A systems biology approach to genetic studies of complex diseases

Momiao Xiong[a], Carol A. Feghali-Bostwick[c], Frank C Arnett[b], Xiaodong Zhou[b,*]

[a] *Human Genetics Center, University of Texas School of Public Health, Houston, TX 77030, USA*
[b] *Division of Rheumatology and Clinical Immunogenetics, Department of Internal Medicine, University of Texas Medical School at Houston, 6431 Fannin, MSB5.270, Houston, TX 77030, USA*
[c] *Dorothy P. and Richard P. Simmons Center for Interstitial Lung Disease, Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA*

Received 20 May 2005; revised 27 July 2005; accepted 18 August 2005

Available online 13 September 2005

Edited by Robert B. Russell

Abstract **Revealing mechanisms underlying complex diseases poses great challenges to biologists. The traditional linkage and linkage disequilibrium analysis that have been successful in the identification of genes responsible for Mendelian traits, however, have not led to similar success in discovering genes influencing the development of complex diseases. Emerging functional genomic and proteomic ('omic') resources and technologies provide great opportunities to develop new methods for systematic identification of genes underlying complex diseases. In this report, we propose a systems biology approach, which integrates omic data, to find genes responsible for complex diseases. This approach consists of five steps: (1) generate a set of candidate genes using gene–gene interaction data sets; (2) reconstruct a genetic network with the set of candidate genes from gene expression data; (3) identify differentially regulated genes between normal and abnormal samples in the network; (4) validate regulatory relationship between the genes in the network by perturbing the network using RNAi and monitoring the response using RT-PCR; and (5) genotype the differentially regulated genes and test their association with the diseases by direct association studies. To prove the concept in principle, the proposed approach is applied to genetic studies of the autoimmune disease scleroderma or systemic sclerosis.**

Keywords: Systems biology; Genetic networks, omics; Regulation

1. Introduction

The traditional paradigm for genetic studies of diseases is to connect DNA variation with phenotypic variation [1]. The major tools for identification of genes influencing traits are linkage analysis and association studies [2]. Although linkage analysis and association studies are successful in localizing genes responsible for Mendelian diseases, their applications to identification of genes causing complex diseases have not led to similar success [3]. The discovery of causative genes for complex diseases poses great challenges to biologists because disease develops as a consequence of interactions between multiple DNA variants and exposures to environmental agents varying over time and space, which are organized into networks [4].

Emerging genomic, transcriptomic, proteomic, and metabolomic ('omic') resources and technologies are revolutionizing biomedical research and allow a transition from the traditional paradigm for genetic studies of complex diseases to a new paradigm based on systems biology. The systems biology approach to genetic studies of complex diseases has several remarkable features. First, the systems biology approach assumes that the majority of genes function through complex networks. Biological networks abstractly represent biological systems and capture their systemic properties [5]. In the systems biology approach, complex traits will not only be dissected by individually studying some components of the networks, but also will be investigated by studying networks as a whole. Second, the data from any single approach may contain incomplete information due to the occurrence of false negatives and false positives [6]. The data from distinct omic sources may be complementary to each other. Therefore, the systems biology approach integrating DNA variation, gene expression, protein–protein interaction and phenotypic variation will increase the reliability of discovering causative genes for complex diseases.

One of the key issues in developing a systems biology approach to genetic studies of complex diseases is how to efficiently integrate various omic data sets and to maximally extract disease-relevant information. To address this issue, we propose the following schemes for applying a systems biology approach to unraveling the genetic mechanisms of complex diseases.

First, we can select and model candidate genetic networks. Widely used genomic and transcriptomic methods for complex disease studies have focused on candidate gene approaches. Most genetic association studies have been conducted as candidate polymorphism or gene studies [7]. A commonly used method for using gene expression in dissecting the molecular basis of the disease is to identify differentially expressed genes [8]. However, the status of the cell and cellular processes is largely determined by a number of genes interwoven into networks, rather than a few genes [9]. An alternative method to a candidate gene approach is a candidate genetic network strategy. Candidate genetic network approaches will be useful not only for identifying disease genes, but also for elucidating pathogenesis and discovering treatments for diseases. Selection of candidate genetic networks can be accomplished by (1) choosing a set of candidate genes from either linkage analysis or gene–gene interaction data sets, or gene expression analyses,

*Corresponding author.
E-mail address: Xiaodong.Zhou@uth.tmc.edu (X. Zhou).

0014-5793/$30.00
doi:10.1016/j.febslet.2005.08.058

and then reconstructing genetic networks from expression profiles of a selected set of candidate genes or (2) searching literature and network databases. After the candidate genetic networks are selected, we then model quantitatively genetic networks. We propose to use structural equations [10] as a framework for genetic network modeling.

Second, we can identify differentially regulated genetic networks. Differential expression of genes is a widely used concept for identifying genes that are able to discriminate cell phenotypes. However, the level of gene expression does not reflect genetic changes. Causes of differential expression are multiple. Differential expression of genes between normal and abnormal tissues may be due either to mutation of its own gene or the effects of expression changes of other genes in the network. We often observe that the degree of differential expression of one gene due to its own mutations is lower than that of the downstream genes in the network whose expression levels are affected by expressions of upstream genes in the network. Functional mutations in the gene often cause changes in regulation.

The functional mutation of the gene may have more impact on its regulation than on its expression. We expect that due to the accumulation of mutations, the regulation of some genetic networks in abnormal cells will be significantly different from that in normal cells. Uncovering such differences may help to identify the cause of the diseases. Coefficient parameters in the structural equations measure the regulatory effects of one gene on others or the strength of the gene–gene interaction, and form a matrix that is referred to as a regulatory matrix. Identification of differentially regulated genetic networks can be accomplished by measuring differences in a norm of the regulatory matrix between normal and abnormal samples. By identifying differentially regulated genetic networks, we are likely to discover a set of genes and genetic networks that influence the development of diseases.

Third, we validate the regulatory relationship between genes in the network by perturbing the network using RNA *silencing* (*RNAi*) or antisense RNA and measuring the response using RT-PCR. Due to biological and experimental variation, the results of differentially regulated genetic networks from gene expression analysis may not be reliable and need to be confirmed. RNAi coupled with RT-PCR is a powerful tool for changing gene regulation and can be used to examine the accuracy of predictions of regulatory relations between genes in genetic network modeling.

Fourth, we previously assumed that changes in regulatory effects may come from mutations. To test this hypothesis, we can genotype the differentially regulated genes and test for associations of the mutations with the disease as well as with regulatory effect changes.

To prove the principle of concept, the proposed scheme for systems biology approach to complex diseases was applied to genetic studies of scleroderma or systemic sclerosis (SSc).

2. Selection of a set of candidate genes

For ease of presentation, we use our studies of SSc as an example to describe a scheme of a systems biology approach to genetic studies of a complex disease. SSc is a multi-system disease of unknown etiology characterized by cutaneous and visceral fibrosis, microvascular damage and autoimmune phenomena [11]. Although candidate genes are not obvious, there are still multiple ways to select them. The candidate genes can be selected by (1) literature review, (2) linkage and association studies and (3) gene or protein expression data analysis. Here, we select candidate genes by identifying differentially expressed genes from microarray data analysis of skin fibroblasts.

Great biological variability exists within each individual. Causes of differential expressions of genes between normal individuals and patients with a disease can be due to either genetic differences or differences in environmental exposures. To reduce the impact of genetics on the differential expressions of the genes, we conducted twin studies that used 16 pairs of affected and unaffected SSc twins. Among these twins, 11 pairs were monozygotic and 5 were dizygotic. Each pair represents an SSc patient and a normal individual. Fibroblast strains were cultured from skin biopsies of lesional and non-lesional skin of affected twins and normal skin of unaffected twins (total 48 samples). Oligo microarrays containing 16 650 human genes were used in gene expression profiling of cultured fibroblasts of these twins [12].

We postulated that if the expression levels of the genes between monozygotic SSc patients and their normal twins showed no significant differences, but expression levels between dizygotic SSc patients and their normal pairs, or between monozygotic SSc patients and dizygotic normals showed significant differences, then the difference in gene expressions between SSc patients and normal individuals is more likely due to genetic differences. There are four ways to compare the differential expressions of the genes between dizygotic SSc patients and their normal twin pairs or between monozygotic SSc patients and the dizygotic normal twins: (1) comparison between lesional skin of dizygotic SSc patients and that of their paired normal twins; (2) comparison between non-lesional skin of dizygotic SSc patients and their paired twins; (3) comparison between lesional skin of monozygotic SSc patients and dizygotic normals; and (4) comparison between non-lesional skin of monozygotic SSc patients and dizygotic normals. Table 1 lists genes whose *P*-values showed significantly differential expression equal or less than 0.05 in at least three of the above comparisons.

Collagens are important components of the extracellular matrix (ECM) and connective tissue growth factor (CTGF) is a cysteine-rich secreted protein. Earlier studies have shown that transforming growth factor (TGF)-β induces CTGF expression and that TGF-β pathways, including CTGF as a downstream mediator, can induce collagen production [13]. The TGF-β pathways regulate multiple biological processes, including inflammation, skeletal development, wound repair, differentiation and apoptosis [14].

Because of the complexity of TGF-β pathways, the complete structure of the network has not been elucidated. Table 2 lists the genes that are directly or indirectly involved in TGF-β transduction pathways and which show differential expression in five comparisons of our SSc twin studies. Collagen types I, III and XI, SPARC, MAD3, CTGF and CREB demonstrated significantly differential expression in at least one comparison between SSc patients and normal controls, but showed no significant differential expressions between lesional skin of monozygotic patients and their paired normal twins. This implies that differential expression of these genes may be due to genetic differences in one or more of them. Below we will study how to use a structural equation model as a simplified representation

Table 1
Genes showing significantly differential expression in at least three comparisons

Gene name	Monozygotic lesion/normal pair		Dizygotic lesion/ normal pair		Monozygotic lesion/ dizygotic normal		Dizygotic non-lesion/ normal pair		Monozygotic non-lesion/ dizygotic normal	
	P-value	Fold	*P*-value	Fold	*P*-value	Fold	*P*-value	Fold	*P*-value	Fold
COL XIA1	0.3667	0.7032	0.0404	4.3987	0.018	8.4834	0.0157	21.899	0.0148	22.369
OCRL	0.1698	1.3261	0.0371	3.4459	0.0462	2.4577	0.0076	4.8684	0.0006	5.6031
PNMT	0.106	1.0773	0.0113	3.8834	0.0131	2.53	0.0178	2.0229	0.0609	2.1629
CTGF	0.1908	1.0901	0.0311	3.7412	0.0619	2.9199	0.0085	6.329	0.0309	6.0569
PRKAA2	0.4613	0.8258	0.011	3.4701	0.0837	2.124	0.0321	3.8	0.0271	6.2326
CPR8	0.2456	0.6822	0.0468	3.393	0.1686	1.6322	0.0276	2.531	0.0387	2.2136

*Gene names: COL, collagen; CRL, oculocerebrorenal syndrome of Lowe; PNMT, phenylethanolamine *N*-methyltransferase; CTGF, connective tissue growth factor; PRKAA2, protein kinase, AMP-activated, alpha 2 catalytic subunit, CPR8, cell cycle progression 8 protein.

Table 2
P-values of 10 genes in TGF-β pathways showing significantly differential expressions

Gene name	Dizygotic lesion/ normal pair		Monozygotic lesion/ dizygotic normal		Dizygotic non-lesion/dizygotic normal		Monozygotic non-lesion/ dizygotic normal		Monozygotic lesion/ normal pair	
	P-value	Fold	*P*-value	Fold	*P*-value	Fold	*P*-value	Fold	*P*-value	Fold
COL XIA1	0.0404	4.3987	0.018	8.4834	0.0157	21.899	0.0148	22.3686	0.3667	0.7032
SPARC	0.2002	1.6255	0.0501	1.5577	0.0008	3.9864	0.0097	4.2166	0.4531	0.7764
TGFB1	0.1445	1.4403	0.473	0.8009	0.08	1.6504	0.2289	1.1788	0.2889	0.6138
TGFB2	0.4069	0.8584	0.2541	1.3879	0.3167	0.9645	0.1235	0.8419	0.0130	1.9115
MAD3	0.3762	0.897	0.4454	1.1587	0.032	0.5716	0.1359	0.7637	0.4955	0.9746
CTGF	0.031	3.7412	0.0619	2.9199	0.0085	6.3289	0.0309	6.0569	0.1908	1.0902
CREB	0.4672	0.9837	0.0244	0.6021	0.1938	0.7474	0.1083	0.7344	0.442	1.1187
Plasminogen	0.1093	1.1769	0.2963	0.9447	0.0739	1.3732	0.0655	1.8568	0.4981	0.8668
COL 1A2	0.0143	0.5627	0.0408	0.7861	0.14	0.8912	0.4224	1.2339	0.4588	0.7034
COL 3A1	0.3083	2.8334	0.3872	0.8992	0.0145	3.8956	0.2015	2.9321	0.1194	0.3658

of TGF-β pathways and to estimate the strength of regulatory interactions between these ten genes.

3. Reconstruction and modeling of genetic networks

We start with modeling of genetic networks based on some known networks. A genetic network can be represented by a path diagram. The path diagram consists of nodes represented by letters, and edges represented by lines. The nodes of the path diagram correspond to variables. The directed edges between nodes denote the direction of the regulatory relationship between the nodes (variables) connected by the edges and indicate a directed regulatory influence of one gene on another. The directed edges can represent either activation (positive control) or inhibition (negative control).

Variables in path diagrams can be classified into two basic types: observed variables that can be measured and the residual error variables that cannot be measured and represent all other un-modeled causes of the variables. Most observed variables (e.g., gene expression levels) are random. Some observed variables might be non-random or control variables (e.g., drug doses) whose values remain the same in repeated random sampling or might be manipulated by the experimenter. The observed variables will be further classified into exogenous variables, which lie outside the model, and endogenous variables, whose values are determined through joint interaction with other variables within the system. All non-random variables and some of the gene (or protein) expression data (e.g., initiators of pathway) can be viewed as exogenous variables. Most of the gene (or protein) expression data are viewed as endogenous variables. The terms exogenous and endogenous are model specific. It may be that an exogenous variable in one model is endogenous in another. The observed variables are enclosed in boxes and the error variables are not enclosed at all.

Linear structural equations can be used to model quantitatively genetic networks. Let Y be a vector of the p endogenous variables and X be a vector of q exogenous variables. Occasionally, one or more of the X's are non-random. We denote the errors by e. We assume that $E[e] = 0$ and that e is uncorrelated with the exogenous variables in X. We also assume that e_i is homoscedastic and non-autocorrelated [10]. Then, gene expressions in the genetic network are modeled by the following linear structural equations:

$$Y = BY + \Gamma X + e, \tag{1}$$

where B is a $p \times p$ matrix and Γ is a $p \times q$ matrix. The elements of the coefficient matrices B and Γ describe the regulatory (or causal) effects of one gene on the other, or a non-random variable on the gene, which are a direct regulatory influence of one variable on the other. Therefore, throughout the paper, the matrices B and Γ are referred to as the *regulatory matrices*. Since the genetic networks are not fully connected, many elements in the matrices B and Γ will be zero. The matrices B and Γ are, in general, sparse. The matrix B can describe feedback relations in the path diagram. The structural equations can model directed cyclic graphs and hence genetic networks with feedback loops [10].

The basic hypothesis of the general structural equation is

$$\Sigma = \Sigma(\theta),$$

where Σ is the population covariance matrix and $\Sigma(\theta)$ is the covariance matrix. The above equation implies that each element of the covariance matrix is a function of model parameters. To ensure that parameter estimators are consistent and unbiased, the estimation procedures derive from the relation of the covariance matrix of the observed variables to the structural parameters. The unknown parameters are estimated so that the implied covariance matrix Σ is close to the sample covariance matrix S. To know when our estimates are as "close" as possible, we must define "close", that is, we require a function that is to be minimized. The most widely used fitting function is based on the method of maximum likelihood (ML) defined by maximizing the likelihood function or its log:

$$F_{\mathrm{ML}} = \log|\Sigma(\theta)| + \mathrm{Tr}(S\Sigma^{-1}(\theta)) - \log|S| - (p+q),$$

where p and q are the number of endogenous and exogenous variables, and Tr denotes the trace of a matrix. The fitting function F_{ML} compares the difference between the observed and predicted covariance matrices. In general, F_{ML} is a complicated non-linear function of the structural parameters, and explicit solutions are not always found. Instead, a Newton unconstrained optimization procedure is employed to find solutions [11].

A linear structural equation model was applied to analyzing the expression profiles of ten genes in these twin studies. The tissue samples include lesional and non-lesional skin of 11 monozygotic and 5 dizygotic patients, and unaffected skin of their 16 normal pairs. Since we assume that the structure of the networks with ten candidate genes are unknown, we repeatedly applied genetic algorithms to the data set 200 times. The path diagram of the network with the largest fitting probability $P = 0.8864$ is shown in Fig. 1. The structural equations for the network were given by

$$\mathrm{SMAD} = -0.2242\mathrm{TGF}-\beta1 - 0.1242\mathrm{TGF}-\beta2,$$
$$\mathrm{CTGF} = -0.6359\mathrm{SMAD} - 0.4031\mathrm{CREB} + 1.3750\mathrm{SPARC},$$
$$\text{Collagen I} = -0.1720\mathrm{CREB} - 0.2679\mathrm{SPARC},$$
$$\text{Collagen III} = 0.5916\mathrm{CTGF} - 0.6155\text{Collagen I},$$
$$\text{Collagen XI} = 0.9124\mathrm{CTGF},$$
$$\textit{Plas}\ \min\ \textit{ogen} = 0.1331\mathrm{SPARC} - 0.1581\mathrm{TGF}-\beta2.$$

The coefficients in the equations measure the magnitude of influence of one gene on the expression of another gene and hence are referred to as the regulatory effects of the genes. The positive and negative regulatory effects of the gene indicate activation and inhibition, respectively.

The proposed algorithm correctly identified the structure of the network. The regulatory relations between the genes in the reconstructed network can be confirmed by the experiments. Numerous studies have shown that TGF-β families initiate activation and the transduction of MAD3 proteins by binding to TGF-β receptors type I and type II [15]. Since the regulatory effects of TGF-β1 and TGF-β2 on MAD3 and the regulatory effect of MAD3 on CTGF were negative, TGF-β1 and TGF-β2 inhibit MAD3, which in turn increased expression of CTGF. This was supported by the report that TGFβ increases expression of CTGF markedly in human fibroblasts [16]. CTGF was reported as a downstream mediator of bioactivities of TGF-β [17]. CTGF also is reported to enhance expression of collagen [18]. SPARC is shown to regulate the expression of collagen type I in mesangal cells [19]. Previous data also confirm that CREB blocks expression of CTGF and collagen type I [20].

4. Differentially regulated genetic networks

Differential regulation is a useful concept of genetic networks for identifying mutations causing diseases. Before we investigate how to use differentially regulated genetic networks

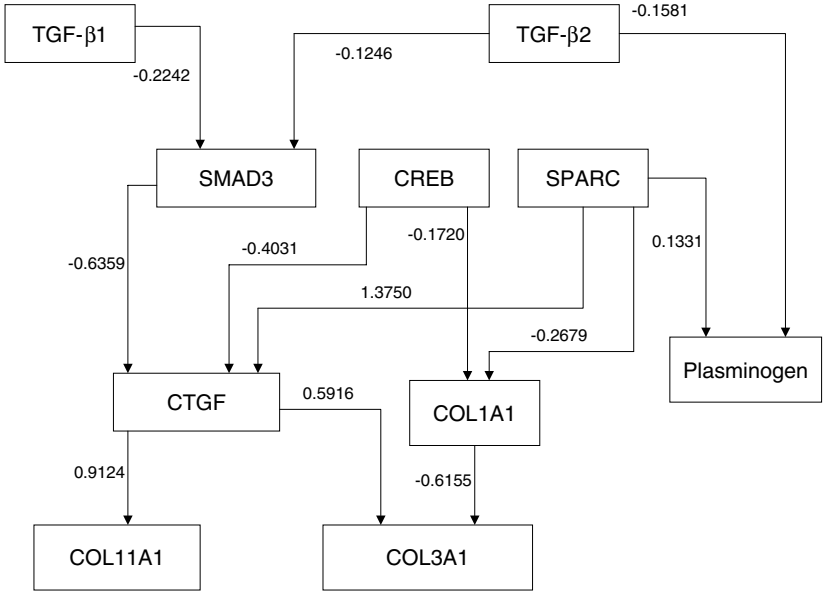

Fig. 1. The scheme of path diagram for TGF-β pathways with 10 genes reconstructed from gene expression data in SSc studies of total 16 abnormal tissue samples and 32 normal tissue samples. The number along the edges was the estimated regulatory effect of one gene on another.

to reveal pathogenesis of the disease, we study how to test differential regulation of the gene between normal and abnormal samples. Let $A = [B\ \Gamma]$ be a coefficient matrix of structural equations for modeling a genetic network. Let A_1 and A_2 be its corresponding coefficient matrices in the normal and abnormal tissue samples. Let $W = A_1 - A_2$ and w_{ij} be an element of the matrix W. Since w_{ij} is a parameter in the network, its asymptotic standard deviation can be calculated from the square root of the main diagonal of the asymptotic covariance matrix of the estimated parameters in the network and denoted by $S_{W_{ij}}$. We define the test statistic T_G as follows:

$$T_G = \frac{W_{ij}}{S_{W_{ij}}}.$$

Although the exact distribution of T_G is unknown, its asymptotical distribution can be approximated by a t distribution with $N - 2$ degrees of freedom. This statistic can be used to test the difference of the regulatory effect of one gene on another between normal and abnormal tissues.

The difference of the regulatory effect of one gene on another cannot measure the difference in the global behavior of the genetic networks between normal and abnormal tissues. A simple quantity to measure the difference in global behavior of genetic networks between the normal and abnormal tissues is the largest absolute value of the difference of the regulatory effect of one gene on another in the network between the normal and abnormal tissues, i.e., $w_0 = \max_{i,j} \mid w_{ij} \mid = \mid w_{i_0 j_0} \mid$. The statistic T_G for testing the difference of individual regulatory effects can be used to test the difference in global behavior of genetic networks. Specifically, the statistic for testing the differential regulation of the genetic networks is given by

$$T_{G_0} = \frac{w_{i_0 j_0}}{S_{w_{i_0 j_0}}}.$$

The P-value is calculated by a permutation test. The gene expression profile matrix is randomly permuted, and the structural equation model and genetic algorithms are applied to randomly permutated gene expression data to reconstruct the genetic network hundreds or thousands of times. Then, we calculate T_{G_0} and obtain an empirical distribution of T_{G_0}. The P-value of the test is then defined as the probability that T_{G_0} exceeds its observed value. The statistic T_{G_0} can be used to measure the difference in regulation of the genetic network.

Identification of differentially regulated genetic networks consists of three steps. First, we reconstruct genetic networks using structural equations and gene expression data in all available samples. Second, we fix the structure of the genetic networks and then estimate network parameters by using gene expression data of normal and abnormal samples. Third, we rank the genetic networks according to some statistics, which measure the extent of the difference in regulatory effects of the genetic networks between normal and abnormal tissue samples.

In twin studies, for the reconstructed genetic network shown in Fig. 1, we estimated regulatory effects of the genes in the network by using gene expression data of abnormal samples consisting of lesional skin of 11 monozygotic SSc patients and 5 dizygotic SSc patients and putative normal samples consisting of non-lesional skin of 11 monozygotic and 5 dizygotic SSc patients and skin of 11 normal monozygotic and 5 dizygotic pairs, respectively. The results are shown in Fig. 2. The number in parenthesis along the edges was the regulatory effect of the genes in the normal tissue samples. The largest difference of the regulatory effects of the genes was 1.3553 (T_{G_0} = 12.4387, P-value = 1.1102×10^{-16}), which was associated with the regulation of CREB on CTGF, where the P-value was obtained by a permutation test. In the normal tissues, the function of CREB is to inhibit the expression of Collagen type I and CTGF, which in turn regulates the expression of collagen III. As Fig. 2 shows, in the abnormal tissues, CREB increased the expression of collagen type I and CTGF, which in turn increased the expression of collagen type III. From Fig. 2, we

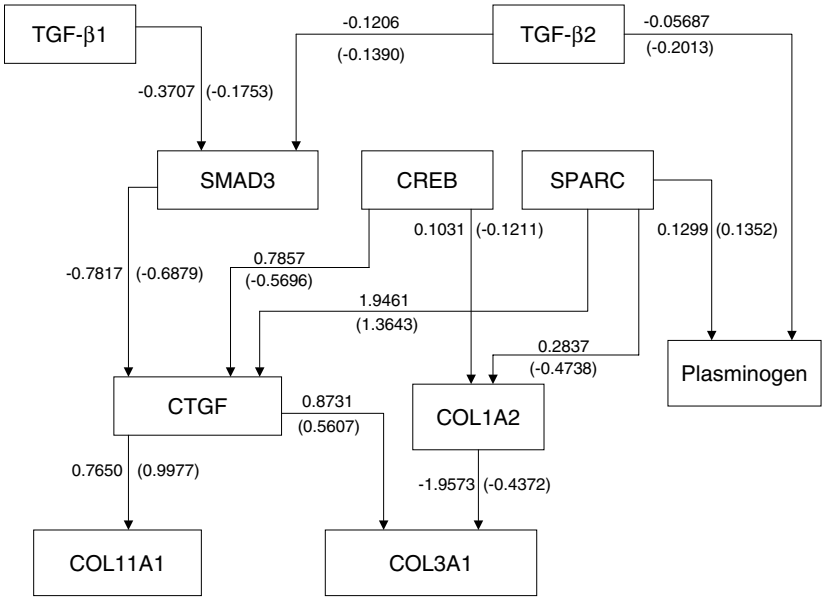

Fig. 2. The scheme of path diagram for TGF-β pathways with 10 genes reconstructed from gene expression data of total 32 normal tissue samples and 16 abnormal tissue samples. The number along the edges was the estimated regulatory effect of one gene on another from abnormal tissue samples. The number in parenthesis along the edges was the estimated regulatory effect of one gene on another from normal tissue samples.

can also see that SPARC changed regulatory roles from negative control to positive control when the normal skin tissues became lesional. In the normal tissues, SPARC inhibited expression of collagen type I, but increased production of collagen type I.

The above observations can be well supported by previous experiments and explained by the pathogenesis of SSc. SSc is a chronic multi-system disease. It is believed that widespread tissue fibrosis is due to expansion of fibrogenic clones of tissue fibroblasts, which produce excessive amounts of ECM components, such as collagens and fibronectin [11]. Growth factors and matricellular proteins are believed to play major roles in the maintenance of the homeostasis of the ECM. In SSc tissues, increased activities of growth factors or cytokines, such as TGF-β and CTGF, are well documented [21]. TGF β signaling in SSc tissue is believed to play important role in fibrotic process [21]. The CTGF is a downstream gene in TGF-β signaling [17]. Transfection of the CTGF gene into normal fibroblasts induced an autocrine fibrotic phenotype including over-production of collagens [22]. SPARC is a matricellular protein and an important regulator of cell–matrix interaction. Our previous studies have demonstrated an over-expression of SPARC gene in SSc fibroblasts and an increased level of SPARC protein in both cellular lysates and culture media in SSc [23]. It is also reported that SPARC-null cells showed decreased expression of collagen type I and addition of recombinant SPARC to SPARC null cells restored the expression of collagen type I to 70% [19].

5. Validation by perturbing networks and genetic association studies

The inferred regulatory relations between the genes in the network should be validated by perturbing network and analyzing its response to perturbation. The changes of regulatory roles of the genes from activation to inhibition or vice versa from inhibition to activation due to affection of tissues also should be validated by perturbing network. Several methods, for example, antisense RNA and RNA interference, can be used to perturb networks.

We used SPARC siRNA and CTGF siRNA to suppress the expression of SPARC and CTGF and RT-PCR to measure changes of expressions of collagen type I and type III for investigations of regulatory relations between genes in the network. Fig. 3 shows effects of SPARC siRNA and CTGF siRNA and TGF-β1 in transfected and normal fibroblasts. Several features emerged from Fig. 3. First, it showed in normal fibroblasts that TGF-β1 increased expression of CTGF and collagen type III. This observation was consistent with prediction of the structural equation model for the network. In both lesional and non-lesional skin tissues, regulation of TGF-β1 on CTGF through negative control of TGF-β1 on MAD3 and MAD3 on CTGF. Therefore, the structural equation model predicted that addition of TGF-β1 would increase expression of CTGF, which in turn increased the expression of collagen type III.

In fibroblasts transfected with SPARC siRNA, we observed that expression of both collagen type I and type III were decreased. The structural equation model precisely predicted a large reduction of expression of collagen type I and type III by inhibition of the SPARC gene. Significant changes in regulation of a gene may imply the occurrence of a mutation or gene variations within the gene, which provides information for identifying disease genes. This was supported by association studies of SPARC with SSc. Our previous genetic association studies in an isolated population with a high prevalence of SSc (Choctaw Indians), as well as in a multi-ethnic cohort of SSc patients, strongly suggest that the SPARC gene influenced disease susceptibility [23].

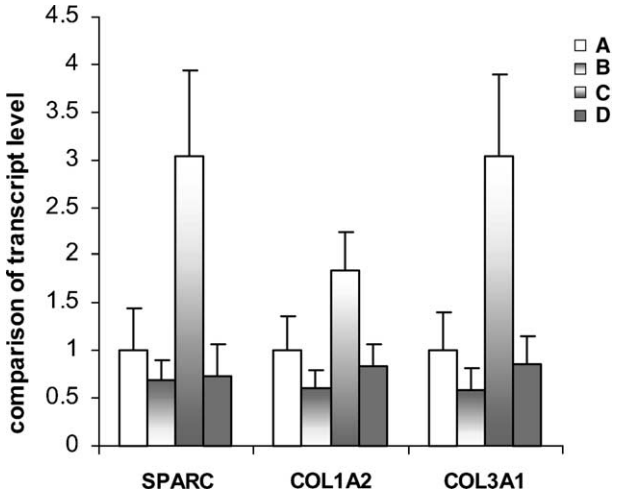

Fig. 3. Comparison of transcript levels of SPARC, COL1A2 and COL3A1 in different conditions. (A) Cultured fibroblasts with transfection media. (B) Cultured fibroblasts with SPARC siRNA transfection 10 μg/ml for 72 h. (C) Cultured fibroblasts treated with TGF-β1 10 ng/ml. (D) Cultured fibroblasts with SPARC siRNA transfection 10 μg/ml for 24 h before addition of TGF-β1 10 ng/ml. Error bars represent standard deviation in four assays.

6. Discussion

We present gene microarray studies of disease-discordant SSc twins as an example to illustrate the power of a systems biology approach to genetic studies of complex diseases. The traditional paradigm for identifying disease susceptibility genes is positional cloning that connects DNA variation to phenotypic variation. However, there is an intermediate 'omic' world between the DNA genotypes and the disease phenotypes – end point observations. The system from occurrence of DNA mutations to phenotypes through molecular events at the gene and protein level is highly likely to be organized into complicated biological networks. The number of paths from DNA genotypes to the end point observations – phenotypes are numerous. This may explain why positional cloning can only lead to limited success in identification of causative disease genes. It is therefore necessary to relate DNA variation to phenotypic variation through an 'omic' world organized into biological networks.

Systems biology that integrates genetic, transcriptomic, proteomic and metabolomic data to understand a whole biological system provides an exciting new paradigm for genetic studies of diseases. DNA genotypes provide only partial information on the connection of DNA to phenotypic variation. It is unlikely to directly map DNA genotypes to their corresponding phenotypes using DNA data alone. In the past decade, great

progress in molecular biology has paved the way to generate vast amounts of valuable data, including genotyping, profiling of mRNA, protein expression, gene–gene interaction, protein–protein interaction and biological networks. The 'omic' data provide the basis to conduct systems biology analyses for genetic studies of complex diseases. The analyses integrating 'omic' data offer insight into the pathogenesis of complex diseases that could not be gained by using each type of 'omic' data independently.

Key elements of a systems biology approach include (1) start working points, (2) network analysis, and (3) integration. Systems biology approaches can begin with any one type of 'omic' data: genotypes, gene expression, protein expression, and metabolic profiles, and then correlate each type of 'omic' data with end point observations, phenotypes. Positional cloning begins with genotype data and attempt to establish linkage or association of genotypes with phenotypes. In SSc studies, we started with gene expression data and intended to associate differentially expressed genes with SSc. By analyzing gene expression data in studies of discordant SSc twin pairs, we found that collagens, which are components of the extracellular matrix, and CTGF, which induces persistent fibrotic tissue formation, are associated with SSc. The results in the initial stage of gene discovery in the studies of complex diseases are used to identify a set of candidate genes.

To further analyze data for expanding the list of candidate genes and establishing formal burden of proof, we suggest performing network analysis. A gene does not work alone, but rather functions together with other genes interwoven into the network. The network analysis has two advantages. First, it will open a new way to identify causative genes for diseases. Second, it can provide insights into the pathogenesis of the disease. In a positional cloning approach, network analysis is difficult to perform. Although we can incorporate gene–gene interaction into the disease model, DNA data themselves do not provide information on reconstruction of gene networks. We have a long history of studies of genetic and metabolic networks that have been reconstructed mostly by experiments. Genetic and metabolic network databases provide information on the structure of the networks. An advancement of high throughput 'omic' technologies, interests in reconstruction of genetic and metabolic networks using mathematical models coupled with experiments are now resurging. In our example, we show how to use structural equations for modeling genetic networks. Starting with an initial set of candidate genes coupled with model selection, we reconstructed networks by searching whole genome gene expression profiles. Although the genes in the TGFβ pathway are known, the proposed structural equations for construction of genetic networks can infer relations between new, novel and uncharacterized genes. The reconstruction of genetic networks will expand the initial set of candidate genes. Alternative to structural equations for modeling genetic networks, Bayesian networks and other statistical methods can be used to model genetic networks. Advantage of the structural equations for construction of genetic networks, compared to Bayesian networks is that structural equations can model feedback structure of gene regulation networks, but in general, Bayesian networks are difficult to deal with feedback regulation. To identify causative genes for disease, we proposed the concept of differentially regulated genetic networks. We postulated that the changes in gene regulation in abnormal cells are due to gene variation. The preliminary results in SPARC genotyping analysis support this assumption.

Directionality of interactions was inferred from model selection, assuming that the network with correct causal relations should best fit the data. The inferred causality or regulatory relations among genes in the network should be validated by perturbing network using gene and/or protein regulatory methods. We propose to perturb networks for examining regulatory relationships between the genes in the network shown in the network model. This step is necessary because of biological and experimental variability in gene and protein expression data. We used RNA interference to perturb the networks and RT-PCR to measure gene expression levels. Total regulatory relations which we inferred are 12. Seven of them were known in the literature and remaining five relations were new. The results show that the identified new regulatory relationships between SPARC and collagen type I, SPARC and CTGF, and collagen type I and collagen type III in the model were supported by perturbation analysis of the network.

Integration of multiple types of data is important in system biology studies. Each type of 'omic' data provides only partial information. More importantly, each 'omic' data is only one level of multiple level organization of the biological system. We should study not only the relationships between the genes within one type of 'omic' data, but also their connections between different types of 'omic' data. For example, we first studied the regulatory relationship between SPARC and collagen type I using gene expression data. Then, we correlated regulation of SPARC on collagen type I to phenotypes by estimating the regulatory effect of SPARC on collagen type I in the structural equation model using gene expression profiles in normal and abnormal tissues separately. We found that regulation of SPARC on Collagen type I was changed from inhibition in normal tissues to activation in SSc disease tissues. We further correlated regulation data with DNA genotypes and found that the regulation changes were likely due to gene sequence variation in the SPARC gene. Integration of 'omic' data allows one to reveal the path from DNA mutation to phenotypes through gene–protein-metabolic interaction, and to gain deeper insights into the developments of the diseases.

The results of TGF-β pathways in SSc are very limited. Like any statistical inferences, the reliability and robustness of the model inference depend on the number of tissue samples used for gene expression profiling, which in turn depends on the number of genes in the networks. There are no theoretic sample sizes in construction of genetic networks. Heuristically, we suggested using as many samples as four times of the number of the genes in the networks. The sample sizes will limit the size of the inferred genetic networks. One way to overcome this problem is to decompose a large network into several modules of the networks with small size. For each module of the network we can make robust inference.

The purpose of this example is to illustrate the basic scheme of systems biology approach to genetic studies of complex diseases. We did not study the interactions between environments and genes (in each 'omic'). Environments will definitely affect transcription, translation, and metabolism of the genes. We should keep in mind that the real biological systems are extremely complicated. However, the information available for biological systems will be increased when we have more 'omic'

data. The increased information in conjunction with development of mathematical models will offer exciting perspectives for uncovering the causes of many diseases.

Acknowledgements: This work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Disease (NIAMS) Specialized Center for Research in Scleroderma Grant IP50AR44888 (F.C.A.), Grant AR050840 (C.A.F.), the Arthritis Foundation (C.A.F.), the Scleroderma Foundation (X.Z.), NIAMS Grant AR050517-01A2 (X.Z.), NIH PHS NCRR GCRC Grant M01RR002558, University of Texas Health Science Center at Houston UCRC CreFF Grant (X.Z.) and NIH Grant ES09912 (M.X.).

References

[1] Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. Science 273, 1516–1517.
[2] Glazier, A.M., Nadeau, J.H. and Aitman, T.J. (2002) Finding genes that underlie complex traits. Science 298, 2345–2349.
[3] Page, G.P., George, V., Go, R.C., Page, P.Z. and Allison, D.B. (2003) "Are we there yet?": deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am. J. Hum. Genet. 73, 711–719.
[4] Sing, C.F., Stengard, J.H. and Kardia, S.L. (2003) Genes, environment, and cardiovascular disease. Arterioscler. Thromb. Vasc. Biol. 23, 1190–1196.
[5] Alon, U. (2003) Biological networks: the tinkerer as an engineer. Science 301, 1866–1867.
[6] Ge, H. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet. 19, 551–560.
[7] Goldstein, D.B., Ahmadi, K.R., Weale, M.E. and Wood, N.W. (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. Trends Genet. 19, 615–622.
[8] Okuda, T., Sumiya, T., Mizutani, K., Tago, N., Miyata, T., Tanabe, T., Kato, H., Katsuya, T., Higaki, J., Ogihara, T., Tsujita, Y. and Iwai, N. (2002) Analyses of differential gene expression in genetic hypertensive rats by microarray. Hypertens. Res. 25, 249–255.
[9] Strohman, R. (2002) Maneuvering in the complex path from genotypes to phenotype. Science 296, 701–703.
[10] Bollen, K.A. (1989) Structural Equations with Latent Variables, John Wiley & Sons, New York.
[11] Claman, H.N. (1991) Endothelial and fibroblastic activation in scleroderma. The myth of the "uninvolved skin". Arthritis Rheum. 34, 1495–1501.
[12] Zhou, X.D., Tan, K.F., Xiong, M., Arnett, F.C., Feghali-Bostwick, C.A. (2005). Arthritis Rheum. 52, (in press).
[13] Yamane, K., Ihn, H., Asano, Y., Jinnin, M. and Tamaki, K. (2003) Antagonistic effects of TNF-alpha on TGF-beta signaling through down-regulation of TGF-beta receptor type II in human dermal fibroblasts. J. Immunol. 171, 3855–3862.
[14] Waite, K.A. (2003) From developmental disorder to heritable cancer: it's all in the BMP/TGF-beta family. Nat. Rev. Genet. 4, 763–773.
[15] Zimmerman, C.M. and Padgett, R.W. (2000) Transforming growth factor beta signaling mediators and modulators. Gene 249, 17–30.
[16] Igarashi, A., Okochi, H., Bradham, D.M. and Grotendorst, G.R. (1993) Regulation of connective tissue growth factor gene expression in human skin fibroblasts and during wound repair. Mol. Biol. Cell. 4, 637–645.
[17] Dammeier, J., Brauchle, M., Falk, W., Grotendorst, G.R. and Werner, S. (1998) Connective tissue growth factor: a novel regulator of mucosal repair and fibrosis in inflammatory bowel disease?. Int. J. Biochem. Cell. Biol. 30, 909–922.
[18] Frazier, K., Williams, S., Kothapalli, D., Klapper, H. and Grotendorst, G.R. (1996) Simulation of fibroblast cell growth, matrix production, and granulation tissue formation by connective tissue growth factor. J. Invest. Dermatol. 107, 404–411.
[19] Francki, A. (1999) SPARC regulates the expression of collagen type I and transforming growth factor-beta1 in mesangial cells. J. Biol. Chem. 274, 32145–32152.
[20] Houglum, K. (1997) Proliferation of hepatic stellate cells is inhibited by phosphorylation of CREB on serine 133. J. Clin. Invest. 99, 1322–1328.
[21] Trojanowska, M. (2002) Molecular aspects of scleroderma. Front. Biosci. 7, 608–618.
[22] Shi-wen, X. (2000) Autocrine overexpression of CTGF maintains fibrosis: RDA analysis of fibrosis genes in systemic sclerosis. Exp. Cell Res. 259, 213–224 Bertsekas, D.P. (1995) Nonlinear Programming, Athena Scientific, Belmont, MA.
[23] Zhou, X., Tan, F.K., Reveille, J.D., Wallis, D., Milewicz, D.M., Ahn, C., Wang, A. and Arnett, F.C. (2002) Association of novel polymorphisms with the expression of SPARC in normal fibroblasts and with susceptibility to scleroderma. Arthritis Rheum. 46, 2990–2999.