Hu-EGR2-Exon1-F;GGTGACCATCTTTCCCAATG;60C 123bp
Hu-EGR2-Exon2-R;GGATATGGGAGATCCAACGA;60C 123bp

Hu-POU3F2-Exon2-F;ACGGCGGCTTGCTCTACT;60C 137bp
Hu-POU3F2-Exon3-R;CTTGAACTGCTTGGCGAACT;60C 137bp

Hu-SP1-Exon3-F;GGCCTCCAGACCATTAACCT;60C 165bp
Hu-SP1-Exon4-R;TCCACCTGCTGTGTCATCAT;60C 165bp

Hu-ELF1-Exon4-F;GCCCTATGCTGGATGAAAAA;60C 160bp
Hu-ELF1-Exon5-R;CCCGGTGAGTCTGCATATTT;60C 160bp

Hu-NFKB1-Exon5-F; ACTGTGAGGATGGGATCTGC;60C 128bp
Hu-NFKB1-Exon6-R; CTCTGTCATTCGTGCTTCCA;60C 128bp

The 18 downstream qPCR target primer sets used in this chapter are:

Hu-YWHAE-Exon4-F;CTTCCACCAACGCATCCTAT; 60C 141bp
Hu-YWHAE-Exon5-R; CAGCGTATCCAGTTCTGCAA; 60C 141bp

Hu-NFKB1-Exon5-F; ACTGTGAGGATGGGATCTGC;60C 128bp
Hu-NFKB1-Exon6-R; CTCTGTCATTCGTGCTTCCA;60C 128bp

Hu-TGFB1-Exon1-F;GAGCCTGAGGCCGACTACTA;60C 131bp
Hu-TGFB1-Exon2-R;CGGAGCTCTGATGTGTTGAA;60C 131bp

Hu-CYC1-Exon2-F;TCTCTTCCTTGGACCACACC; 60C 195bp
Hu-CYC1-Exon4-R;GCATGAACATCTCCCCATCT; 60C 195bp

Hu-PPP2R2A-Exon6-F; TGCAGATGATTTGCGGATTA;60C 127bp
Hu-PPP2R2A-Exon7-R; TGGATGAAATTCTGCTGCTG;60C 127bp

Hu-OAZ1-Exon5-F; GAGCCGACCATGTCTTCATT;60C 179bp
Hu-OAZ1-Exon6-R; CTCCTCCTCTCCCGAAGACT;60C 179bp

Hu-PSME2-Exon6-F;GTGGATTTCTCCCTGGGAAT;60C 124bp
Hu-PSME2-Exon7-R;ATCTTGGGGATCAGGTGTTG;60C 124bp

Hu-CCL5-Exon2-F;TACACCAGTGGCAAGTGCTC;60C 100bp
Hu-CCL5-Exon3-R;TGTACTCCCGAACCCATTTC;60C 100bp

Hu-RBBP4-Exon3-F;TGATGCGTCACACTACGACA;60C 116bp
Hu-RBBP4-Exon4-R;AACGGGCCCTGTTTACTTCT;60C 116bp

Hu-ODC1-Exon8-F;GTGGCTTTCCTGGATCTGAG;60C 120bp
Hu-ODC1-Exon9-R;CGGGCTCAGCTATGATTCTC;60C 120bp

Hu-IFI6-Exon1-F;CTGTGCCCATCTATCAGCAG;60C 146bp
Hu-IFI6-Exon2-R;CCACTGCAAGTGAAGAGCAG;60C 146bp

Hu-PARP1-Exon1-F;AAGAAATGCAGCGAGAGCAT;60C 164bp
Hu-PARP1-Exon2-R;TCAGAGAACCCATCCACCTC;60C 164bp

Hu-THBS1-Exon7-F;AGAATGCTGTCCTCGCTGTT;60C 140bp
Hu-THBS1-Exon8-R;ATCGGTTGTTGAGGCTATCG;60C 140bp

Hu-TOP1-Exon3-F;GAGAAGGACCGGGAAAAGTC;60C 100bp
Hu-TOP1-Exon4-R;AGCTTCCATCTTTGTGTTTGG;60C 100bp

Hu-ACTG2-Exon3-F;AGACAGCTATGTGGGGGATG;60C 156bp
Hu-ACTG2-Exon4-R;GGGTGCTCTTCAGGTGCTAC;60C 156bp

Hu-Crabp2-Exon2-F;AGACAGTGTCCAGTGCTCCA;60C 161bp
Hu-Crabp2-Exon3-R;CACAGCAATCTTCCTCAGCA;60C 161bp

Hu-RELA-Exon4-F;CCACGAGCTTGTAGGAAAGG;60C 162bp
Hu-RELA-Exon5-R;AAGGGGTTGTTGTTGGTCTG;60C 162bp

Hu-IL18-Exon1-F;TGCATCAACTTTGTGGCAAT;60C 169bp
Hu-IL18-Exon3-R;ATAGAGGCCGATTTCCTTGG;60C 169bp

qPCR protocol was conducted using a Roche480 Light Cycler and SYBR Green reagent with the following per sample mix:

Reagent:	**Volume (µL):**
SYBR Green	5
Water	2
Primer (10 pM)	.5
cDNA	2.5
Total	**10**

Recorded cycle numbers were entered in Microsoft Excel in order to analyze relative gene expression, which was normalized to Human Gapdh. The Δ ΔCt method or the comparative method for quantifying relative gene expression data was used for analysis.

Imaging : Microscopy

Light and fluorescent images were taken of transfected cells in culture using Olympus MVX10 and Olympus BX51 microscopes. Software programs were used to adjust image settings, exposure time, etc.

Flow Cytometry

Transfected BJ cells were trypsinized and spun down from culture. Pelleted cells were then washed and suspended in PBS. Standard

protocol per the manufacturer's recommendations were used to conduct flow cytometry with the Millipore Guava easyCyte 8HT machine. Green fluorescence was measured for transfected cells and samples were gated relative to a non-transfected negative control sample.

RESULTS:

Microarray Data Acquisition

To identify transcription factors associated with the potential to activate genes of the Scleroderma disease phenotype, a set of activated genes unique to Scleroderma was determined. Previously, Whitefield et al. obtained skin biopsies from individuals diagnosed with systemic sclerosis, a form of diffuse scleroderma, and used microarray analysis to study the expression of ~12,000 genes. Four patients (two men and two women) underwent two sets of biopsies. These biopsies included 5 mm punch biopsies from the lateral forearm, 8cm proximal to the ulnar styoid. Three biopsies total for each patient were taken from clinically involved skin and another set of three biopsies were taken from the buttock or back for clinically uninvolved skin. In addition, four normal control samples were taken from one man and three women. They underwent the exact same biopsies, with the exception of control individual number 4 that only had biopsies of the lateral forearm. Of the set of three biopsies, 2 were frozen, half of the third biopsy went into 10% formalin for routine histology, and the other half was for fibroblast cell culture.

Preparation of total RNA was done from the two frozen punch biopsies and cRNA synthesis, and hybridization to an Affymetrix Hu95A microarray followed. The raw data produced from this microarray had multiple sets, but two of the sets of interest were the overall biopsy microarray data and the culture fibroblast microarray data. Biopsy microarray data from this experiment was used for analysis. The end result the analysis using the raw data from the Whitfield publication resulted in 165 (163 unique genes) differentially expressed genes between the scleroderma and normal skin samples (**Tables 1 and 2**). Of these genes, 92 (56%) were higher expressed in the normal samples with the remaining 73 (44%) in the scleroderma diseased samples (both numbers include NF-KB1 and RELA, which was highly expressed in both due to probe variation). In addition, this set of differentially expressed genes contains a total of 15 endogenously expressed transcription factors. 12 (71%) of these transcription factors were expressed higher in the normal skin samples, versus 5 (29%) transcription factors in the scleroderma samples. It is important to note that NF-KB and RELA were highly expressed in both scleroderma and normal samples due to probe sets interrogating different exons within the gene, which possibly corresponds to different isoforms of both genes.

Predicting Transcription Factors that Regulate Differentially Expressed Genes

The next step in the process was to begin to predicting transcription factors that regulate these differentially expressed human genes. In order to do this, web server named DiRE was used, which was developed by the National Center for Biotechnology Information (NCBI). Using the DiRE tool, the previously discussed 165 differentially expressed annotated gene names were entered based on the human genome (hg18) and a random set of 5000 genes for the background/control genes was used. The untranslated region (UTR) evolutionary conserved regions (ECRs) and promoter ECRs were selected for target elements. We decided to focus mainly on ECRs within the UTR and proximal promoter region due to the UTR and promoter's proximity to the transcription start site and their known role in transcription.

The search resulted in the finding of 65 regulatory elements from the input genes. 42 (65%) of these elements were found within the promoter region and the other 23 (35%) were located within the UTR. Based upon these 65 regulatory elements a potential 112 candidate transcription factors were returned. The 112 transcription factors

returned were displayed along with two scores, occurrence and importance. Occurrence indicates the fraction of regulatory elements containing a particular TF, while importance is the product of occurrence and the weight assigned to each TF after optimization. Optimization is a process that increases the number of candidate regulatory elements recognized by a profile of transcription factor binding sites (TFBS) in the loci of input genes, while decreasing the number of such predictions in the loci of background genes. During this optimization, TFBSs are assigned weights based upon the TFBS content of the candidate regulatory elements (Gotea et al.). Essentially, importance is occurrence with taking into account the overall promiscuity of the TF in all of the background genes and the higher the importance value, the greater the probability that the TF binds and regulates the gene target.

Of these 112 candidate TFs, 10 TFs were selected with varying occurrence and importance values to ensure an accurate and unbiased distribution of predicted candidate TFs (**Table 3**). This should help ensure that our validation of DiRE is most accurate. Occurrence values ranged from 34% to 3% and importance values ranged from 0.002 to 0.26. These values exemplify an even distribution of the

occurrence and importance values for the 112 candidate TFs returned and the 10 TFs chosen are an accurate representation of the overall profile and characteristics of the original 112 TFs. Tertiary characteristics taken into account for choosing TFs were whether or not these were associated with published data and were known activators/repressors.

Selection of Downstream Gene Targets for 10 DiRE-predicted TFs

With 10 candidate transcription factors already selected from the DiRE results, downstream genes needed to be selected to measure the predicted TFs activity or ability to regulate these targets. **Table 4** details the original 18 downstream genes selected. These targets were selected with the following factors in mind: number (both high and low) of transcription factors binding sites (TFBS) of the selected TF in the gene and the number (both high and low) of total TFBSs for the specified gene. High and low numbers of TFBSs in each situation were chosen to ensure that there was a range of differentiated targets. The variation in target characteristics is best exemplified by the range (0.001-2.780) of the selected genes' regulatory element's associated scores. This score takes into account the impact of multiple TFBS motifs in order to better classify candidate enhancers based on sequence signatures that define gene expression in a particular tissue. This tissue-specific scoring

scheme allows one to enrich for positively scoring candidate enhancers in tissue-specific loci (Pennacchio et al.). In addition, the location of the regulatory element (RE), the type of RE (UTR5, Promoter), and the locus in which it the RE exists are displayed in **Table 4**. Throughout the initial prediction process, it was our goal to create an unbiased profile of selected candidate TFs and downstream gene targets.

Transfecting 10 DiRE-predicted Factors into Dermal Fibroblasts

With 10 TFs predicted and 18 downstream genes selected, our goal was to overexpress these TFs in human cell lines and measure the gene expression of corresponding targets to see if we can alter a subset of the 165 differentially expressed genes. All 10 DiRE-predicted factors were cloned and isolated into one of two possible mammalian overexpression vectors, pCMV-SPORT6 and pcDNA3 (**Figure 1**). Both of these expression vectors are driven by the cytomegalovirus (CMV) promoter. Each of the individual 10 TFs were transiently transfected into cultured dermal fibroblasts along with a GFP reporter plasmid using electroporation. Three control samples were transfected with vector alone and with the GFP reporter plasmid, while another three samples were transfected with the overexpression vector containing the DiRE-predicted TF and a GFP reporter plasmid. After 48 hours from the point

of transfection, total RNA was extracted from the treated cells. 48 hours was chosen as the optimal time point because fibroblasts divide relatively slowly (about once a day) and changes in gene expression could be observed then. First-strand complimentary DNA (cDNA) was synthesized from the each sample's RNA extract and quantitative polymerase chain reaction (qPCR) was conducted to measure the expression of candidate TFs and downstream gene targets.

Initially, a lipofection reagent was used to transfect each TF into dermal fibroblast cells. However, examination by fluorescence microscopy revealed transfection efficiency of less than 5% (**Figure 2A**). In comparison, electroporation resulted in a higher transfection efficiency as seen in **Figure 2B**. Flow cytometry determined that our transfection using electroporation yielded a 27.19% transfection efficiency.

Measuring Expression of DiRE-predicted TFs and Downstream Gene Targets

Once gene expression data was collected from qPCR, sample threshold cycle values (C_T) were used to calculate relative gene expression, which was normalized to Human Gapdh expression in the

same samples. **Figure 3** details the expression data from the 10 Dire-predicted TFs that were transfected into a total of six samples containing cultured dermal fibroblasts. The mock-transfected samples were used as baseline TF expression levels. Six out of ten of the TFs (ETF1, POU3F2, NF-KB1, ELF1, YY1, and EGR2) were overexpressed by greater than 10 times relative to control samples. Four of the remaining TFs, SMAD3, SP1, ELK1, and ZF5, were not significantly overexpressed. Four of the six TFs were overexpressed by greater than 100 times relative to control. EGR2 was the most overexpressed TF with 3000 fold relative expression.

With the successful overexpression of the majority of our DiRE-predicted TFs, we sought to determine whether the TF overexpression was sufficient to activate scleroderma target genes. Three downstream targets that are predicted to be regulated by NF-KB1 and were a part of the 165 differentially expressed genes were analyzed (**Figure 4**). The first gene PSME2, was not differentially expressed between the control and treated samples. The second gene analyzed, CCL5, exhibited a 5-6 fold increase in gene expression in the presence of NF-KB. The increase in expression was determined to be statistically significant with a p-Value of 0.007. This means that there is a 0.7% chance that the change

in gene expression is not due to the treatment, or in this case, the overexpression of NF-KB TF. Two remaining predicted targets PSME2 and YWHAE, were not altered in expression levels in treated and control samples. Thus, for NF-κβ overexpression only 1 out of 3 of the targets were activated.

To further our validation of DiRE and determine whether we would be able to activate additional scleroderma gene targets with DiRE-predicted TFs, we tested and validated the response to the 9 additional predicted TFs. A summary of the transfections of each TF and the associated gene target expression is provided in **Table 5**. The overexpression levels of the TFs that was previously discussed is displayed along with their associated downstream gene targets. A total of 36 targets that were predicted to be regulated by the 10 TFs were examined. From our analysis of the treated samples, a total of 8 out of the 36 total targets were statistically altered from control ($p<0.05$). Unexpectedly, altered expression of four downstream targets was observed even in the absence of clear overexpression of predicted TFs. 4 of the 8 differentially expressed targets were associated with the 4 TFs that were unable to be overexpressed.

In addition to looking at statistical significance, we evaluated all target gene expression to see which targets had a change in expression greater than 50% from baseline. With this cutoff, we wanted to choose a more stringent selection method to see which target's gene expression that we were able to radically change. A total of 3 genes had a change in expression greater than 50% from baseline. These three targets were CCL5, YWHAE, and PPP2R2A, but only CCL5 was also statistically significant. In addition, the 50% change in expression from baseline of YWHAE and PPP2R2A occurred in the ZF5 transfection sample where we did not experience overexpression of the TF. From this data, we can say that most (7/8, or 87.5%) of the statistically significant targets had a change of expression less than 50%. While one target, CCL5, was both statistically significant and had an average change in expression that was 411% increased in the treated sample compared to control.

DISCUSSION:

Our goal with the experiments contained in this chapter was to indentify the major TFs that drive scleroderma-specific gene expression. In order to do this, we used raw microarray data to find differentially expressed genes between scleroderma and normal skin samples. Our analysis resulted in 165 differentially expressed gene targets between both the scleroderma and normal phenotypes. We were able to use the web-based bioinformatic program DiRE to predict 10 TFs to regulate up to 18 of our 165 differentially expressed genes. Shared TFBSs among the 165 target genes were used to predict TFs. Upon switching our transfection method, we were able to successfully transfect dermal fibroblasts and overexpress the majority of predicted TFs.

In our analysis of a total of 36 downstream targets, 10 or 27.7% of those targets were differentially expressed on either a statistically significant basis or with an average chance in expression greater than 50%. The only target to fit both categories was CCL5, which has a 411% increase in expression and was regulated by the NF-KB TF. Of those 10 differentially expressed targets, 6 (YWHAE (2), CCL5, NF-KB, PPP2R2A (2)) of those were upregulated of activated, while 4 (IL18, PSME2, RBBP4, ODC1) were repressed. All of the 6 activated targets were consistent

with scleroderma gene expression, meaning that those 6 targets had increased expression in the scleroderma diseased patient samples. The 4 repressed targets were not consistent with scleroderma gene expression. Those 4 targets were actually increased in scleroderma patient samples and relatively decreased in normal patients samples. The results have important implications for potential future reprogramming of the scleroderma phenotype. Thus, we know that YY1, NF-KB1, ETF1, ZF5, and SP1 were sufficient to activate these 6 scleroderma-specific genes potentially leading to partial activation of the scleroderma phenotype. Conversely, YY1 and ELK1 TFs were found to repress the expression of four scleroderma-specific targets and those TFs could be used to reprogram the normal skin phenotype from scleroderma disease-effected cells.

While 27.7% of analyzed gene targets were differentially expressed, all except CCL5, YWHAE (ZF5 TF), and PPP2R2A (ZF5 TF) were differentially expressed at low levels (below 50% avg. Δ) and only one target was both differentially expressed and statistically significant (CCL5). There are several factors that could have increased our ability to differentially express a larger amount of genes. First, the microarray data analyzed was from punch biopsies, which contains a mixed

population of cells. The positive aspect of analyzing gene expression of a biopsy is that it is a true physiologic snapshot of the disease. However, our experiments were conducted in fibroblasts. It is possible that a subset of the differentially expressed genes are differentially expressed in cells other than fibroblasts, therefore it would be much more difficult to mimic that same gene expression in a different cell type (Machesney et al., 1998). It is also possible that the four TFs we were unable to overexpress were underexpressed due to cellular mechanisms preventing its overexpression, or the cell already had high level of TF expression making it more difficult to overexpress. Nonetheless, the majority of our TFs were overexpressed allowing us to continue to look at their predicted downstream targets. Also, TFs are known to act together, or in networks of gene expression, thus transfecting a single TF may not be substantial enough to evoke larger changes in gene expression. While improvements can be made, this initial DiRE analysis proved itself to be predictive and it deserves further attention.

Table 1: Differentially Expressed Genes Increased In Normal Skin Phenotype (Input Set)

#	Gene	#	Gene	#	Gene	#	Gene
1	PF4	**26**	GSK3B	**51**	CIB1	**76**	RPS6KA1
2	CYP4F2	**27**	EMP1	**52**	UBA7	**77**	AXL
3	GSTM3	**28**	THRA	**53**	NF-KB*	**78**	CGB7
4	KLF1	**29**	CDA	**54**	FGFR4	**79**	EFS
5	NCAM1	**30**	MST1	**55**	MYCL2	**80**	RELA*
6	CYP11B1	**31**	GUCA1A	**56**	FLT3LG	**81**	TNK2
7	PCDHGC3	**32**	MAP4K1	**57**	MME	**82**	MAPK3
8	MAPK11	**33**	RABGGTA	**58**	RAB40B	**83**	EDN2
9	IL3	**34**	GSTM5	**59**	GATA2	**84**	MAP3K11
10	CXCR5	**35**	PTPN9	**60**	MLANA	**85**	PAX8
11	GSTA2	**36**	ANG	**61**	MT3	**86**	EIF4G1
12	WNT10B	**37**	CDC34	**62**	CSPG4	**87**	GRK6
13	NTRK3	**38**	NOS2A	**63**	GPR31	**88**	DUSP1
14	RXRG	**39**	CYP4F3	**64**	PLCG2	**89**	HSPA1A
15	XPA	**40**	LIG3	**65**	RBPMS	**90**	IGFBP5
16	CCR1	**41**	RARG	**66**	SRF	**91**	DDR1
17	CD19	**42**	GSTZ1	**67**	TRA@	**92**	EEF1A1
18	RAP1GAP	**43**	MLL	**68**	CGB		
19	DUSP2	**44**	IL8RB	**69**	POLR2H		
20	GH1	**45**	CDH15	**70**	TCF3		
21	MST1R	**46**	TIMP2	**71**	LTK		
22	ARHGEF16	**47**	PTPRS	**72**	TYRO3		
23	CASP2	**48**	MUC1	**73**	PRKAR1B		
24	E2F1	**49**	QSOX1	**74**	MMP15		
25	TNFRSF25	**50**	COL11A2	**75**	FGFR2		

* Differentially Expressed in Both Scleroderma/Normal Samples (suspected different isoforms)
Yellow Highlight = Transcription Factor

Table 2: Differentially Expressed Genes Increased In Scleroderma Skin Phenotype (Input Set)

#	Gene	#	Gene	#	Gene
93	CCL5	**118**	ISG15	**143**	PSMB4
94	TNFAIP6	**119**	YWHAE	**144**	TCEB1
95	IL10RA	**120**	MAP2K1	**145**	PSMB3
96	LYN	**121**	CDC25B	**146**	CYC1
97	PIK3R1	**122**	POLG	**147**	CRABP2
98	RAB1A	**123**	PSMD11	**148**	ACTG2
99	THBS1	**124**	ERF	**149**	TGFBI
100	MAPK9	**125**	DDB2	**150**	PSMB2
101	TWF1	**126**	PSMD1	**151**	EIF2AK2
102	PTPRZ1	**127**	ODC1	**152**	PSMD2
103	SLA	**128**	PSMB7	**153**	RHOC
104	IL18	**129**	SLC38A2	**154**	UBA1
105	RBBP4	**130**	TIMP3	**155**	PSMD8
106	VDR	**131**	TOP1	**156**	EIF4A1
107	CD44	**132**	PPP2R2A	**157**	HINT1
108	THBS4	**133**	PSME2	**158**	RHOA
109	SLC20A1	**134**	IL23A	**159**	EIF4A2
110	NF-KB*	**135**	IL1R1	**160**	YWHAH
111	PARP1	**136**	CTSC	**161**	HSP90AB1
112	SFRS10	**137**	PTP4A2	**162**	YWHAZ
113	IFI6	**138**	REEP5	**163**	UBB
114	EPHB4	**139**	CALM3	**164**	UBC
115	CTSK	**140**	RELA*	**165**	OAZ1
116	RPA2	**141**	CEBPD		
117	IER3	**142**	TCEA1		

* Differentially Expressed in Both Scleroderma/Normal Samples (suspected different isoforms)
Yellow Highlight = Transcription Factor

Table 3: Characteristics of 10 DiRE-predicted Transcription Factors

#	TF	Occurrence	Importance	Associated Targets
1	SP1	33.85%	0.06873	CRABP2, TGFB1, RELA, NF-KB1, CYC1, PPP2R2A, OAZ1
2	ZF5	27.69%	0.00428	YWHAE, OAZ1, NF-KB1, PPP2R2A
3	ELK1	18.46%	0.26077	PSME2, RBBP4, ODC1, YWHAE
4	ETF1	16.92%	0.09909	YWHAE, OAZ1, NF-KB1, PPP2R2A
5	NF-KB1	13.85%	0.13846	PSME2, CCL5, YWHAE
6	EGR2	12.31%	0.11132	NF-KB1, PARP1, THBS1
7	YY1	10.77%	0.04156	IL18, CYC1, PPP2R2A, OAZ1
8	ELF1	3.08%	0.00275	RELA, NF-KB1
9	SMAD3	3.08%	0.03308	IFI 6, NF-KB1
10	POU3F2	3.08%	0.01138	TOP1, ACTG2

Table 4: 18 Differentially Expressed Gene Targets Used for qPCR Analysis

Gene	Regulatory Element	Type	Score	Locus	Gene	Regulatory Element	Type	Score	Locus
ACTG2	chr2:73973182-73973890	UTR5	0.259	chr2:73943558-74007447	PARP1	chr1:224662407-224662507	UTR5	0.758	chr1:224563828-224803083
CCL5	chr17:31220031-31269179	UTR5	2.780	chr17:31220031-31269179	PPP2R2A	chr8:26204897-26205252	UTR5	0.006	chr8:25958522-26296426
CRABP2	chr1:154942048-154942209	Promoter	1.058	chr1:154915774-154959018	PSME2	chr14:23686135-23686253	Promoter	0.192	chr14:23680640-23687177
						chr14:23686279-23686386	Promoter	0.365	chr14:23680640-23687177
CYC1	chr8:145221875-145221990	UTR5	0.995	chr8:145213120-145225448	RBBP4	chr1:32888754-32888880	Promoter	0.006	chr1:32887984-32918839
IFI 6	chr1:27871158-27871383	UTR5	0.010	chr1:27834354-27925075	RELA	chr11:65187961-65188194	Promoter	0.001	chr11:65175083-65236054
IL18	chr11:111540443-111540613	Promoter	0.055	chr11:111471857-111543248	TGFB1	chr5:135391775-135391964	Promoter	0.319	chr5:135318644-135496433
NF-KB1	chr4:103641372-103641644	UTR5	0.601	chr4:103485410-103771527	THBS1	chr15:37660269-37660505	Promoter	0.012	chr15:37335029-37679369
OAZ1	chr19:2220395-2220597	UTR5	0.010	chr19:2206345-2224569	TOP1	chr20:39090242-39090412	Promoter	0.007	chr20:38751306-39199562
ODC1	chr2:10505707-10506001	UTR5	1.533	chr2:10485358-10627882	YWHAE	chr17:1250155-1250503	UTR5	0.889	chr17:1151418-1272173

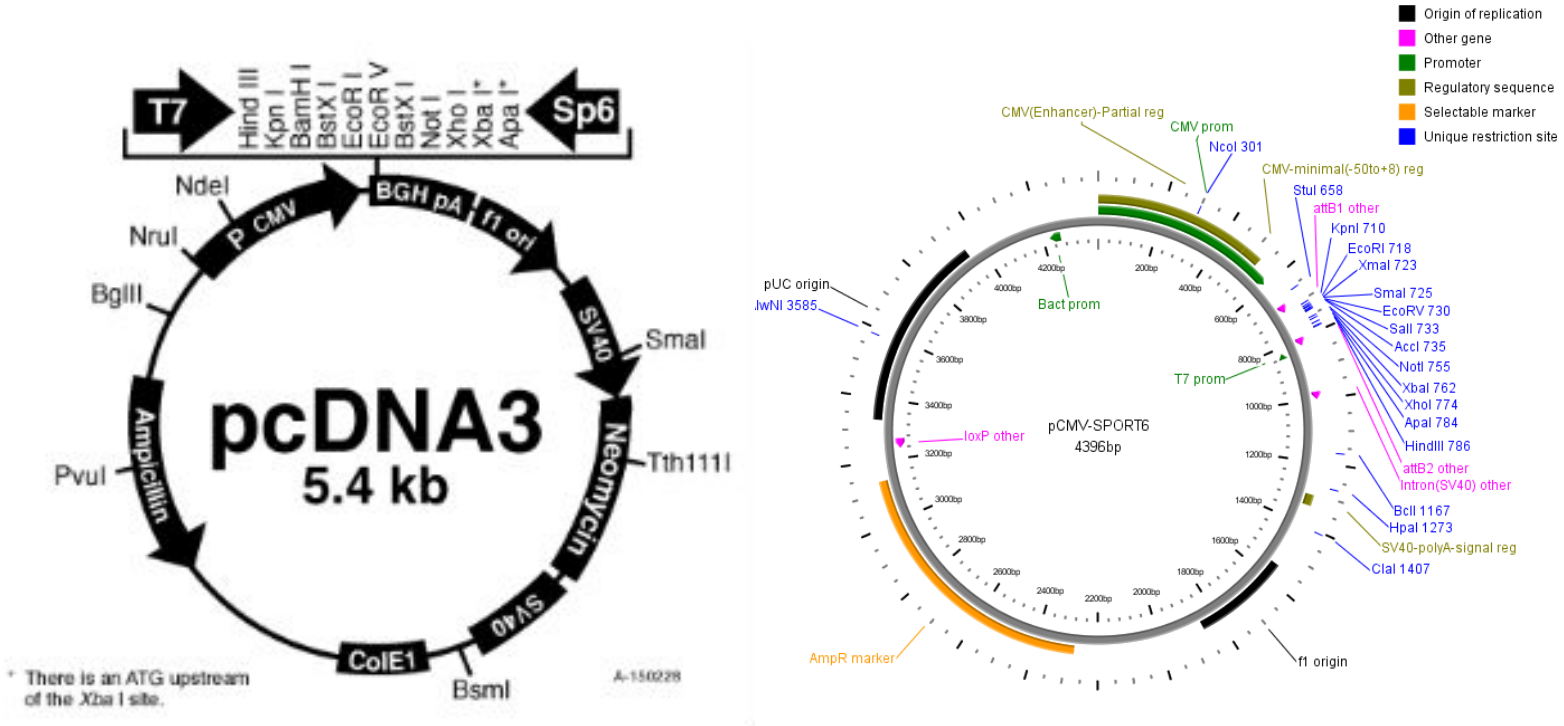

Figure 1: pcDNA3 and pCMV-Sport6 Overexpression Vectors Used to Transfect TFs

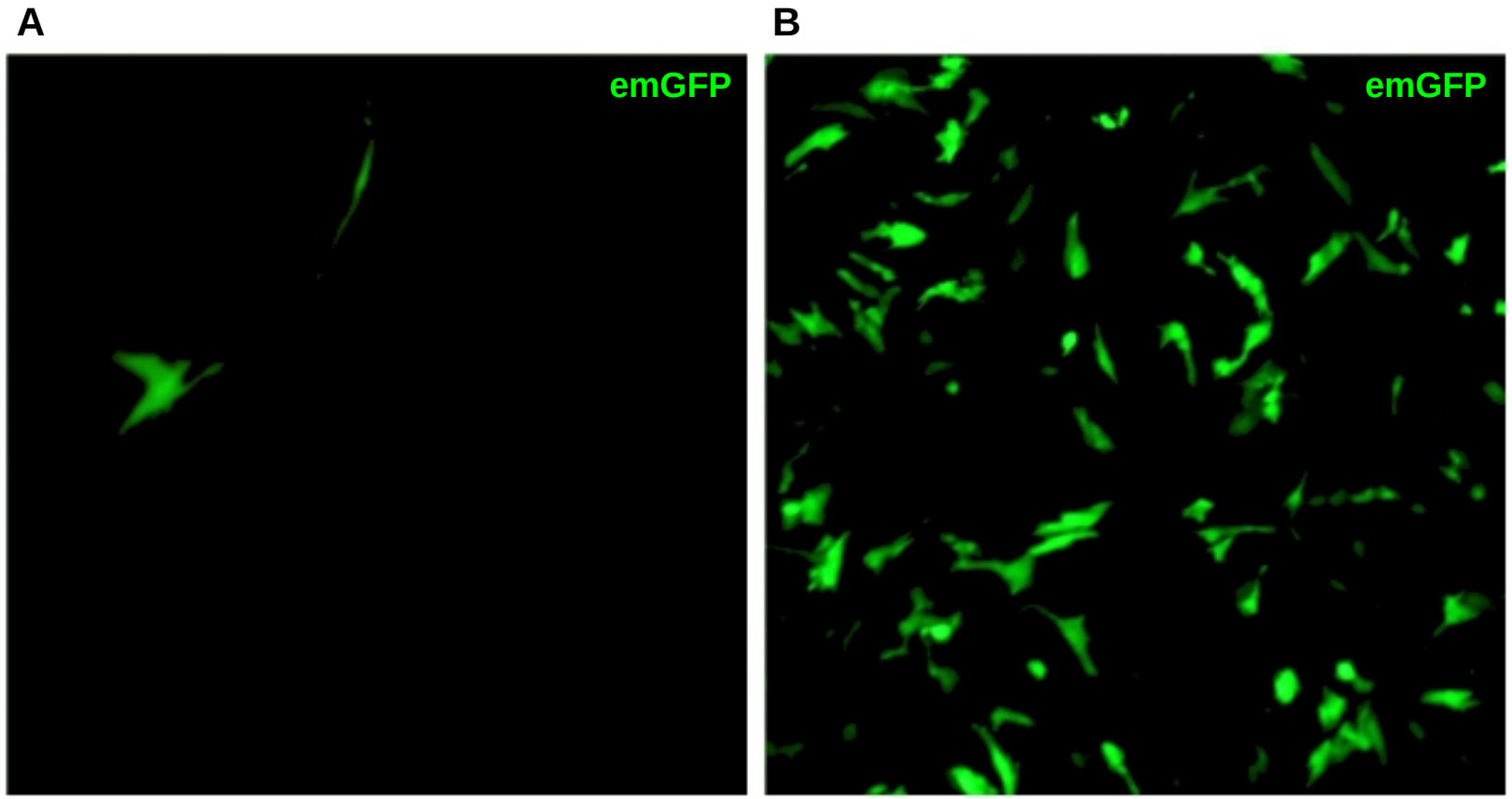

Figure 2: Fluorescent Images of BJ Fibroblasts Transfected with emGFP: Lipofectamin2000 vs. Neon

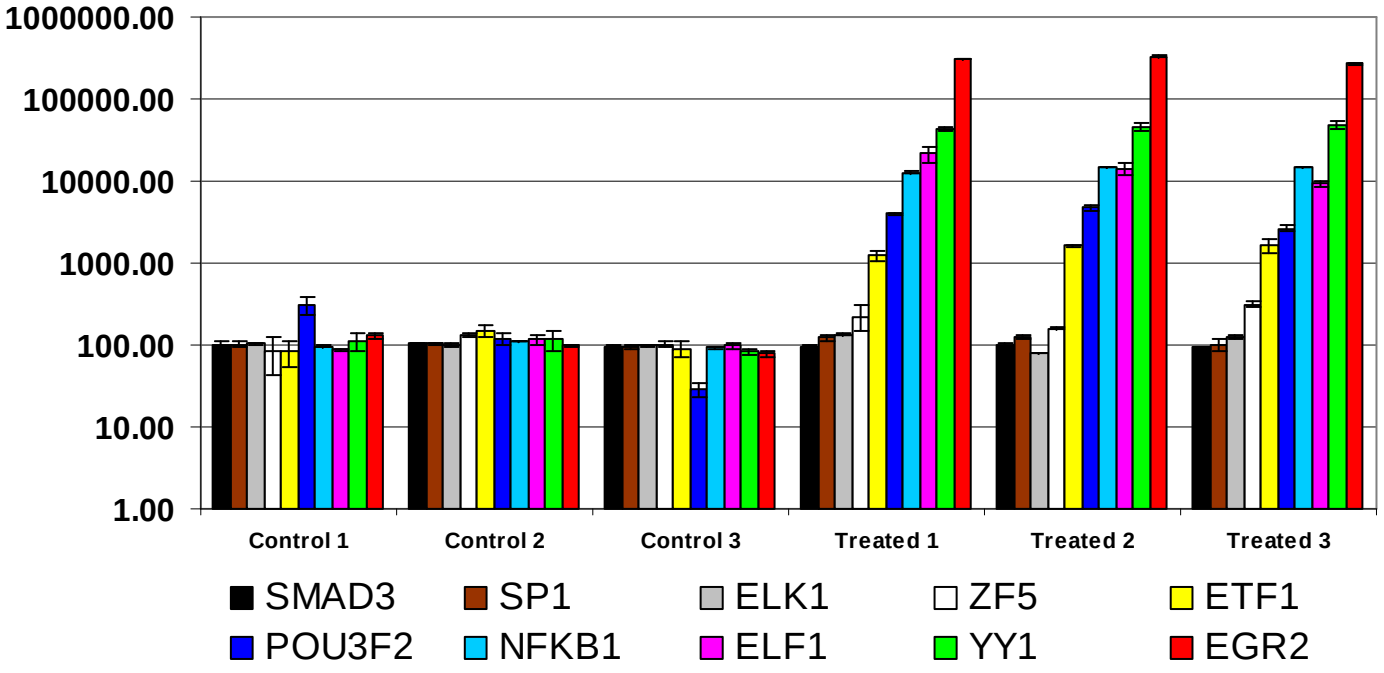

Figure 3: Overexpression Levels of 10 DiRE-predicted Transcription Factors

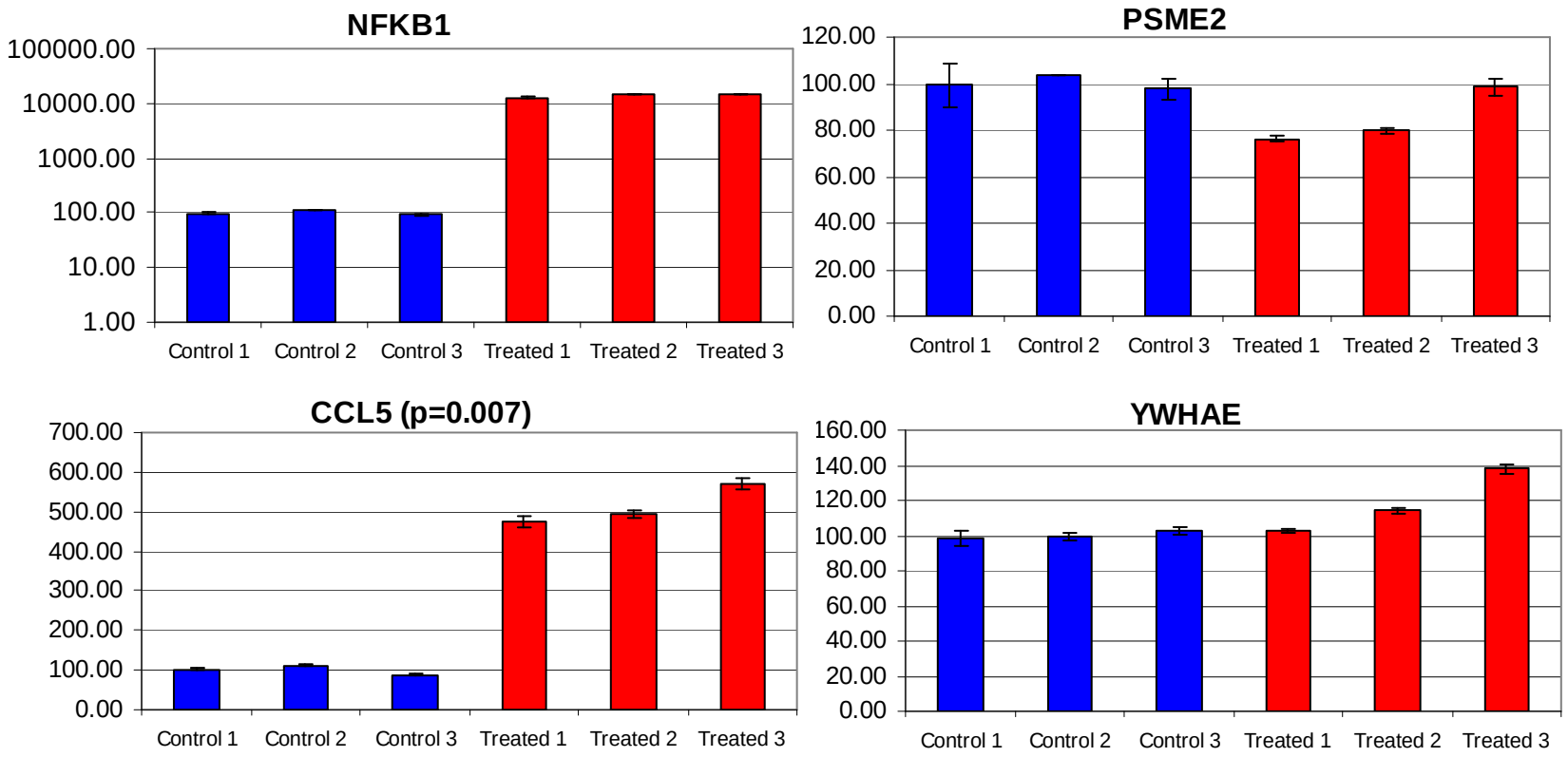

Figure 4: Downstream Gene Target Expression Following NF-κβ1 Overexpression

Table 5: Summary of Expression Levels of 10 DiRE-predicted TFs and Their Downstream Targets

T.F.	O.E. Level	Targets	p≤0.05	50% Δ in Expression
EGR2	3022X	NF-KB1, PARP1, THBS1	0/3	0/3
YY1	455X	IL18↓*, CYC1, PPP2R2A, OAZ1, YWHAE↕*	2/5	0/5
ELF1	150X	RELA, NF-KB1	0/2	0/2
NF-KB1	140X	PSME2, CCL5↕*†, YWHAE	1/3	1/3
POU3F2	38X	TOP1, ACTG2	0/2	0/2
ETF1	15X	YWHAE, OAZ1, NF-KB1↕*, PPP2R2A	1/4	0/4
ZF5	2.3X	YWHAE↕†, OAZ1, NF-KB1, PPP2R2A↕†	0/4	2/4
SP1	1.2X	CRABP2, TGFB1, RELA, NF-KB1, CYC1, PPP2R2A↕*, OAZ1	1/7	0/7
ELK1	1.1X	PSME2↓*, RBBP4↓*, ODC1↓*, YWHAE	3/4	0/4
SMAD3	1X	IFI 6, NF-KB1	0/2	0/2

* p≤0.05, † 50% Avg. Δ, ↑↓ Up/Down Δ in Expression, ↕ Consistent with Scleroderma

CHAPTER 3

Combinatorial Effects of TFs on Scleroderma Gene Targets

Transcription Factors Work Together to Provide Synergistic Effects

Transcriptional activators bind to otherwise silent genes to stimulate their expression. The specificity of transcription factors lies in their ability to physically bind to specific short DNA sequences known as cis-regulatory elements and to recruit numerous co-factors, necessary to recruit RNA Polymerases and other proteins to initiate and maintain transcription. Recurrence of cis-elements at multiple promoters results in activation of all genes that share the same DNA binding sequence. Thus, the concept of gene batteries has emerged. In addition, while some TFs work independently, the majority of TFs work in cooperation with others. In bacteria and yeast experiments, multiple factors are required to form the transcription initiation complex and begin the process of transcription (Ptashne et al., 1997). Typically, the recruitment of multiple components or factors for transcription provides synergistic benefits and increases the specificity of gene activation. TFs are also known to regulate other TFs and it has been suggested that there is an important additional contribution from cooperative DNA binding. In a study by Takahashi et al., the successful induction of Oct4, Sox2, Nanog, and Klf4, which leads to induced pluripotent stem (iPS) cells from fibroblasts was reported. Their experimental approach involved pooling

transcription factors and using the process of elimination to find the core essential transcription factors.

Clustering of Transcription Factor Binding Sites are Important Elements in Cis-regulatory Modules

There have been many studies and experiments showing that regulatory elements are evolutionary conserved and that they are usually enriched in combinations of transcription factor binding sites (TFBSs). The co-occurrence of TFBSs or grouping of binding sites for the same transcription factors into homotypic clusters of TFBSs (HCTs) has only been observed in invertebrates, mainly *Drosophila* until recently (Gotea et al., 2010). While prior evidence for the functional contribution of HCTs in transcription seems somewhat limited, in a paper published by Gotea et al. these advantages are laid out. HCTs favor lateral diffusion of TFs along a regulatory region, favor high affinity of cooperative binding of TFs, and provide functional redundancy. These very properties led us to examine the effect of pooling TFs to see if there would be an observed synergistic effect on our ability to regulate differentially expressed genes between scleroderma and normal phenotypes.

A cis-regulatory module (CRM), also known as a regulatory element, is a sequence of DNA several hundred base pairs long that TFs bind to and regulate nearby genes. Clustering of TFBSs is a common feature of CRMs; however HCTs in the human genome have not been extensively discovered. Gotea et al. report that evolutionary conserved HCTs occupy nearly 2% of the human genome with more than half of the promoters of human genes containing HCTs. HCTs are normally distributed around the transcription start site (TSS) and almost half of the 487 experimentally validated enhancers contain HCTs, which is 25 fold higher than expected by chance. They have also shown that there is a strong correlation between HCTs and the binding of the enhancer-associated co-activator protein Ep300 (p300). Taken together, these results suggest that HCTs play a powerful role in regulatory elements or CRM. We plan to exploit that characteristic in our efforts to predict downstream targets of endogenously expressed TFs.

Predicting Downstream Gene Targets of Known TFs

SynoR (Genome Miner for Synonymous Regulation) is a web-based bioinformatic tool that allows the end user to carry out genome-wide scans for REs using evolutionary conserved TFBSs (cTFBSs) motifs. SynoR uses known TFBS structures of REs, which are defined as a cluster

of TFBSs and their specific spatial order and distribution, to search for novel REs that are different in sequence from original REs, but are synonymous in regulation. Synonymous gene regulation is ultimately defined as regulatory elements that drive shared spatial/temporal aspects of gene expression (Ovcharenko et al., 2005). It is believed that synonymous gene regulation is predicated on regulatory elements that contain similar modules or clusters of TFBSs. SynoR identifies synonymous regulatory elements (SREs) in vertebrate genomes and performs a *de novo* identification of these SREs using patterns of TFBSs in known regulatory elements. Alternatively from our previous approach to identify TFs that regulate scleroderma downstream gene targets, we aimed to predict the downstream targets of endogenous TFs differentially expressed between normal and scleroderma disease phentoypes.

MATERIALS & METHODS:

Predicting Downstream Targets of Endogenous TFs with SynoR (a Genomic Miner for Synonymous Regulation) (http://synor.dcode.org/)

TFs of interest, NF-KB1 and POU3F2, were put individually into the TFBS cluster specifications box. Count, strand, and distance limitations between neighboring binding sites were all optimized to return fewer than 1000 TFBS clusters. The base genome used was the Human genome build 18 (hg18) and the comparison genome used was the Mouse genome build 9 (mm9). Output data was analyzed and downstream gene targets were selected based upon criteria discussed in this chapter.

Cloning Transcription Factors and Vectors Used

All 10 TFs used in this chapter were previously cloned and grown up using the methods described in Chapter 2.

Cell Culture and Transfections

Dermal fibroblasts (BJ) and HaCaT keratinocytes cell lines were thawed from cell stock and cultured in D10 media (DMEM plus 10% Fetal Calf Serum (FCS)) with penicillin-streptomycin bacterial antibiotic. Cells were passaged using Trypsin and stored in a 5% CO_2, 37°C

incubator. Dermal fibroblasts and HaCaT cells were transfected using the standard electroporation provided by Invitrogen for the Neon electroporation unit. A 24-well cell-specific optimization was completed to determine the optimal parameters for the Neon, which was 1 pulse with a pulse width of 20ms, and a pulse voltage of 1650 volts (also optimized by Invitrogen). The same parameters were also used for the HaCaT cell line.

Quantitative Polymerase Chain Reaction (qPCR)

Total RNA was extracted from BJ and HaCaT cells using the Zymo RNA Purification Kit. cDNA was synthesized using the Fermentas Maxima First Strand cDNA Synthesis Kit. Primers were designed using sequence information from Ensembl and Primer3 to optimize conditions. The qPCR primers were designed to span exon-exon gap junctions to eliminate non-specific binding to possible genomic DNA contamination. Primers for downstream targets of DiRE-predicted TFs were the same as in Chapter 2. The 5 qPCR primer sets used in this chapter are:

"Species-Gene-Exon-Direction; Sequence"

Hu-TRPV5-Exon7-F: GGAGCTTGTGGTCTCCTCTG
Hu-TRPV5-Exon8-R: GAAACTTAAGGGGGCGGTAG

Hu-BCAS3-Exon2-F: CGTGAGCAACCCAACAGTAA
Hu-BCAS3-Exon3-R: TTGCTGGTACCTACGGGAAG

Hu-SDC4-Exon4-F: CCACCGAACCCAAGAAACTA
Hu-SDC4-Exon5-R: AGGAAGACGGCAAAGAGGAT

Hu-KREMEN2-Exon2-F: GAATGCTTCCAGGTGAATGG
Hu-KREMEN2-Exon3-R: AGATGCCCTCCTCTGTCTCA

Hu-PDZD2-Exon1-F: CAGCTGATGGTTGGAGTTGA
Hu-PDZD2-Exon2-R: GTCACCCAGCTCCAAGGTAG

qPCR protocol was conducted using a Roche480 Light Cycler and SYBR Green reagent with the following per sample mix:

Reagent:	**Volume (µL):**
SYBR Green	5
Water	2
Primer (10 pM)	.5
cDNA	2.5
Total	**10**

Recorded cycle numbers were input in Microsoft Excel in order to analyze relative gene expression, which was normalized to Human Gapdh. The Δ ΔCt method or the comparative method for quantifying relative gene expression data was used for analysis.

Imaging : Microscopy

Light and fluorescent images were taken of transfected cells in culture using Olympus MVX10 and Olympus BX51 microscopes. Software programs were used to adjust image settings, exposure time, etc.

RESULTS:

Pooling and Transfecting 10 DiRE-predicted Transcription Factors

In previous experiments, single DiRE-predicted transcription factors were transfected into dermal fibroblasts and the expression of 8 of the total 36 targets, or approximately 22%, were significantly altered. While 8 of those targets were differentially expressed, only one of those targets had an average change in expression greater than 50%. To determine if the predicted TF might act in synergy, multiple TFs (10) were introduced into dermal fibroblasts, to see if they would have a combinatorial effect and ultimately evoke greater changes in the expression of the downstream gene targets relative to transfecting them individually.

Figure 5 describes the shared predicted downstream targets and the 10 DiRE-predicted TFs. Five target genes, including CCL5, TGFBI, YWHAE, PPP2R2A, and RELA, appeared to have the most overlap as potential targets. There is a variance in the number of TFs predicted to regulate these chosen targets. While up to 7 of the 10 TFs are predicted to regulate YWHAE, only 1 TF is predicted to regulate both CCL5 and TGFBI. If the hypothesis that pooled transcription factors has an additive or synergistic effect on scleroderma-specific targets, then we would

expect no improvement in CCL5 or TGFBI, while those targets with multiple TFBSs should respond in increased activation. In previous experiments using individual TFs, CCL5, PPP2R2A, and YWHAE were found to be statistically significant. Also, CCL5, PPP2R2A, and YWHAE had average changes in expression greater than 50% from the control. However, only CCL5 was both statistically significant and had an average change in expression greater than 50% in the same TF transfected sample. Thus, pooling all 10 TFs would allow us to test whether improved gene expression would be observed for these targets relative to our previous single TF experiment.

Transient co-transfections of GFP expression vector plus overexpression vectors were performed in dermal fibroblasts. Three control and treated biological replicates were used, so three samples had GFP plus an empty overexpression vector and the other three had GFP plus overexpression vector containing a TF. A difference was that by pooling transfections, a smaller amount of each overexpression vector was used than when using single TF vectors. Nevertheless, their expression relative to previous transfections and the 6 previously overexpressed TFs were overexpressed at similar levels. After 48 hours,

total RNA extract was isolated and mRNA expression levels of associated targets were measured via qPCR.

Observing Downstream Gene Target Expression of Pooled TFs in Dermal Fibroblasts

As discussed previously, pooling DiRE-predicted TFs was tested to see if multiple TFs would provide a combinatorial effect thereby increasing or decreasing the expression of downstream gene targets more so than the single TFs alone. **Figure 6** shows the gene expression for the 5 selected downstream targets. There are 3 biological replicates for both control and treated samples. It is clear that there were only subtle changes in gene expression among all 5 targets. In fact, the only target that had a statistically significant ($p=0.009$) change in expression was CCL5, with a minor increase in expression in the TF-treated sample versus control.

Observing Downstream Gene Target Expression of Pooled TFs in Keratinocytes

It is not known what cell type is responsible for the scleroderma phenotype and in Chapter 2 we discussed that the original microarray data we used to find our 165 differentially expressed genes is derived

from a punch biopsy, meaning that it includes a mix of all cell types found in skin. We first focused on fibroblasts, which exist in the dermis layer of the skin. The dermis layer of the skin also has large amounts of collagen and extracellular matrix, which is primarily synthesized from fibroblasts. Collagen and extracellular matrix are proteins that are upregulated in the scleroderma phenotype. While it makes sense to look at fibroblasts as a primary cell type, it is also quite possible that keratinocytes, which exists in the epidermis, could have a functional role in the scleroderma phenotype. Keratinocytes are the predominant cell type in the epidermis making up approximately 95% of all cells (McGrath et al.). To test the hypothesis that our DiRE-predicted TFs may operate differently in another cell type to differentially express the scleroderma gene targets, we repeated the transfection of all 10 pooled TFs in a keratinocyte cell line, called HaCaT.

A summary of the 10 pooled TF transfections in both BJ and HaCaT cell lines is contained in **Table 6**. Subtle changes in gene expression were observed in HaCaT cells as well as no targets had a 50% average change in gene expression between both cell lines. However, 2 out of the 5 targets measured in HaCaT cells had statistically significant changes in gene expression. Those two targets were

PPP2R2A and RELA. Both of these targets were downregulated in the treated samples suggesting that one or more of the 10 pooled TFs are responsible for repressing their gene expression.

Testing a New Bioinformatic Approach: SynoR

SynoR predicts gene regulation is the opposite direction of DiRE, meaning that instead of predicting TFs from a set of genes, it predicts downstream gene targets that are regulated by the TFs that one inputs. SynoR uses the concept of evolutionary conservation to predict regulatory elements. However, instead of analyzing a TF's ability to bind once in a regulatory element, it can detect multiple transcriptional factor binding sites (TFBSs) in a given RE. This allows for one to search for multiple binding sites or clusters of binding sites in a regulatory element belonging to either one or multiple TFs, therefore predicting synonymous regulation (SynoR). Theoretically, it makes sense that this method may be more predictive than DiRE due to the simple fact that multiple TFBSs would appear to be more powerful than one.

Using the bioinformatics tool SynoR, TFs NF-KB and POU3F2 were entered into the web-based software in order to find downstream gene targets that contained regulatory elements with clusters of either NF-KB

or POU3F2 TFBSs. Once the output was received, the following three characteristics were used to select the downstream targets. The median length of a cluster of TFBSs is 597 base pairs (bps), the median numbers of TFBSs in a cluster is 5, and most clusters of TFBSs exists in close proximity to the transcription start site (TSS) (Gotea et al., 2010). With this in mind, gene targets were selected that had clusters of 5 TFBSs or more, were less than 600 bps in length, and exists in the promoter region of the gene. Picking regulatory elements that exist in the promoter region is also consistent with previous analysis. Two targets were selected for NF-KB and three were selected for POU3F2 as shown in **Table 7**. Also portrayed, are the locus in which the cluster exists, the length of the cluster, and the number of TFBSs contained in the cluster.

In order to perform an initial test for this method, transient co-transfections of NF-KB and POU3F2 that were identical to those in Chapter 2 were performed. These two TFs were transfected individually into dermal fibroblasts, total RNA was extracted after 48 hours, and gene expression was measured via qPCR.

Analysis of SynoR-predicted Target Gene Expression

Of the 5 targets analyzed, PDZD2 and TRPV5 were not abundantly expressed high in dermal fibroblasts to accurately assay change in gene expression. Therefore, 3 initial SynoR targets were left to analyze. Of the remaining 3 targets (**Figure 7**), KREMEN2 and BCAS3 both had over a 50% change in average gene expression. KREMEN2 and BCAS3 had increases in expression relative to control of 435% and 181%, respectively. Although, out of the three targets KREMEN2, BCAS3, and SDC4, only BCAS3 had a statistically significant change in gene expression ($p=0.009$). For an initial validation of 3 targets, SynoR appears to be somewhat predictive for determining downstream gene targets of TFs.

DISCUSSION:

Previously, it was proven that DiRE had the potential to predict TFs that would differentially regulate our set of downstream gene targets. In our experiment in this chapter, we wanted to see whether pooling all 10 DiRE-predicted TFs would provide a combinatorial effect, thereby increasing the percentage of genes differentially expressed in our experiment. From the transfection of the 10 pooled TFs, we observed that 1 out of 5 (20%) targets were differentially expressed. This target was CCL5 and it was differentially expressed as determined by statistical significance. The increase in CCL5 expression was also consistent with the scleroderma phenotype, as it is increased in scleroderma patient samples. Moreover, none of the 5 targets analyzed had an average change in expression greater than 50%, which was observed in our original single TF transfection experiment. Also, YWHAE and PPP2R2A were differentially expressed in our initial experiment and they were not with the pooled TFs. From this data, we can conclude that pooling the TFs had no effect on differentially expressing downstream targets via TF regulation relative to transfecting single TFs. However, in order to determine statistical significance between the two methods, additional targets would need to be analyzed.

We decided to test our pooled TFs in an additional cell line to see if regulation of a portion of the downstream targets was specific to a certain type of cells. The largest cell type (95% of cells) of the epidermis is keratinocytes and that sparked our interest to repeat the previous experiment in the HaCaT cell line, which are immortalized human keratinocytes. In this experiment, 2 out of the 5 targets (PPP2R2A and RELA) were differentially expressed. Interestingly enough, RELA was unable to be differentially expressed in previous BJ cell line experiments and PPP2R2A expression decreased as opposed to the increase observed previously. These results support the hypothesis that multiple cell lines might be responsible for the total change in gene expression between diseased and normal patient samples.

In this section, we also conducted an initial test of a new bioinformatic tool called SynoR, which stands for synonymous regulation, or predicts downstream gene targets based upon clustering of TFBSs. Of the three SynoR downstream targets measured, one had a statistically significant change in expression (BCAS3). On top of that, 2 (KREMEN2 and BCAS3) of the 3 targets had an average change in expression greater than 50%. These initial results seem promising and they force one to ask additional questions regarding the approach.

Does the number of binding sites in a cluster, or in general, correlate with change in gene expression? Does the closeness in base pairs of TFBSs within a cluster effect the level of gene expression? These are questions that need to be answered by conducting additional experiments and testing more SynoR targets.

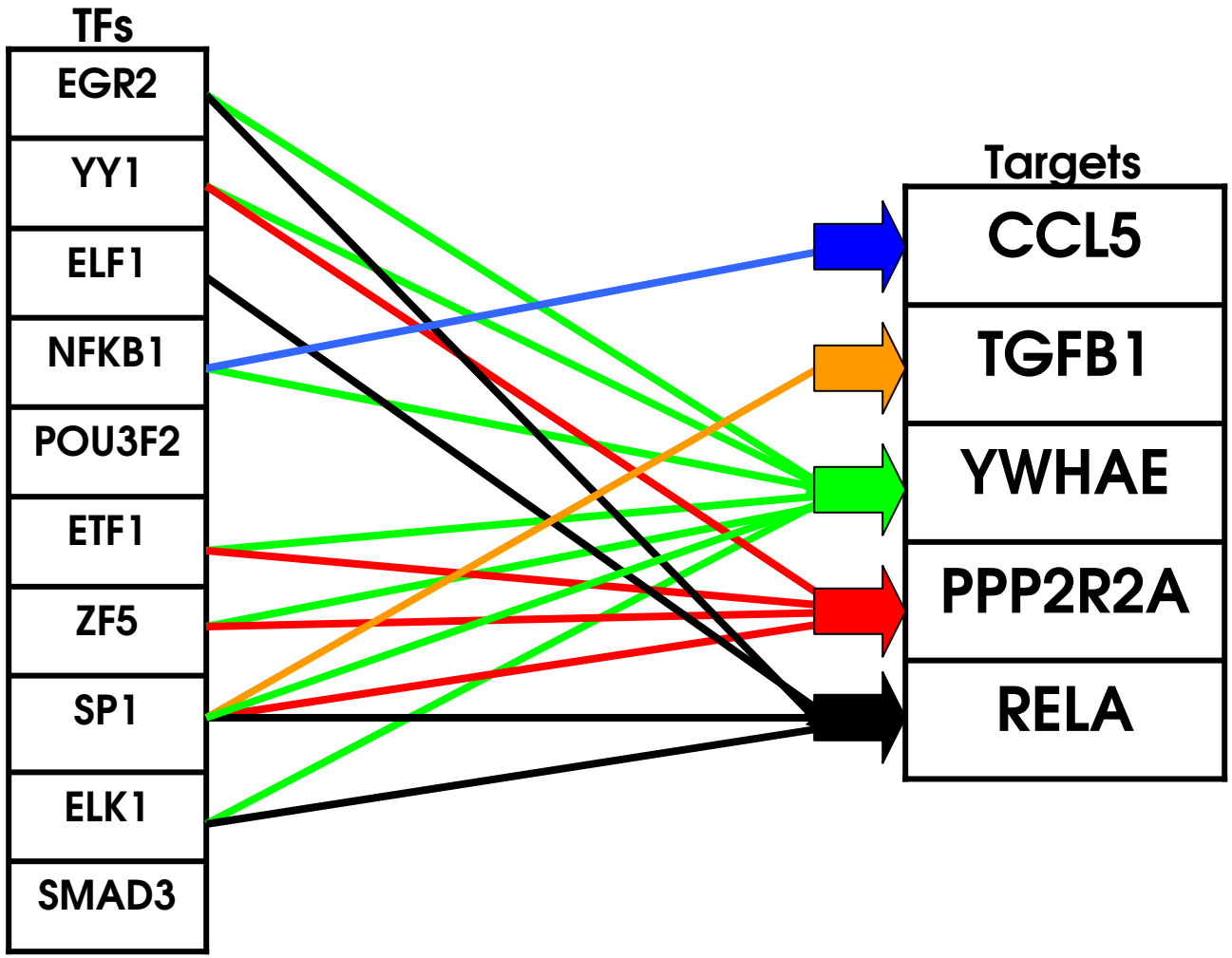

Figure 5: Relationships of Gene Regulation between 10 Pooled TFs and Downstream Targets

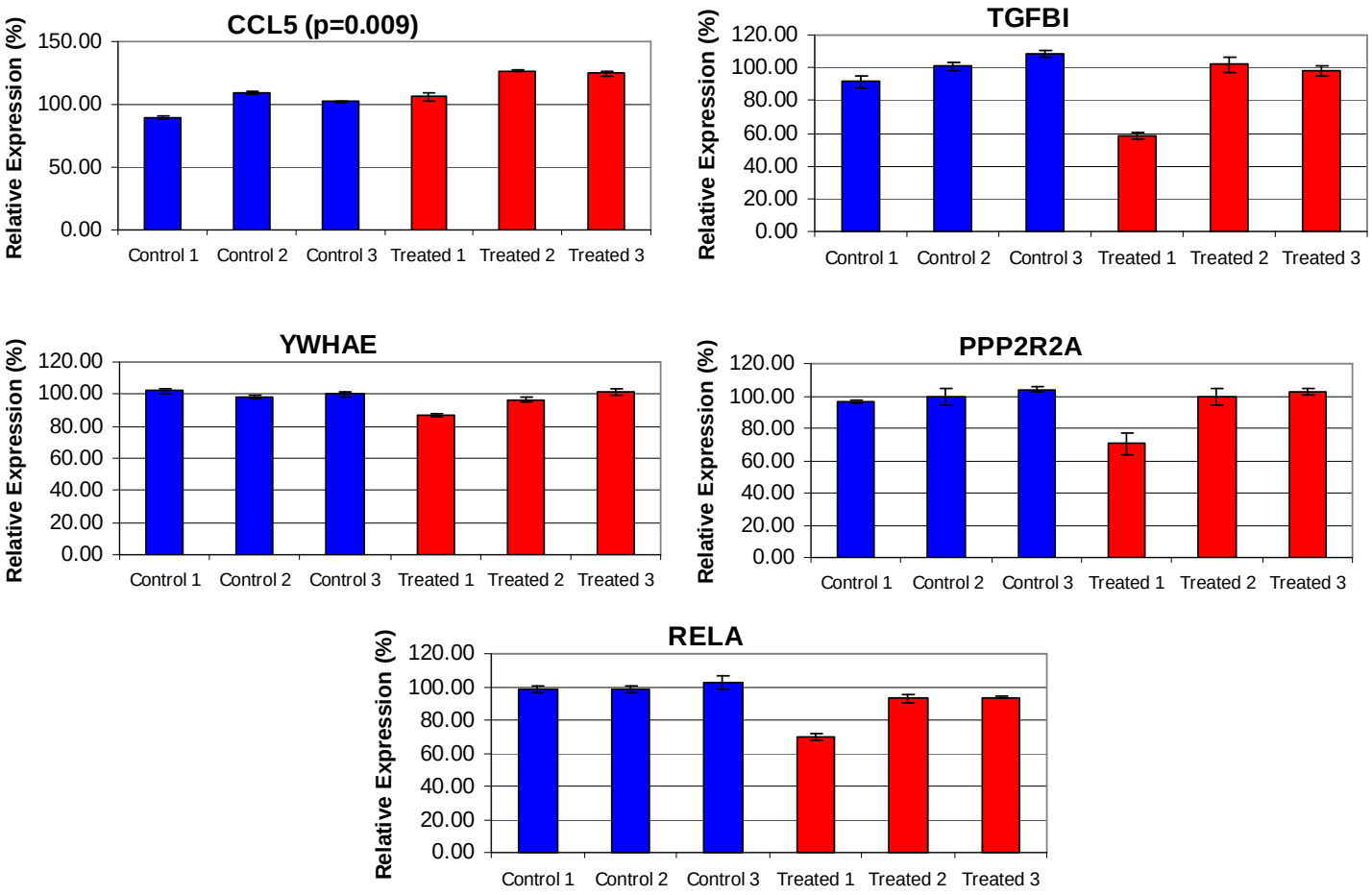

Figure 6: Gene Expression Levels of 5 Downstream Targets in BJ Cells with 10 Pooled TFs

Table 6: Summary of Downstream Gene Target Expression from 10 Pooled TFs in BJ and HaCaT Cells

TFs	Cell Line	Targets	$p \leq 0.05$	50%Δ in Expression
EGR2, YY1, ELF1, NF-KB1, POU3F2, ETF1, ZF5, SP1, ELK1, SMAD3	BJ	YWHAE, PPP2R2A, RELA, CCL5↕*, TGFB1	1/5	0/5
	HaCaT	YWHAE, PPP2R2A↓*, RELA↓*, CCL5, TGFB1	2/5	0/5

* $p \leq 0.05$, ↑↓ Up/Down Δ in Expression, ↕ Consistent with Scleroderma

Table 7: Properties of SynoR Predicted TFBS Clusters and Downstream Targets

TF	Target	Locus	Length	# TFBSs
NF-KB1	KREMEN2	chr16:2953432-2953522	91 bps	13
	PDZD2	chr5:31674868-31674920	53 bps	12
POU3F2	TRPV5	chr7:142342303-142342348	46 bps	9
	BCAS3	chr17:56108782-56108827	46 bps	9
	SDC4	chr20:43411443-43411484	42 bps	8

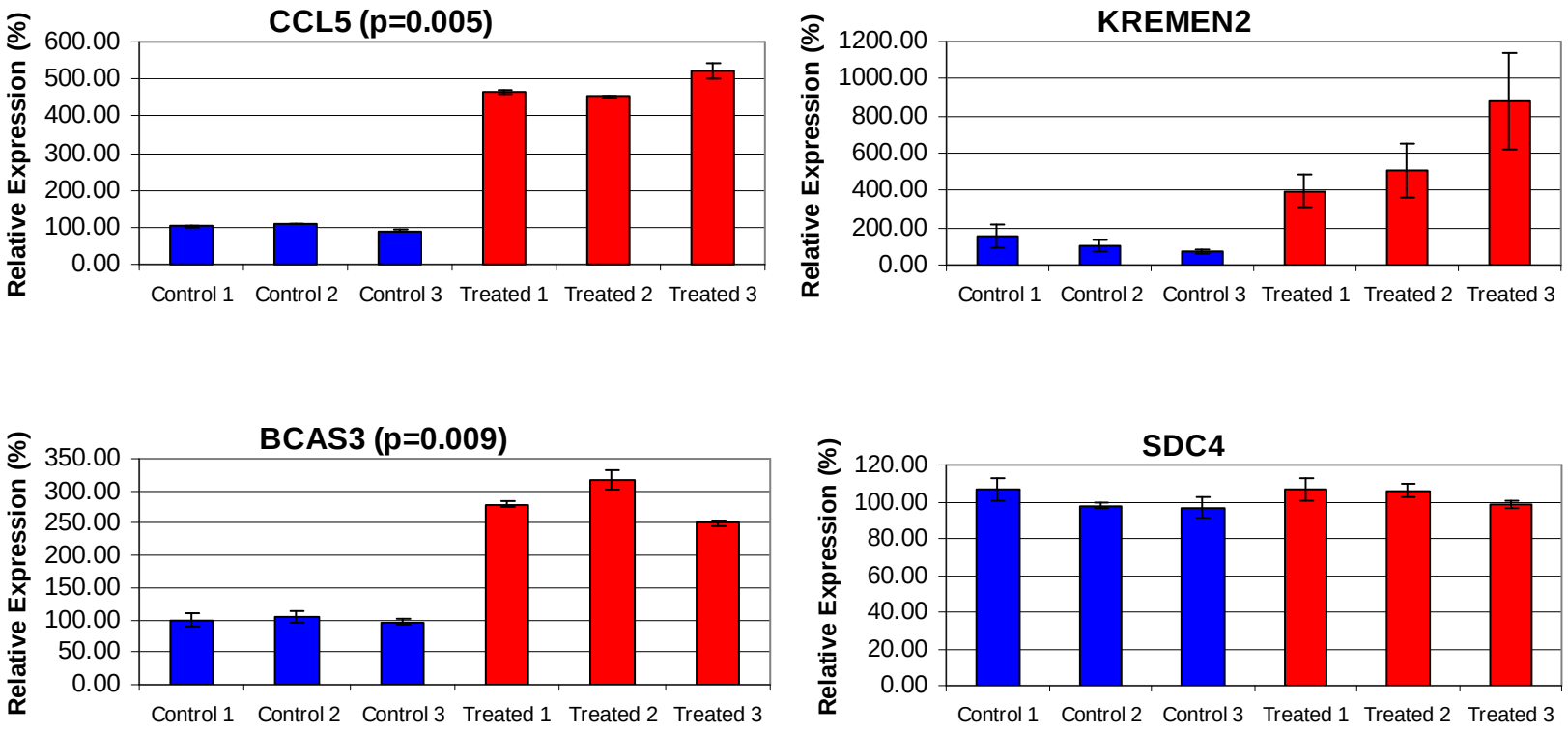

Figure 7: CCL5 and 3 SynoR-predicted Downstream Targets

CHAPTER 4

An Epigenetic Approach: Testing Expression of Identified Transcription Factors and Downstream Targets in Human Embryonic Stem Cells

Embryonic Stem Cells Represent an Early Epigenetic State

Recently, scientists have been able to isolate embryonic stem (ES) cells from the inner cell mass of blastocysts and culture them. A characteristic feature of stem cell is their ability to proliferate indefinitely while maintaining the ability to differentiate into various cell types. Mouse ES cells are so potent that an entire mouse can be produced from an ES cell. While ES cells are transcriptional active, most protein expression is at low levels. As these ES cells differentiate into various cell types, transcription activity actually decreases enhancing the effect of relatively higher expressed genes and protein expression increases resulting in a change in phenotype. Numerous studies have indicated that TFs are responsible for the change in gene expression that results from the differentiation of an ES cell to another cell lineage.

Epigenetics has been defined as "the study of changes of gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence" (Bibikova et al., 2008). Epigenetic factors that help define a particular cell state include: DNA methylation, transcription factors, changes in chromatin structure (through binding of histones), and possibly microRNAs. Understanding how these epigenetic factors affect various cell types is crucial to

elucidating the vary mechanisms that define differentiation. Scientists in the Yamanaka Lab have recently showed that introducing transcription factors to change cell epigenetics can results in changing a cell state, such as changing a fibroblast to an induced pluripotent (iPS) cell. It is known that ES cells have a non-condensed chromatin structure. This characteristic defines an ES cell as an undifferentiated cell or a cell in an early state that is transcriptionally active and is partially responsible for its ability to turn into any cell type or cell lineage.

Gene Expression is Regulated by Epigenetic Modifications

DNA methylation and histone modifications are two of the largest epigenetics mechanisms thought to regulate gene expression. DNA methylation is known to establish a silent chromosome state by collaborating with other proteins that help modify nucleosomes. DNA methylation can act to either activate or repress gene expression. Histone modifications also effect gene expression by keeping the organism's DNA in an "open" or "closed" state having a profound effect on transcription. From the point at which the embryo is fertilized rapid demethylation across the genome occurs. As the zygote starts to mature and eventually differentiate into different cell types, *de novo* methylation occurs making those tissues less transcriptionally active.

Epigenetic factors such as DNA methylation and histone modifications are becoming increasingly understood and they have been shown to have significant effects on overall gene expression patterns in various cell types. We postulate that by introducing our predicted and endogenous TFs into Human embryonic stem cells (hESC), there will be fewer road blocks that will allow for better access of our TFs to DNA, thereby allowing us to better mimic the pattern of scleroderma-specific gene expression.

MATERIALS AND METHODS:

Predicting Downstream Targets of Endogenous TFs with SynoR (a Genomic Miner for Synonymous Regulation) (http://synor.dcode.org/)

TFs of interest ETF1 and VDR, were put individually into the TFBS cluster specifications box (NF-KB1 and POU3F2 were done previously in Chapter 3). Count, strand, and distance limitations between neighboring binding sites were all optimized to return fewer than 1000 TFBS clusters. The base genome used was the Human genome build 18 (hg18) and the comparison genome used was the Mouse genome build 9 (mm9). Output data was analyzed and downstream gene targets were selected based upon criteria discussed in the previous chapter 3.

Cloning Transcription Factors and Vectors Used

Cloning primers were designed and amplified from genomic complimentary DNA (cDNA). Linkers specific to the overexpression vector multiple cloning site (MCS) were added along with the Kozak Consensus Sequence, which plays a major role in the initiation of translation. The amplified gene of interest was extracted from 1% agarose gel using standard protocol for the Zymoclean Gel DNA Recovery and Cleanup Kits. The overexpression vector of interest,

(**Figure #**) pcDNA 3.0, was digested at its MCS and the VDR gene was ligated into it. The resulting overexpression vector was grown up and checked with sequencing. Of the 7TFs used, VDR was cloned and the other 6 were cloned or produced previously (Chapter 2).

Cell Culture and Transfections

Human embryonic stem cells (hESC) were thawed from cell stock and cultured in mTESR1 media by Stem Cell Technologies (mTESR1 Basal Medium plus 20% 5X mTESR1 Supplement) with penicillin-streptomycin bacterial antibiotic. Cells were passaged using Dispase and stored in a 5% CO_2, 37°C incubator on matrigel-coated plates. hESC were transfected using the standard Lipofectamin 2000 protocol provided by Invitrogen.

Quantitative Polymerase Chain Reaction (qPCR)

Total RNA was extracted from hESC cells using the Zymo RNA Purification Kit. cDNA was synthesized using the Fermentas Maxima First Strand cDNA Synthesis Kit. Primers were designed using sequence information from Ensembl and Primer3 to optimize conditions. The qPCR primers were designed to span exon-exon gap junctions to eliminate non-specific binding to possible genomic DNA contamination. Primers

for downstream targets of DiRE-predicted TFs, overexpressed TFs, and SynoR targets were the same as in Chapter 2 and 3. The 1 TF and 8 additional SynoR targets used in this chapter have the following qPCR primer sets:

"Species-Gene-Direction-Exon"	"Sequence"
Hu-VDR-F-Exon4	TGACCCTGGAGACTTTGACC
Hu-hVDR-R-Exon5	GTTGAAGGGGCAGGTGAATA
Hu-GAA-F-Exon1	GATGAGGCAGCAGGTAGGAC
Hu-GAA-R-Exon2	TCACTCCCATGGTTGGAGAT
Hu-PCDH7-F-Exon1	AAAACACCAGGCCGTACAAG
Hu-PCDH7-R-Exon2	TGGAATTCAGCCAAACACAG
Hu-DIVA-F-Exon1	CCCCACTGTACACCAAGGTC
Hu-DIVA-R-Exon2	AACTCTGAATGGGGCACATC
Hu-ARID2-F-Exon4	GTTCATCATTTTGGGGAGGA
Hu-ARID2-R-Exon5	CCACTTCATTTGGGAGTCCA
Hu-PCDH1-F-Exon2	GCTTGACACCAATGACAACG
Hu-PCDH1-R-Exon3	CAGGGTGCTTAGGTCCTCAC
Hu-NOVA2-F-Exon3	CAGCTTTATTGCCGAGAAGG
Hu-NOVA2-R-Exon4	ACCCATGCTCCTGACTGTTC
Hu-QSER1-F-Exon5	ACTGGCGGTAACAGTCCATC
Hu-QSER1-R-Exon6	ACCCACAGTGGTTGAGGAAG
Hu-PRKCD-F-Exon2	TCTGTGCCGTGAAGATGAAG
Hu-PRKCD-R-Exon3	CACGGTCACCTCAGACACTG

qPCR protocol was conducted using a Roche480 Light Cycler and SYBR Green reagent with the following per sample mix:

Reagent:	**Volume (µL):**
SYBR Green	5
Water	2
Primer (10 pM)	.5
cDNA	2.5
Total	**10**

Recorded cycle numbers were input in Microsoft Excel in order to analyze relative gene expression, which was normalized to Human Gapdh. The Δ ΔCt method or the comparative method for quantifying relative gene expression data was used for analysis.

Imaging : Microscopy

Light and fluorescent images were taken of transfected cells in culture using Olympus MVX10 and Olympus BX51 microscopes. Software programs were used to adjust image settings, exposure time, etc.

Immunohistochemistry

Transfected hESC were removed from matrigel-coated plates with Trypsin in order to break hESC colonies into single cells. hESC were then washed and resuspended in PBS, and then cytospinned onto glass coverslips. Standard immunocytochemistry and immunofluorescence protocol by Abcam was used. Samples were fixed in 3-4% paraformaldehyde in PBS at pH 7.4. PBS containing 0.25% Triton X-100

was used for permeabilization of cells. 1% BSA in PBST was used for blocking overnight at 4°C. Monoclonal rat VDR and rabbit NF-κβ (p65) primary antibodies were used in a 1:100 dilution. Goat anti rat Alexa Fluor 568 and goat anti rabbit Alexa Fluor 488 secondary antibodies were diluted 1:500 in PBS and used for visualization. Coverslips were mounted and dried overnight at 4°C in the dark.

RESULTS:

Overexpressing DiRE-predicted and Endogenous TFs into Human Embryonic Stem Cells

Thus far, DiRE and SynoR have proven to be somewhat predictive with 12 out of 49, or roughly 24.5% of downstream gene targets differentially expressed from a statistically significant viewpoint. While this is better than expected, only 5 of the 49, or approximately 10% of the total targets had an average change in expression greater than 50% and just one gene was both significantly differentially expressed and had an increase in gene expression greater than 50%. One of the classic problems affecting gene expression involves epigenetic marks and chromatin modeling. Epigenetic modifications made to DNA over time or throughout development can determine what genes are expressed. A unique characteristic about stem cells are that they exhibit an earlier epigenetic state, therefore they have different epigenetic modifications and their chromatin is less condensed. This allows stem cells to have a more open form of DNA relative to somatic cells, thus allowing TFs better access to regulate downstream gene targets. Due to these properties of stem cells, we decided to transfect our DiRE-predicted and endogenous transcription factors into Human

embryonic stem cells (hESC) to discover the effect that it would have on the expression of downstream targets.

A total of 7 TFs were chosen for this experiment and they included: YY1, ETF1, ELF1, NF-KB1, EGR2, POU3F2, and VDR. 6 (excluding VDR) of the 7 TFs were DiRE-predicted, but 2 (NF-KB1 and VDR) of the 7 were also TFs that were a part of the 165 differentially expressed targets. 6 of the 10 original DiRE-predicted TFs were used for the experiement, because those were able to be overexpressed. The fact that they were able to be overexpressed makes it easier to draw a correlation between overexpression of the TF and change in downstream gene expression. As determined from microarray analysis, both NF-KB1 and VDR were expressed higher in the scleroderma disease patient samples. If in fact, overexpressing scleroderma-specific endogenous TFs is the key to reprogramming the scleroderma disease phenotype as opposed to altering gene targets that were differentially expressed, then applying SynoR to these two TFs would be an interesting method to predict their downstream gene targets. Finally, in this experiment both DiRE-predicted and endogenous TFs were combined not just to further validate DiRE and SynoR, but in order to see if hESC were more receptive to reprogramming.

All 7 TFs were pooled and transiently transfected into hESC along with a GFP reporter plasmid. After 24 hours, total RNA was extracted and gene expression was measured. 24 hours (different from 48 with fibroblasts) was chosen as a time point here, because hESC divide faster and they appear to differentiate after 48 hours of transfection (**Figure 8A**). In the initial experiment conducted, the hESC looked elongated and stringy similar to fibroblasts as opposed to small compact hESC that form a colony. If the hESC were allowed to differentiate and their cell state changed, then it is fair to say that their epigenetic state and gene expression would be altered, which would greatly affect the outcome of the experiment. **Figure 8B** displays the hESC after 24 hours of transfection maintaining their small, round, and compact morphology that is characteristic to stem cells (Note **Figure 8A** is magnified many time relative to **Figure 8B**). **Figure 9** displays the gene expression of each of the 7 transfected TFs. All 7 TFs were overexpressed more than 10 fold, with most of them being overexpressed at least 100 fold relative to control samples. 5 TFs were expressed greater than 100 fold, 3 were greater than 1000 fold expression, and the highest express TF was VDR with an approximate 8500 fold overexpression.

The main purpose of this hESC experiment was to observe if cells that are in an earlier epigenetic state are more receptive to TFs and altering gene expression. First, the overexpression of the 6 TFs that were transfected into both dermal fibroblasts and hESC were compared. **Table 8** details the overexpression of the 6 TFs in both BJ and hESC. Two of the 6 TFs had a decrease in overexpression in hESC relative to BJ Cells. The decrease in EGR2 overexpression was modest with an 11% decrease in overexpression, while the YY1 TF had an 89% decrease in overexpression in hESC relative to dermal fibroblasts, or the overexpression in hESC was 11% of that observed in dermal fibroblasts. The minor fluctuation in EGR2 overexpression might be due to different transfection methods (Electroporation versus Liquid Transfection) or the variation in overexpression plasmid amounts used between the two protocols. However, the drastic decrease in overexpression of YY1 is too significant to be attributed to protocol variations alone. It is possible that hESC have a large amount of YY1 expression relative to Bj cells making it more difficult to overexpress YY1. In fact, qPCR cycle numbers confirmed this as YY1 control samples had higher YY1 expression relative to the Gapdh control in hESC versus BJ cells. Cycle differences were approximately 2-3 cycles between cell lines.

Turning back to the larger picture, 4 out of the 6 TFs had drastic increases in expression in hESC relative to dermal fibroblasts, with POU3F2 having the largest increase in overexpression of 10,561%. 2 of those 4 TFs had different endogenous expression levels between cell lines. NF-KB1 was higher expressed in BJ cells and ETF1 was higher expressed in hESC. Part of the 432% change in overexpression of NF-KB may be due to the fact that it is lower expressed in hESC, while ETF1 being higher expressed in hESC should have actually limited the change in overexpression. Worth noting, for the three changes of endogenous TF expression between cell lines we observed cycle differences of 1-3 cycles, so this should not cancel out the overall increase of TF expression in hESC relative to BJ cells. The average change in overexpression of these 6 TFs relative to BJ cells was observed to be 1952%.

Observing Downstream Gene Target Expression in hESC Transfected with 6 DiRE-Predicted TFs

Now that it is established that all 6 DiRE-predicted TFs were overexpressed in hESC, we wanted to quantitate the expression of the associated downstream gene targets to determine if their expression

was differentially regulated in hESC as opposed to dermal fibroblasts. One advantage of pooling the 6 TFs as discussed in Chapter 3, is that several TFs have common downstream gene targets. Therefore, by analyzing the expression of one gene, multiple targets are being analyzed simultaneously. **Table 9** displays the 6 TFs, their overexpression levels, and the gene targets that DiRE predicts them to regulate. There are 25 targets in total and genes such as YWHAE are predicted to be regulated by 4 out of the 6 TFs.

From our analysis of 12 downstream genes (25 targets total), 2 genes, CCL5 and ACTG2, were differentially expressed. Expression graphs of both CCL5 and ACTG2 are displayed in **Figure 10**. CCL5 and ACTG2 have p values of 0.02 and 0.01, respectively. **Table 10** shows p values for all 12 genes analyzed in both hESC and BJ cells (data from Chapter 2). While 2 genes (CCL5 and ACTG2) were differentially expressed in hESC, 3 genes (CCL5, YWHAE, and IL18) were differentially expressed in the original dermal fibroblast experiment with the same 6 TFs. This change in the expression of the 12 gene targets could be due to the two different cell lines, or the pooling of the 6 TFs. Additionally, between both cell line-specific data sets, only one TF, CCL5, was consistently statistically significant and differentially expressed. Also,

between both data sets, only CCL5 had an average change in expression greater than 50%. The 5-fold increase of expression of CCL5 when treated with NF-KB1 was consistent between both data sets with an average increase in gene expression of 427%.

Measuring SynoR-predicted Downstream Gene Target Expression of hESC Transfected with Endogenous Transcriptions Factors

From our initial transfection and analysis of SynoR gene targets (Chapter 3) we concluded that 1/3 targets were statistically significant and differentially expressed and that 2/3 targets had an average change in gene expression greater than 50%. While these initial results are promising, we wanted to look at more SynoR predicted targets to continue to validate the method for reprogramming and observe if changes in gene expression were noticed in hESC relative to dermal fibroblasts.

Table 11 contains the 4 TFs (NF-KB1, ETF1, VDR, POU3F2) used for our SynoR analysis, their overexpression level in our hESC experiment, and their SynoR-predicted downstream gene targets. While only 2 TFs (NF-KB1 and VDR) were differentially expressed in scleroderma microarray data, we provided SynoR analysis for 2 additional TFs (ETF1

and POU3F2) to increase the numbers of targets analyzed for validation purposes. It is worth noting that overexpression of NF-κβ1 and VDR is consistent with the scleroderma phenotype, since they are both high expressed in scleroderma diseased samples. Also provided for each SynoR target, is the locus where the TFBS cluster exists, the length in base pairs of the cluster, and number of TFBSs in the cluster. The same selection criteria used in Chapter 3 for selection of targets was used for these targets. The one exception is that two targets have 4 TFBSs in the cluster, instead of 5 or more.

P values for all ten targets measured are provided, except for DIVA, which was expressed too low in hESC to analyze. 3 out of 9 targets analyzed displayed statistically significant changes in expression and their expression graphs are shown in **Figure 11**. The three differentially expressed targets were QSER1, ARID2, and BCAS3. The statistically significant change in gene expression of BCAS3 was consistent with our earlier findings (Chapter 3) however, BCAS3 and KREMEN2 did not have an average change in gene expression greater than 50% in the hESC experiment like they did in the earlier dermal fibroblasts experiment. The cause of the decrease in expression of these two targets is unknown, since both NF-KB1 and POU3F2 were

significantly more overexpressed in hESC relative to the BJ cell experiment. It is possible that the gene expression differences in hESC are great enough to affect the previous relationships observed before between NF-KB1 and POU3F2 and their SynoR-predicted targets. The only gene target that had an average change in expression of 50% or greater in this experiment was PCDH7, which had an increase in gene expression relative to control of 63%. From the data we gathered with this hESC experiment, SynoR appears to be somewhat predictive for downstream targets as 3 out of 9 targets analyzed were differentially expressed.

Immunohistochemistry: Confirming the Expression of TF Protein

In order for a TF factor to regulate a downstream gene target, the TF gene needs to be transcribed into messenger RNA (mRNA), then that mRNA needs to be translated into a TF protein, and the TF protein physically binds to its downstream gene and recruits various factors to begin transcription, so that the gene may be expressed. For our analysis of gene expression via qPCR, mRNA transcripts are isolated and measured to detect levels of gene expression. To confirm that observed changes in gene expression are due to the overexpression of TFs, Immunohistochemistry was performed to detect expression of TF

protein. 2 out of the 7 transcription factors transfected in this experiment were used for immunohistochemistry. After the hESC transfection was complete, all hESC were stained for NF-KB1 and VDR.

Figure 12 provides ultraviolet (UV) light microscopy images of both control hESC (A and C) and hESC treated with 7 TFs (B and D) stained with primary antibodies (Abs) specific for NF-KB1 and VDR protein. All four images are also stained for 4',6-diamidino-2-phenylindole (DAPI), which is a fluorescent stain that binds strongly to DNA. The DAPI stain is used to stain the nucleus of the hESC and it can be used a reference point when localizing NF-KB and VDR expression. The top two images, A and B, are stained green for NF-KB protein. From these images, it is clear that NF-KB protein is being made and overexpressed in the TF overexpression sample (B), relative to control (A). Also, images C and D are hESC stained for VDR protein in red. Similarly to NF-KB, VDR protein is clearly being produced and overexpressed in the 7 TF treated sample (D) relative to control sample (C) where little or no VDR protein is visualized. From these immunofluorescent images, it is clear that NF-KB and VDR protein was produced in hESC after transfection with overexpression vectors.

DISCUSSION:

The purpose of the set of experiments contained in Chapter 4 was to examine the effect Human embryonic stem cells would have on our ability to alter gene expression using DiRE-predicted TFs and SynoR-predicted downstream targets. Our hypothesis was that hESC, which are in an earlier cellular state and have a less condensed chromatin structure, would be more "open" to reprogramming or the alteration of endogenous gene expression. All 7 of the TFs transfected into hESC were able to be overexpressed. In fact, when the overexpression of the TFs in hESC was compared to their overexpression in dermal fibroblasts, we found that the TFs had an average change in overexpression of 1952% in hESC. These initial TF expression results support the tested hypothesis that hESC are more "open" than somatic cells, or they allow for greater expression of external TFs.

In this hESC experiment, 2 TFs (CCL5 and ACTG2) out of 12 were differentially expressed and statistically significant. Of those two, only CCL5 had an average change in expression greater than 50%, which was consistent with previous results. Worth noting, ACTG2 had an increase in expression, which is consistent with the scleroderma phenotype. When comparing the gene expression of these 12 targets

with earlier results in dermal fibroblasts, the only target between the two cell lines to yield the same observations in expression was CCL5. It is possible that the difference in expression is due to the change in cell type. Of the 12 targets analyzed, 3 were differentially expressed in the dermal fibroblast single TF experiment and 2 in the hESC experiment. Unfortunately, there is not a large enough variation in differentially expressed targets between the two datasets to determine statistical significance or say that hESC allow for greater alteration of gene expression. To do this, a larger sample size is needed and more targets would need to be analyzed.

Additional analysis of the SynoR approach that was previously tested in Chapter 3 was performed. This approach is interesting because it can allow one to overexpress TFs differentially expressed between scleroderma and normal skin phenotypes and predict downstream targets that they regulate. This approach provides an alternative to DiRE and our first attempt at reprogramming in that it is possible that the 15 endogenous differentially express TFs are responsible for the change in phenotype as opposed to the 150 other differentially expressed genes. This approach gives a method to predict the 15 TF's downstream targets. The two endogenous differentially

expressed TFs between scleroderma and normal patient samples that were overexpressed were NF-KB and VDR. Both of these TFs were increased in scleroderma samples. Hence, when overexpressed, the downstream gene targets they regulate might have a role in the scleroderma disease phenotype. To increase our number of SynoR targets for analysis, we also examined SynoR targets of ETF1 and POU3F2 TFs.

3 (QSER1, ARID2, and BCAS3) out of the 9 SynoR gene targets were differentially expressed and statistically significant. An average change in expression greater than 50% was only observed in PCDH7, which was not statistically significant. The differential expression of BCAS3 was consistent with earlier finding in Chapter 3, but it did not have a change in expression greater than 50% as observed in the earlier experiment. KREMEN2 also did not exhibit this increase. This may be due to the "pooling effect" of the TFs that was observed in Chapter 3. It is impossible to say whether the number of TFBSs plays a role in differential expression of gene targets as there was a large variation (4, 9, and 10). On the other hand, the length of base pairs of the TFBS cluster may have some correlation with gene regulation. In the differentially expressed targets, their cluster lengths were 34, 46, and 62

base pairs. The ratio in differential expression was 1:3 SynoR gene targets in both the BJ and hESC experiments, thus it is impossible to say whether hESC had a significant effect on our attempt to alter gene expression. Nonetheless, our results indicate that SynoR can be used to predict downstream gene targets for endogenous TFs.

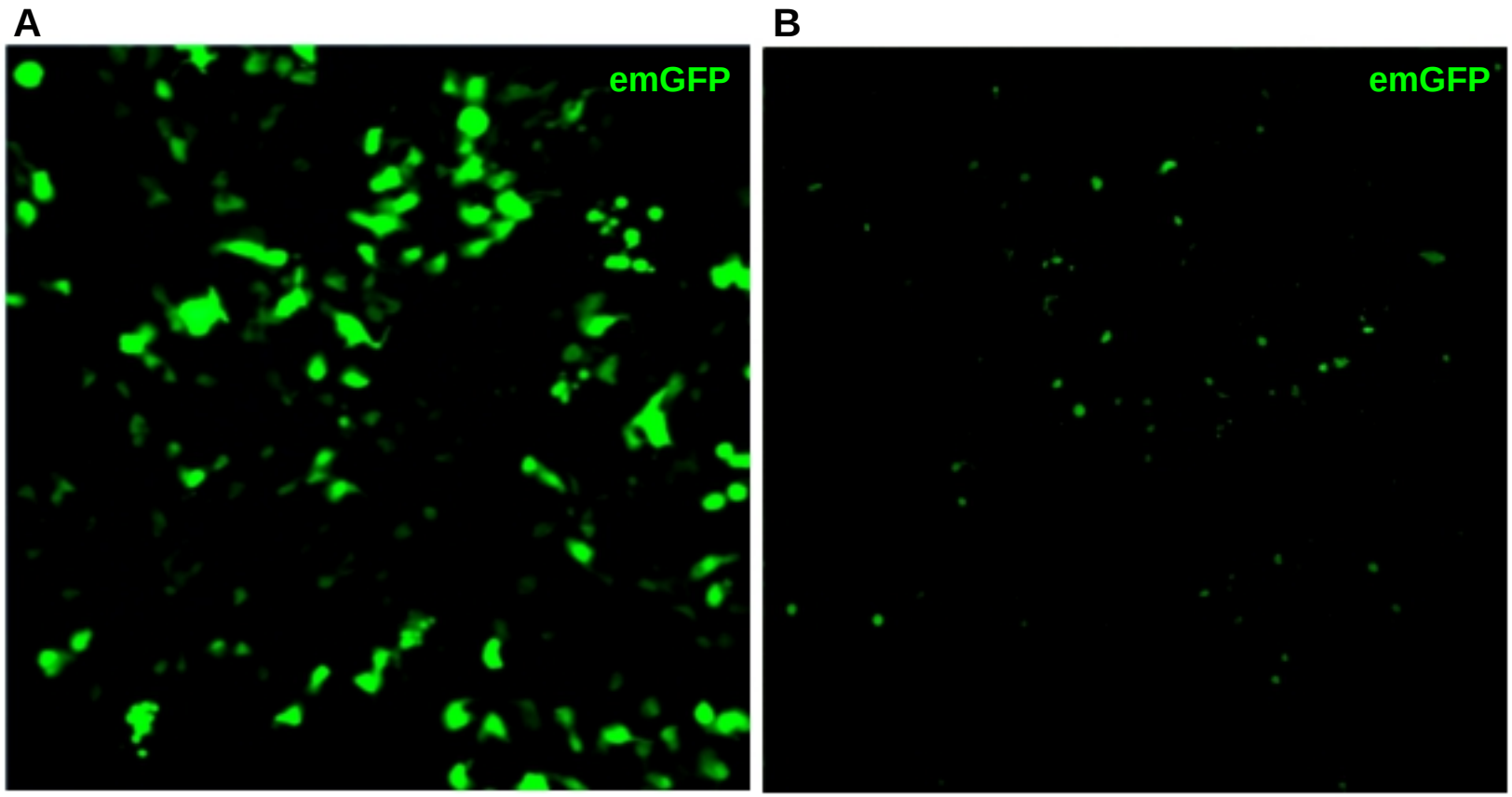

Figure 8: hESC Transfected with emGFP: 48 hours vs. 24 hours

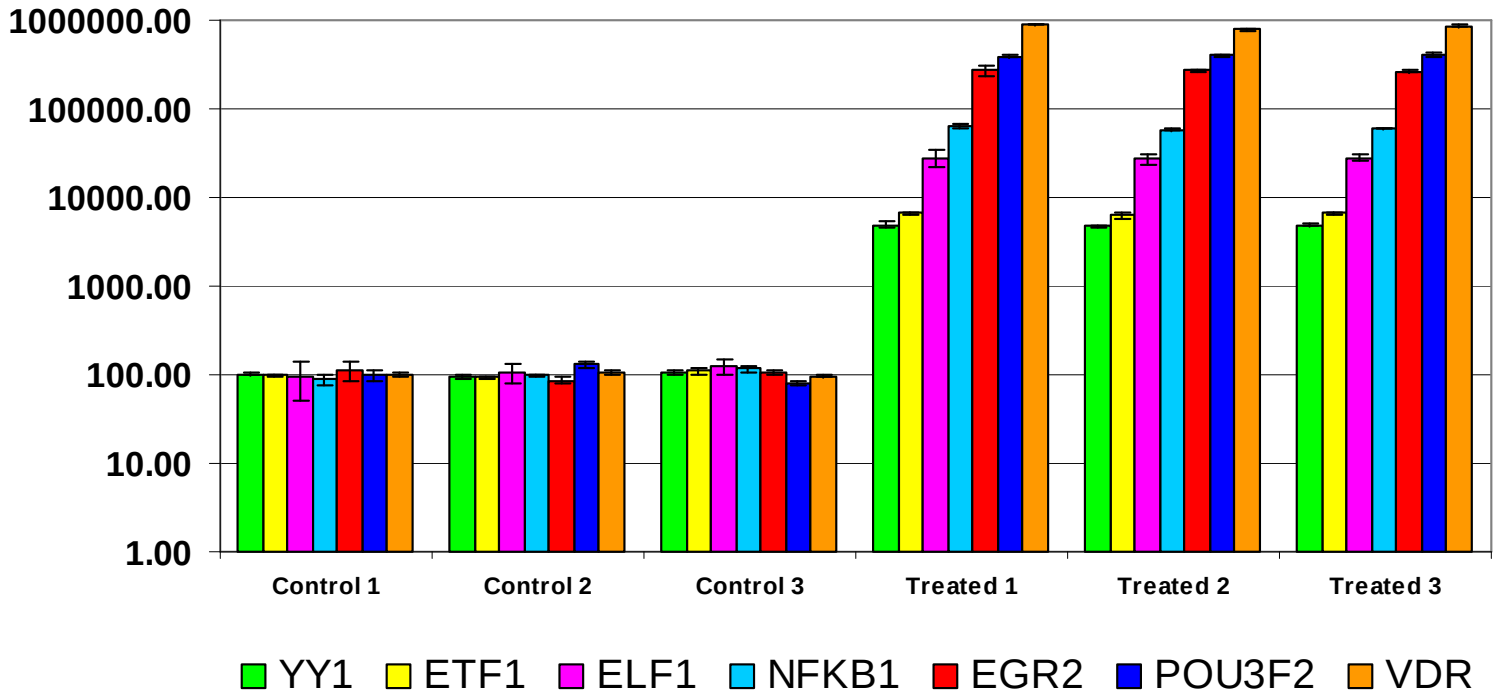

Figure 9: Overexpression Levels of 7 Transcription Factors in hESC

Table 8: Comparing Overexpression Levels of TFs in BJ Cells and hESC

T.F.	BJ O.E. Level	hESC O.E. Level	Δ O.E.
EGR2	3022X	2700x	89%
YY1	455X	49x*	11%
ELF1	150X	276x	184%
NF-KB1	140X*	605x	432%
POU3F2	38X	4013x	10561%
ETF1	15X	65x*	433%
Average Change in Overexpression in hESC Relative to BJ=			**1952%**

* TF had higher endogenous expression in that cell line

Table 9: Quantitative Overexpression Levels and 25 Downstream Targets of 6 DiRE TFs

T.F.	O.E. Level	25 Downstream Targets
EGR2	2700x	NF-KB1, PARP1, THBS1, RELA, YWHAE
YY1	49x	IL18, CYC1, PPP2R2A, OAZ1, YWHAE, ACTG2
ELF1	276x	RELA, NF-KB1
NF-KB1	605x	PSME2, CCL5, YWHAE, RELA, PPP2R2A
POU3F2	4013x	TOP1, ACTG2
ETF1	65x	YWHAE, OAZ1, NF-KB1, PPP2R2A, THSB1

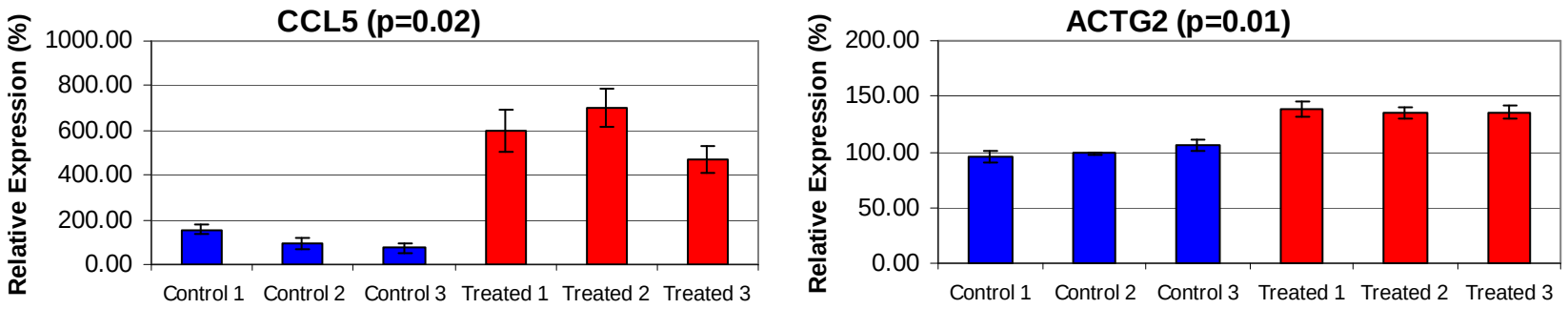

Figure 10: Gene Expression of 2 Differentially Expressed Targets Downstream of DiRE TFs in hESC

Table 10: Comparing Gene Expression of Downstream Targets for DiRE TFs in BJ Cells and hESC

DiRE Target Measured	hESC p-Value	50% Δ	BJ p-Value	50% Δ
YWHAE	0.59		0.01 ↕	
OAZ1	0.23		0.50	
PPP2R2A	0.90		0.08	
PSME2	0.74		0.20	
CCL5	0.02 ↕	√	0.01 ↕	√
PARP1	0.06		0.30	
THBS1	0.83		0.59	
IL18	0.92		0.03 ↓	
CYC1	0.58		0.29	
RELA	0.40		0.89	
TOP1	0.41		0.07	
ACTG2	0.01 ↕		0.67	

↑↓ Up/Down Δ in Expression, ↕ Consistent with Scleroderma

Table 11: Summary of Endogenous TFs and SynoR-predicted Targets

TF	O.E. Level	SynoR Target	Locus	Length	# TFBSs	p-Value	50% Δ
NF-KB1	605x	KREMEN2	chr16:2953432-2953522	91 bps	13	0.12	
		QSER1	chr11:32870262-32870323	62 bps	4	0.05	
		PRKCD	chr3:53169917-53169964	48 bps	4	0.64	
VDR	8453x	DIVA	chr3:15180666-15180843	178 bps	10	LOW	LOW
		GAA	chr17:75708463-75708921	459 bps	13	0.37	
		PCDH7	chr4:30330973-30331093	121 bps	5	0.07	√
ETF1	65x	ARID2	chr12:44409718-44409751	34 bps	10	0.02	
		PCDH1	chr5:141238160-141238196	37 bps	11	0.97	
POU3F2	4013x	NOVA2	chr19:51168551-51168581	31 bps	5	0.08	
		BCAS3	chr17:56108782-56108827	46 bps	9	0.01	

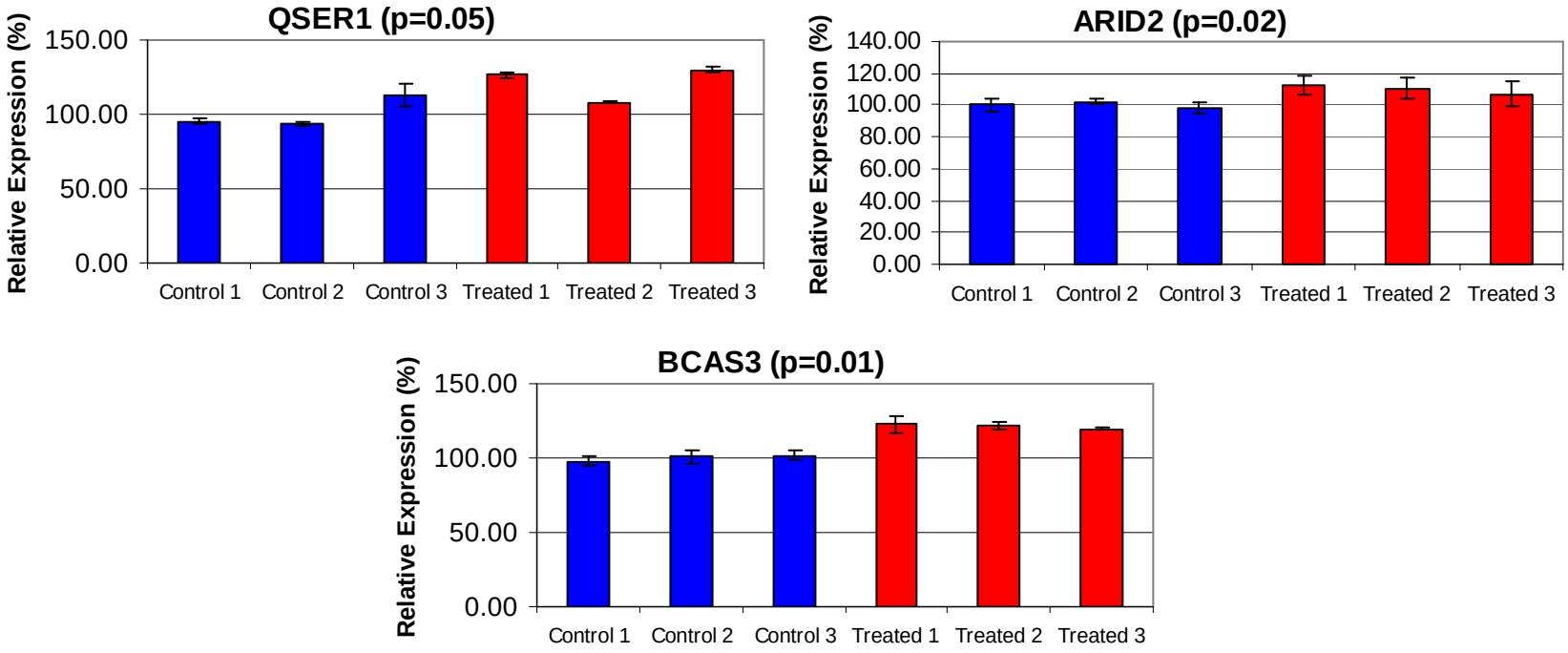

Figure 11: 3 Differentially Expressed SynoR-predicted Downstream Targets

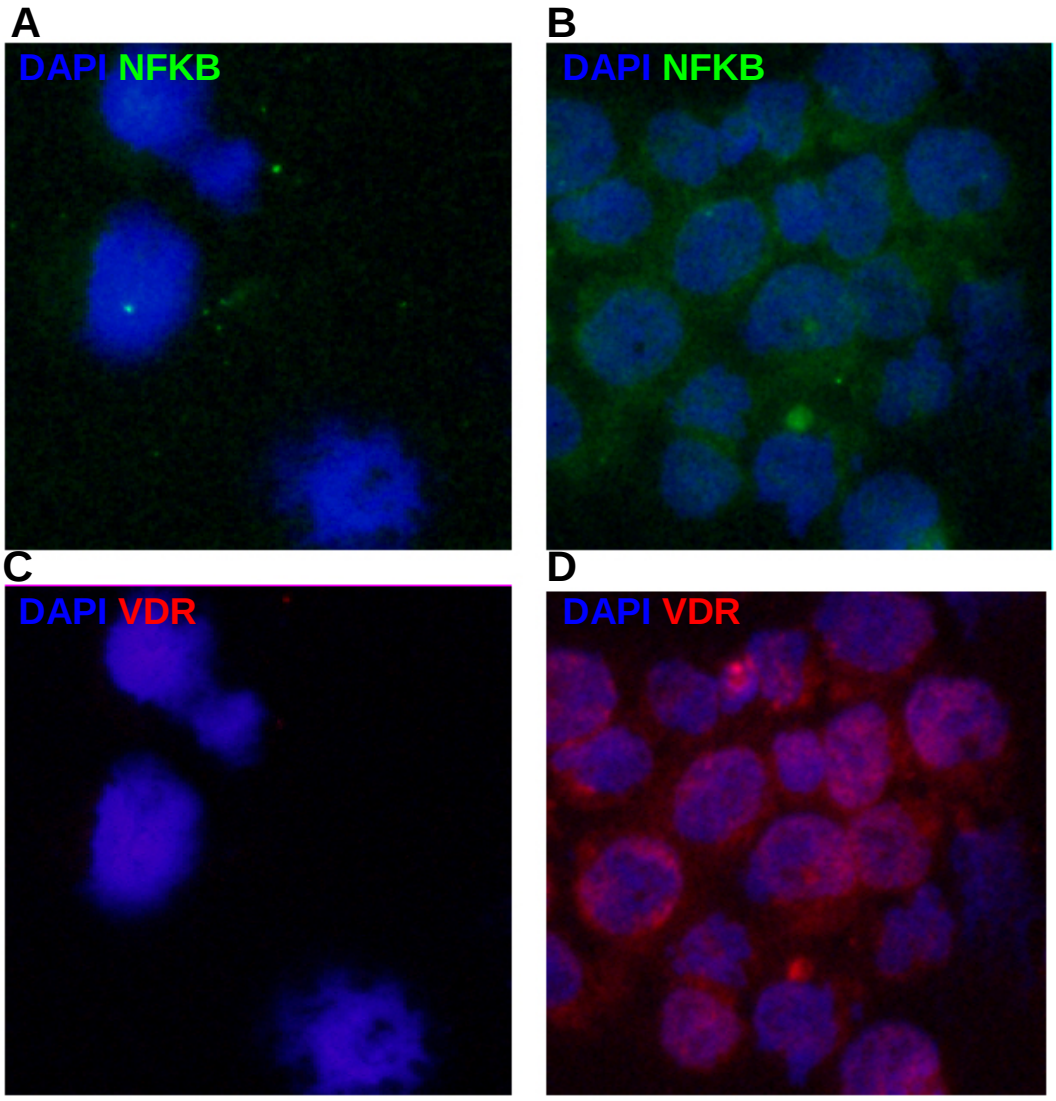

Figure 12: Immunohistochemistry of hESC Transfected with 7TFs: Staining for NFKB and VDR Transcription Factor Proteins

CHAPTER 5

Conclusion

The overall goal of this project was to induce a diseased/non-diseased gene expression pattern. Our hypothesis is that TFs may be sufficient to reprogram a diseased/non-diseased gene expression pattern and possibly a diseased phenotype. In order to do this, our aims were to: 1) Find what TFs are predicted to interact with scleroderma-specific genes, 2) Discover whether single or multiple TFs are sufficient to induce scleroderma phenotypes, and 3) Find the downstream targets of TFs that are differentially expressed in scleroderma and fibrotic cells.

DiRE-predicted TFs were able to Induce Scleroderma-specific Gene Expression

Our initial attempt to induce scleroderma-specific gene expression proved to be successful. We were successful at overexpressing our DiRE-predicted TFs in dermal fibroblasts via electroporation. By transfecting 10 individual DiRE-predicted transcription factors into dermal fibroblasts, we were able to differentially express 8 out of 36 total downstream targets. However, out of 8 of those targets, only CCL5 had an average increase in expression greater than 50%. This suggests that NF-κβ1 overexpression is able to evoke a greater change in gene expression of CCL5 relative to the other TF-gene target relationships. Differential gene expression

determined by a p-value less than or equal to 0.05 was the standard of comparison, since that was the basis of our analysis of the original microarray data. Previous studies have shown that relatively small changes in gene expression are powerful enough to evoke significant phenotypic changes.

In an effort to increase the efficiency of our ability to induce scleroderma-specific gene expression, we pooled all 10 DiRE-predicted TFs to see if we can observe combinatorial or synergistic effects. In BJ cells, 1/5 or 20% of targets were differentially expressed. This target was CCL5, which was consistent with our previous findings. From this data, we did not observe a significant difference between pooling TFs and introducing them into cells individually.

Additionally, we decided to test the 10 pooled TFs in keratinocytes to test whether or not the scleroderma gene expression pattern was cell-type specific. While fibroblasts are the major cell type of the dermis, keratinocytes occupy 95% of cells in the epidermis layer. Thus, it is plausible to assume that keratinocytes may have a function role in scleroderma disease gene expression patterns. After pooling and introducing the 10 TFs into the HaCaT (keratinocyte) cell line we did not

observe a change in CCL5 gene expression. Alternatively, changes in PPP2R2A and RELA gene expression were observed. While a statistically significant change in PPP2R2A gene expression was measured in initial dermal fibroblast single TF experiments, the same was not observed for RELA. This preliminary data suggest that multiple cell types may be responsible for the overall scleroderma gene expression pattern. In other words, in order to reprogram the scleroderma phenotype with TFs, multiple cell types may be required.

SynoR Downstream Targets were Differentially Expressed by Endogenous Upstream Regulators (TFs)

In further efforts to increase our ability to induce gene targets defined by the scleroderma phenotype, we employed a tool called SynoR, which predicts downstream regulators of endogenous TFs, or TFs we know to be differentially expressed between normal and scleroderma disease phenotypes. SynoR specifically analyzes the genes with regulatory elements that contain clusters of TFBSs. Initial analysis proved to be promising with 1 out of 3 targets differentially expressed. We continued this analysis of SynoR adding 7 targets in Chapter 4. Of the 9 targets analyzed in Chapter 4, 3 targets or 33% were differentially expressed. The only target that had an average 50% increase in gene

expression was BCAS3 when POU3F2 was overexpressed from an individual transfection (non-pooled method). The increased ratio of differentially expressed targets relative to the DiRE method (33% versus 22%) suggests that predicting targets based upon using sequence information for clustering of TFBSs as opposed to a single TFBS is more predictive.

Meta Analysis of SynoR Using PPARa Microarray Data

With the initial SynoR results we analyzed, the bioinformatic approach appeared to be predictive. To further validate SynoR, we used raw microarray data to complete a meta analysis to determine the overall accuracy of SynoR. Raw microarray expression data was collected from Finck et al. Two sets of expression data were analyzed. One set was gene expression data for a wild-type mouse and the other was data from a transgenic mouse overexpressing PPARa. First we sorted out the differentially expressed genes between these two datasets and found 2,200 such genes. Next, we conducted a SynoR analysis of PPARa and found a total of 274 TFBS clusters. The 274 genes containing PPARa clusters were then sorted out of the 2,200 differentially expressed genes and analyzed.

Of the 274 SynoR targets for PPARa**,** the differentially expressed genes are displayed in **Table 12**. The genes are sorted by the region in which the cluster is located (promoter, UTR, coding sequence, intron) and the expression ratio of PPARa overexpressing mice relative to wild-type is displayed. The majority of the differentially expressed genes are upregulated. The portion of the 274 SynoR targets that had clusters within intergenic regions was excluded, because expression data specifically for these intergenic regions was not on the chip or included in the raw data file. **Table 13** displays the results of our global validation of SynoR sorted by regions in which the clusters were located. Of the 210 SynoR targets analyzed (excluding 64 intergenic targets), 203 or 96.67% of the SynoR targets were on the chip and had gene expression data. "Hit" means that the SynoR target was differentially expressed, while "miss" means that it was not. Of the 203 SynoR targets we had gene expression data for, 27 or 13.3% of SynoR-predicted targets were in fact differentially expressed. Therefore, this global analysis of gene expression data proved SynoR to be predictable or for every 20 SynoR targets looked at, about 3 of them should prove to be differentially expressed.

Inducing Gene Expression with TFs in Human Embryonic Stem Cells

The latest set of experiments carried out in this project involved human embryonic stem cells (hESC). Our hypothesis was that hESC would be more amenable to inducing gene expression via TFs due to their relative open chromatin structure when compared to mature somatic cells. After initial analysis of TF overexpression, we observed an average increase of 1952% in TF overexpression in hESC relative to BJ fibroblasts. Several epigenetic factors such as chromatin modifications and DNA methylation may have played a role in this increase in overexpression. This increase also provides support for our hypothesis that hESC would have fewer roadblocks to reprogramming. Of the 12 downstream targets analyzed in hESC, 2 of them were differentially expressed. 3 of those same targets were differentially expressed in dermal fibroblasts. From this data, we can not say with certainty that hESC are significantly better than dermal fibroblasts for reprogramming. More data would need to be analyzed. Interestingly, only one differentially expressed target (CCL5) was in common between the two sets. It is possible that this end result is due to the difference in gene expression between the two cells types. Besides the increase in TF overexpression, no significant effect on our ability to induce

scleroderma-specific gene expression was observed in hESC in this experiment.

Experimental Conclusions

The three aims we set out to explore were: 1) Predict TFs, or upstream regulators of scleroderma disease gene targets, 2) Determine whether multiple or single TFs are required to induce scleroderma gene expression, and 3) Predict downstream gene targets of TFs differentially expressed in scleroderma. Our goal with this project was to quantify our ability to induce scleroderma gene expression. Overall, the introduction of predicted TFs into dermal fibroblasts, keratinocytes, and hES cell types were successful in inducing scleroderma expressed genes. However, the efficiency is poor with 1 out of 5 or approximately 20% of targets being differentially expressed. Also, primary sequence analysis of clusters of TFBSs (cTFBS) using SynoR is more predictive than just evolution or conservation-based approaches to predict relationships in gene expression.

More importantly, this project represents a novel approach to treating or even curing disease by reprogramming the cellular state of a disease cell. These are the first steps taken to identify the TFs

necessary to reprogram the scleroderma phenotype. The hope of stably converting a disease phenotype back to a normal phenotype is of major interest to our lab. For future research, multiple cell types should be interrogated as they showed different capacities to overexpress TFs and activate downstream targets. Additionally, increasing the length of the experiment would allow scientists to look for phenotypic changes that could possibly be caused by these changes in gene expression.

Every disease is represented by a change in phenotype. That subsequent change in phenotype is a result of a change in gene expression. Therefore, as gene expression prediction methods continue to improve, the possibility of curing disease by changing gene expression increases and becomes closer to a reality.

Methods:

Global Analysis of PPARa Microarray Expression Data

The GEO dataset GDS2289 was analyzed and wilde-type mice were compared to PPARa-overexpressing mice. Genes with differential expression with a p-value less than or equal to 0.1 were selected. This selection method resulted in 2200 such genes. These genes were then searched for in the results of our SynoR PPARa analysis of downstream gene targets. From there, were provided statistics on how many of the SynoR predicted downstream genes were differentially expressed sorted by region that the cTFBS existed in.

Table 12: List of Differentially Expressed Genes in SynoR Meta Analysis

Region	Gene	Expression Ratio
Promoter	Lyl1	1.11
Promoter	Tceal3	0.87
Promoter	Cckbr	1.26
UTR	Mll2	1.23
UTR	Phkg2	1.10
UTR	Igf2	1.15
Cds	Adcy6	1.02
Cds	C77623///Hip1r	1.07
Cds	Slc7a4	1.07
Cds	Tmcc2	0.94
Cds	Pcbp4	0.92
Cds	Ces3	1.32
Cds	Ces3	1.31
Cds	D10Ertd610e	1.04
Cds	Chuk	1.13
Cds	Pclo	1.06
Intron	Sgpl1	1.07
Intron	Slc12a7	1.09
Intron	Bcl9l	1.04
Intron	Ttyh2	1.45
Intron	Tm9sf2	0.89
Intron	Ptprs	0.91
Intron	Sh3gl3	0.85
Intron	Ebf2	1.12
Intron	Wnt4	1.12
Intron	Dock9	1.07
Intron	3110079O15Rik	1.02

Table 13: Summary of SynoR Meta Analysis by Region

	Promoter	UTR	Cds	Intron	Total Targets
Hits	3	3	10	11	27
Misses	12	17	98	49	176
Total	15	20	108	60	203
Hit %	20.00%	15.00%	9.26%	18.33%	**13.30%**
On Chip	15	20	108	60	203
Not On Chip	1	1	3	2	7
% On Chip	93.75%	95.24%	97.30%	96.77%	**96.67%**
Total	16	21	111	62	**210**

REFERENCES

Bibikova, Marina, Louise C. Lourent, Bing Ren, Jeanne F. Loring, and Jian-Bing Fen. "Unraveling Epigenetic Regulation in Embryonic Stem Cells." *Cell Stem Cell* 2.February (2008): 123-34.

Bioconductor: open software development for computational biology and Bioinformatics. Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang and Jianhua Zhang. Genome Biology 2004, 5:R80

Chang, H.Y. and Cotsarelis, G. Turning skin into embryonic stem cells. Nat. Med. 13(7):783-4 (2007).

Choi, Kyung-Dal, Junying Yu, Kim Smuga-Otto, Giorgia Salvagiotto, Willaim Rehrauer, Maxim Vodyanik, James Thomson, and Igor Slukvin. "Hematopoietic and Endothelial Differentiation of Human Induced Pluripotent Stem Cells." *Stem Cells* 27.3 (2009): 559-67.

Finck BN, Bernal-Mizrachi C, Han DH, Coleman T, Sambandam N, LaRiviere LL, Holloszy JO, Semenkovich CF, and Kelly DP. A potential link between muscle peroxisome proliferator- activated receptor-alpha signaling and obesity-related diabetes. Cell Metab 2005 Feb;1(2):133-44. PMID: 16054054

Gabrielli, Armando, Enrico V. Avvedimento, and Thomas Krieg. "Scleroderma." *The New England Journal of Medicine* 360.19 (2009): 1989-2003.

G. Loots and I. Ovcharenko, rVista 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Research, 32 (Web Server Issue), W217-W221 (2004).

Gotea, Valer, Axel Visel, and John M. Westlund. "Homotypic Clusters of Transcription Factor Binding Sites Are a Key Component of

Human Promoters and Enhancers." *Genome Research* 20 (2010): 565-77.

Gotea V. and Ovcharenko I. DiRE: identifying distant regulatory elements of co-expressed genes. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W133-9.

Jaenische, Rudoplh, and Adrian Bird. "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals." *Nature Genetics* 33.March (2003): 245-54.

Kirk T.Z., Mark M.E., Chua C.C., Chua B.H., and Mayes M.D. Myofibroblasts from Scleroderma Skin Synthesize Elevated Levels of Collagen and Tissue Inhibitor of Metalloproteinase (TIMP-1) with Two Forms of TIMP-1. J. of Biological Chemistry. 1995 Feb 17; 270(7): 3423-28.

Knowledge and Information Systems, Vol. 5, No. 4. (1 November 2003), pp. 416-438.

Latchman, David S. "Transcription Factors: An Overview." *Int. J. Biochem. Crll Biol.* 29 (1997): 1305-312.

Lee, Gabsang, Eirini P. Papapetrou, Hyesoo Kim, Stuart M. Chambers, Mark J. Tomishima, Christopher A. Fasano, Yosif M. Ganat, Jayanthi Menon, Fumiko Shimizu, Agnes Viale, Viviane Tabar, Michel Sadelain, and Lorenz Studar. "Modeling Pathogenesis and Treatment of Familial Dysautonomia Using Patient Specific IPSCs." *Nature* 461.7262 (2009): 402-06.

Li, En. "CHROMATIN MODIFICATION AND EPIGENETIC REPROGRAMMING IN MAMMALIAN DEVELOPMENT." *Nature* 3.September (2002): 662-73.

Lorena D., Uchio K., Costa A.M.A., and Desmouliere A. Wound Repair & Regeneration. 10(2):86-92, March/April 2002.

Machesney M, Tidman N, Waseem A, Kirby L, Leigh I (1998) Activated keratinocytes in the epidermis of hypertrophic scars. Am J Pathol152: 133–1141.

McGrath JA, Eady RAJ, Pope FM. (2004). "Anatomy and Organization of Human Skin". In Burns T, Breathnach S, Cox N, Griffiths C.. *Rook's Textbook of Dermatology* (7th ed.). Blackwell Publishing. pp. 4190. doi:10.1002/9780470750520.ch3. ISBN 9780632064298.

Ovcharenko, Ivan, and Marcelo A. Nobrega. "Identifying Synonymous Regulatory Elements in Vertebrate Genomes." *Nucleic Acids Research* 33 (2005): 403-07.

Pennacchio, Len A., Gabriela G. Loots, and Marcelo A. Nobrega. "Predicting Tissue-specific Enhancers in the Human Genome." *Genome Res.* 17 (2007): 201-11.

Ptashne, Mark, and Alexander Gann. "Transcriptional Activation by Recruitment." *Nature* 386.April 10 (1997): 569-77.

Reik, Wolf, Wendy Dean, and Jorn Walter. "Epigenetic Reprogramming in Mammalian Development." *Science* 293.August (2001): 1089-093.

"Scleroderma Foundation - Medical Overview of Scleroderma. What is it?" Scleroderma Foundation - Home Page. 31 May 2009 <http://www.scleroderma.org/medical/overview.shtm#myths>.

Takahashi K., Tanabe K., Ohnuki M., Narita M., Ichisaka T., Tomoda K., and Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007 Nov 30;131(5):861-72.

Takahashi K., Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 Aug 25;126(4):663-76.

Uitto J., Bauer E.A., and Eisen A.Z. Increased Biosynthesis of Triple-helical Type I and Type III Procollagens Associated with Unaltered Expression of Collagenase by Skin Fibroblasts in Culture. J Clin Invest. 1979 October; 64(4): 921-930.

Wang, Lisheng, Li Li, Farbod Shojaei, Krysta Levac, Chantal Cerdan, Pablo Menendez, Tanya Martin, Anne Rouleau, and Mickie Bhatia. "Endothelial and Hematopoietic Cell Fate of Human

Embryonic Stem Cells Originates from Primitive Endothelium with Hemangioblastic Properties." *Immunity* 21.1 (2004): 31-41.

Werner S. and Grose R. Regulation of Wound Healing by Growth Factors and Cytokines. Phys Rev. 2003 82:835-70.

Wernig M., Meissner A., Foreman R., Brambrink T., Ku M., Hochedlinger K., Bernstein B.E., Jaenisch R. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature. 2007 Jul 19;448(7151):318-24.

Whitfield M.L., Finlay D.R., Murray J.I., Troyanskaya O.G., Chi J.T., Pergamenschikov A., McCalmont T.H., Brown P.O., Botstein D., Connolly M.K. Systemic and cell type-specific gene expression patterns in scleroderma skin. Proc Natl Acad Sci U S A 2003 Oct 14; 100(21):12319-24.

Yu J., Vodyanik M.A., Smuga-Otto K., Antosiewicz-Bourget J., Frane J.L., Tian S., Nie J., Jonsdottir G.A., Ruotti V., Stewart R., Slukvin I.I., Thomson J.A. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007 Dec 21;318(5858):1917-20.

Zhang, Donghui, Wei Jiang, Meng Liu, Xin Sui, Xiaolei Yin, Song Chen, Yan Shi, and Hongkui Deng. "Highly Efficient Differentiation of Human ES Cells and IPS Cells into Mature Pancreatic Insulin-producing Cells." *Cell Research* 19 (2009): 429-38.