

Learning More from Microarrays: Insights from Modules and Networks

David J. Wong and Howard Y. Chang

Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California, USA

Global gene expression patterns can provide comprehensive molecular portraits of biologic diversity and complex disease states, but understanding the physiologic meaning and genetic basis of the myriad gene expression changes have been a challenge. Several new analytic strategies have now been developed to improve the interpretation of microarray data. Because genes work together in groups to carry out specific functions, defining the unit of analysis by coherent changes in biologically meaningful sets of genes, termed modules, improves our understanding of the biological processes underlying the gene expression changes. The gene module approach has been used in exploratory discovery of defective oxidative phosphorylation in diabetes mellitus and also has allowed definitive hypothesis testing on a genomic scale for the relationship between wound healing and cancer and for the oncogenic mechanism of cyclin D. To understand the genetic basis of global gene expression patterns, computational modeling of regulatory networks can highlight key regulators of the gene expression changes, and many of these predictions can now be experimentally validated using global chromatin-immunoprecipitation analysis.

Key words: genomics/microarray/computational biology/gene expression
J Invest Dermatol 125:175–182, 2005

Biological research has been revolutionized by the complete genome sequences of hundreds of organisms, giving rise to the new science of genomics. Microarray analysis takes advantage of the vast amount of sequence information and allows investigators to examine alterations in the gene expression of thousands of genes simultaneously in a single experiment. The basic experimental setup and procedures in standard microarray experiments have been discussed in several excellent reviews (Brown and Botstein, 1999; Churchill, 2002). Global gene expression patterns have the potential to better define biologic phenomena and human disease states at the molecular level. For example, based on the differential expression of hundreds to thousands of genes, many cancers previously thought to be homogenous are now recognized to consist of distinct molecular subtypes and are often associated with significantly different clinical outcomes (Golub *et al*, 1999; Alizadeh *et al*, 2000). Microarray analysis have also been applied to several dermatologic diseases, including melanoma, mycosis fungoides, cutaneous B cell lymphoma, psoriasis, atopic dermatitis, alopecia areata, and scleroderma (Bittner *et al*, 2000; Clark *et al*, 2000; Bowcock *et al*, 2001; Oestreicher *et al*, 2001; Zhou *et al*, 2001, 2003; Carroll *et al*, 2002; Nomura *et al*, 2003; Storz *et al*, 2003; Tracey *et al*, 2003; Whitfield *et al*, 2003).

Making biological sense out of the laundry list of up- and down-regulated genes from microarray experiments is,

however, often difficult. How does one interpret what biological processes underlie the expression changes of hundreds to thousands of genes, and which one of the myriad genes is the key regulator that allows the investigator to experimentally manipulate the underlying biology? We will discuss two new bioinformatic strategies that have been developed to improve interpretation of microarray data: gene module and regulatory network analysis. These methods aim to identify the relevant pathways and key regulators from microarray data of any biologic process or disease state. Resources for these methods are listed in Table I.

In the Beginning

Traditional microarray analysis aims to identify individual genes that are consistently up- or down-regulated among samples that are known to be different, or to identify consistent variation across a set of heterogeneous samples. Therefore, the expression value of each gene across all samples is analyzed with *t* tests or analysis of variance; the significance of the difference in expression is adjusted for the number of times the statistical test is done. (If 10,000 genes were examined, 500 genes should be significant at $p = < 0.05$ based on chance alone. Therefore, more stringent criteria or a calculation of *false discovery rate* given the sample size are necessary. If 1000 genes were identified at $p = < 0.001$, the false discovery rate will be $(10,000 \times 0.001)/1000 = 0.01$). Investigators also typically focus on genes that change in expression beyond a threshold level (typically 2-fold) relative to the control sample. Investigators then examine the selected genes to develop hypotheses

Abbreviations: cDNA, complementary DNA; CDK, cyclin-dependent kinase; ChIP-chip, chromatin-immunoprecipitation followed by microarray analysis

Table I. Informatic resources on the Internet

Tool	Function	Web site
Cluster	Performs hierarchical clustering. Genes and samples in microarray experiments are organized by similarity.	http://rana.lbl.gov/EisenSoftware.htm
GenePattern	Analysis and data visualization software from Broad Institute at MIT. Performs sequence and microarray analysis, including Kolmogorov–Smirnov score test.	http://www.broad.mit.edu/cancer/software/genepattern/
GeneXpress	(i) Implements Gene Module Map.	http://robotics.stanford.edu/~erans/cancer/
	(ii) Identify enriched transcription factor binding sites in the promoters of gene modules.	
	(iii) Implements module networks.	
Gene Ontology Term Finder	Gene Ontology classifies each gene in the genome using a controlled vocabulary. Tool identified enriched GO terms within groups of select genes.	http://search.cpan.org/dist/GO-TermFinder/
GeneHopping	Identifies sets of co-regulated genes between organisms and provides visualization of modules.	http://barkai-serv.weizmann.ac.il/Software/GeneHopping/Hopping.html
Onto-Tools	Web-based program to (i) identify GO term enrichment, (ii) map probes among different array platforms, (iii) retrieve annotations of specific genes.	http://vortex.cs.wayne.edu:8080/ontoexpress/servlet/UserInfo

about the biological mechanism behind the disease process. For example, Clark *et al* found the RhoC gene, a GTPase that regulates cytoskeletal organization, to be up-regulated in melanoma metastases relative to the primary tumor. They subsequently demonstrated that metastasis was enhanced by RhoC overexpression and inhibited with dominant-negative RhoC (Clark *et al*, 2000). Tracey *et al* (2003) identified genes in the TNF signaling pathway to be preferentially expressed in mycosis fungoides lesions relative to other inflammatory disorders. Although these gene-by-gene analysis methods are powerful and will continue to generate new biological insights, new bioinformatic strategies have been developed to learn more from microarray data.

Gene Modules

Gene module analysis is based on the simple idea that genes typically work together in groups, such as in enzymatic pathways or regulatory cascades. Thus, the unit of analysis in a microarray experiment should be groups of functionally related genes, termed modules, and one assigns more significance to a group of genes having coordinate regulation than just one member of a group being regulated in the experiment (Mootha *et al*, 2003; Segal *et al*, 2004). Using previous biological knowledge, gene modules can be defined by sets of genes that are members of the same biological pathway, have a shared structural motif, are expressed in a specific tissue, or are induced by a specific stimulus. A biological pathway is typically defined by two gene modules: one for the up-regulated genes in the pathway and one for the down-regulated genes. Gene modules can have different numbers of member genes, and each gene can belong to multiple modules. In a microarray experiment, gene module analysis searches for coordinate regulation of genes that belong to these *a priori* defined

gene modules; a statistical test is performed for each module relative to all other genes on the microarray to calculate whether the degree of coordinate regulation is more than one would expect by chance.

The gene module approach has several advantages over our current gene-by-gene methods of analysis. First, gene module analysis can detect meaningful expression patterns that would otherwise go undetected (Fig 1). Coordinated small magnitude regulation of gene expression of many genes in the same pathway can be biologically more important than a large magnitude change that is discordant with other members of the pathway; however, this type of regulation is often missed by the gene-by-gene approach. Moreover, the large number of genes examined in the gene-by-gene approach necessitates significant penalties for multiple hypothesis testing; many biologically meaningful changes can be missed. Gene module analysis takes advantage of the power of groups of genes to detect those genes that have biologically significant, albeit subtle, expression changes. Secondly, because modules are defined by groups of genes known to share certain biological function or characteristics, defining the unit of analysis by modules improve the investigator's mechanistic interpretation of the biology underlying the gene expression changes. For example, modules can consist of groups of genes previously found to be coordinately regulated in other microarray experiments. In this way, gene module analysis allows one to compare each new microarray experiment to every previously performed experiment to identify commonalities and unifying mechanisms. By analogy with sequence searches where a newly cloned gene is compared to genes in the database for blocks of sequence similarity, gene module analysis allows one to discover features of gene expression patterns that have been observed in other microarray experiments.

It should be noted that the strength of gene module analysis, which lies in previous biological knowledge, is also

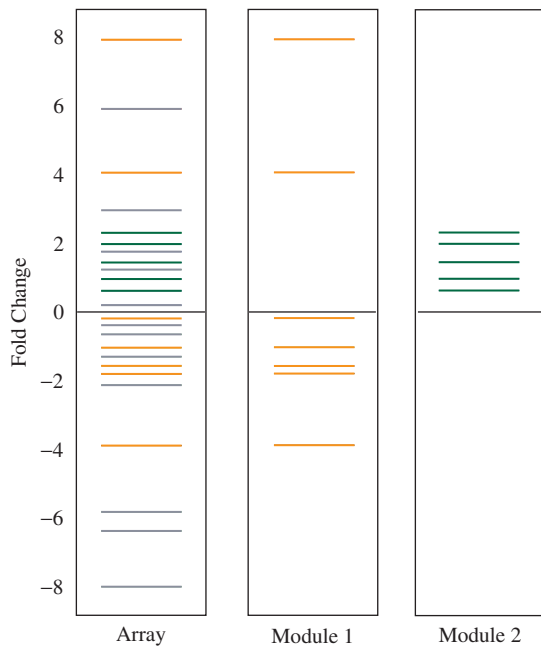


Figure 1
Gene module analysis can detect subtle changes in gene expression from microarray experiments. Each line within the left-hand box labeled “Array” represents a gene positioned along the y-axis according to fold change in expression relative to a reference in a microarray experiment. “Module 1” and “Module 2” are *a priori* defined sets of genes that are co-activated in two distinct biological pathways; their respective boxes contain their members arranged according to their relative fold change in expression in the “Array” and labeled with the colors orange and green, respectively. Individual gene approaches would detect the two highly upregulated genes in “Module 1” but would not detect any significant change in any of the genes in “Module 2”. Gene module analysis, however, demonstrates that “Module 2” may be more significant than “Module 1” in this experiment because “Module 1” has genes that are both upregulated and downregulated, whereas all of the genes in “Module 2” are upregulated.

its limitation. Many genes in the human genome have unknown or poorly defined function and thus are not included in gene modules. Many biological processes have not been examined by previous genetic or expression profiling experiments, and thus gene modules that correspond to these processes need to be experimentally defined. In addition, poorly defined gene modules may result in flawed data interpretation; module analysis needs to start from primary expression data that has sound experimental design and good technical quality. High quality expression data are derived from well-defined biological samples, have adequate median intensity signal, and are validated with replicates and other independent methods such as RT-PCR. Keeping these caveats in mind, we highlight several examples to illustrate the power and diverse uses of gene module analysis.

Three Examples of Gene Module Analysis

At its simplest level, gene module analysis is no more than known gene expression patterns that serve as positive controls to interpret new data. Whitfield *et al* (2003) compared global gene expression patterns of skin from patients with systemic sclerosis and normal volunteers. Interestingly,

the gene expression pattern of clinically uninvolved skin from scleroderma patients was more similar to that of clinically involved skin of scleroderma patients than skin from normal volunteers. Because skin tissue is composed of various types of cells, the authors determined the contribution of each cell type to the pathologic gene expression pattern in scleroderma using a rudimentary form of module analysis. Whitfield *et al* determined the global gene expression patterns of 11 different cell lines, including epithelial cells, fibroblasts, endothelial cells, smooth muscle cells, and B lymphocytes, to construct canonical gene expression patterns of each cell type. Comparison with the scleroderma gene expression pattern demonstrated a strong contribution of the B lymphocyte signature, suggesting a potential role for B cells in the pathogenesis of scleroderma (Whitfield *et al*, 2003). Consistent with this result, B cells also appear to have a pathogenic role in the tight-skin mouse model of scleroderma (Saito *et al*, 2002). These studies support the idea that rituximab, an anti-CD20 monoclonal antibody which targets B cells, may be useful in the treatment of scleroderma.

In the second example, Mootha *et al* (2003) pioneered a type of gene module analysis (which they termed gene set enrichment analysis or GSEA) to discover the biological pathways underlying type II diabetes mellitus. They compared global gene expression patterns of skeletal muscle biopsies from individuals with normal glucose tolerance, impaired glucose tolerance, and type II diabetes mellitus. After rigorous statistical tests on a gene-by-gene basis (and suffering the concomitant multiple hypothesis testing penalty), they found no single gene with a significant difference in expression. Mootha *et al* noticed, however, that many genes that showed the most consistent changes encoded enzymes involved in mitochondrial oxidative phosphorylation. To test the significance of this observation, the authors implemented gene module analysis by constructing 149 modules of various metabolic pathways or co-regulated genes. The authors sorted all genes on the microarray into a ranked list, from the one best able to distinguish diabetes *versus* normal to the least informative. They then asked if the distribution of genes on this list is surprising given the membership of genes in modules.

Specifically, the authors applied the Kolmogorov–Smirnov running sum statistic: Beginning with the gene at the top of the ranked list, the running sum increases when a gene that is a member of the gene set is encountered and decreases otherwise. The maximum enrichment score is the greatest positive deviation of the running sum across all genes. To determine the statistical significance of the maximum enrichment score and validate that the results are unlikely to arise by chance alone, permutation testing is performed, comparing the maximum enrichment score using the actual data to that seen in each of 1000 permuted data sets.

GSEA revealed that a module of genes involved in oxidative phosphorylation was significantly downregulated in patients with diabetes. Each gene in the oxidative phosphorylation gene module was transcriptionally downregulated by roughly only 20%, and thus was not clearly detected at the individual gene level. Independent work by Shulman and colleagues using magnetic resonance

spectroscopy confirmed that defective mitochondrial oxidative phosphorylation is strongly associated with glucose intolerance and appears to be a strong predictor for development of diabetes (Petersen *et al*, 2003).

In addition to looking at whether each gene module is significantly regulated in the experiment, gene module analysis also examines which particular genes within a module are contributing to the regulation. This information can refine the gene module and lead to additional mechanistic insight. For instance, Mootha *et al* (2003) noticed that not all genes in the oxidative phosphorylation module were equally downregulated in diabetes mellitus; the subset of genes that were downregulated consisted of many known targets of the transcriptional coactivator PGC-1 α in muscle cells (Mootha *et al*, 2003). Analysis of shared promoter elements of genes that comprise the refined oxidative phosphorylation module has identified two transcription factors, estrogen receptor related α and GA-repeat binding protein, as key regulators that cooperate with PGC-1 α to regulate expression of this gene module and cellular energy metabolism (Mootha *et al*, 2004). Thus, gene module analysis has generated a model for impaired glucose tolerance and diabetes mellitus in which the downregulation of PGC-1 α function in skeletal muscle results in the downregulation of genes involved in oxidative phosphorylation.

Thirdly, gene module analysis has also been used for exploratory discovery of the shared biological pathways that underlie different human cancers. Using a strategy termed Gene Module Map, Segal *et al* (2004) performed a comprehensive analysis of 1975 previously published microarrays with 2849 gene modules. These gene modules included tissue-specific genes, co-regulated genes, and genes that function in the same process, act in the same pathway, or share similar subcellular localization. Each microarray experiment was also annotated using a controlled vocabulary for several hundred biological and clinical conditions that it represents, including tumor type, stage, and clinical outcomes. For each gene module and array, they calculated the fraction of genes from that module that was induced or repressed in that array and asked if this fraction of enrichment was surprising based on chance alone, estimated using the hypergeometric distribution. A similar algorithm was applied to the clinical annotations, and clinical annotations that were enriched for each gene module were identified.

In this fashion, the large number of microarray experiments and their associated clinical information was distilled to a core set of relationships that defined each cancer by a specific combination of gene modules, many of which provide insight into molecular mechanisms underlying cancer phenotypes. For example, poorly differentiated tumors of many histologic types were found to share an activation of the spindle checkpoint and M phase modules, which have been previously associated with chromosomal instability and aneuploidy. Many modules that were specific to particular types of cancer or even stages of the same disease were also identified, such as deactivation of a growth inhibitory module of dual specificity phosphatases in acute lymphoblastic leukemia and repression of an intermediate filament module in breast cancers (Segal *et al*, 2004). Although it is impressive that many of the newly described

relationships between gene modules and their respective cancers can be supported by the literature, the significance of the majority of gene modules in human cancers awaits experimental validation.

Gene Modules Across Evolution

To truly assess the biological significance of gene modules, one would need functional studies where one can perturb gene co-expression and observe the biological effects. Are such mutant organisms less fit across diverse environmental and physiologic conditions? In fact, this functional experiment has been carried out in a comprehensive fashion by evolution. The functional importance of co-regulated genes can be gleaned by evolutionary conservation of their co-expression in multiple organisms. Genes that must be co-expressed together, such as genes that encode subunits of the ribosome or the proteasome, will be under evolutionary pressure to maintain their coordinated expression. Thus, co-expression across evolutionary time can better define functionally important sets of co-regulated genes compared to using data from only a single species.

Stuart *et al* (2003) and Bergmann *et al* (2004) demonstrate this concept by mapping all orthologues between four and six species, respectively, and searched for coordinate regulation of genes in the published microarray data from humans, *Drosophila*, *Caenorhabditis elegans*, *Arabidopsis*, *Saccharomyces cerevisiae*, and *Escherichia coli*. These multi-species analyses defined new gene modules based on co-expression across these diverse organisms. Through this process, they discovered potential functions for novel genes based on their co-expression with genes that have known functions. They identified groups of co-expressed genes that defined multiple essential cell processes; one of the largest identified by Stuart *et al* was a module enriched for genes involved in the cell cycle and cell proliferation. Based on evolutionary co-expression, novel genes in this module are predicted to also have functions in cell proliferation. Stuart *et al* selected a subset of the novel genes for validation experiments in two different organisms. The novel genes predicted to have functions in cell proliferation genes were found to be overexpressed in human pancreatic cancer relative to normal tissue, and RNA interference of one of these genes in *C. elegans* resulted in excess nuclei in the germ line, thus supporting its role in regulating cell division. These results reinforce the concept that systematic comparison of gene expression patterns across diverse experiments and even organisms can highlight the functional roles and molecular relationships between genes.

Genomic Methods of Hypothesis Testing

Expression profiling has previously been used as an exploratory research tool to suggest new models and hypotheses. Expression profiling also can, however, be used for hypothesis testing on a genome-wide scale (Lamb *et al*, 2003; Chang *et al*, 2004). Lamb *et al* used expression profiling and gene module analysis to test alternative

hypotheses of the oncogenic mechanism of cyclin D1 overexpression in human cancer: the well-known effects of activation of cyclin-dependent kinases (CDK), leading to phosphorylation of Rb and derepression of E2F target genes, versus alternative effects of CDK-independent activation on gene expression. Lamb *et al* first identified an expression profile of cultured human mammary epithelial cells ectopically expressing cyclin D1 and defined the induced genes in this experiment as a cyclin D1 gene module. Interestingly, the cyclin D1 gene module did not include E2F target genes and the expression signature was recapitulated by expression of a cyclin D1 mutant unable to activate CDK4. Lamb *et al* used published datasets of global gene expression patterns from 190 primary human tumors to determine whether the cyclin D1 module defined *in vitro* was found in human tumors. They ordered the expression pattern of all genes in the genome according to its similarity to the expression pattern of cyclin D1 across all of the tumors; they then applied the Kolmogorov–Smirnov running sum statistic to test if members of the cyclin D1 gene module were overrepresented at the top of the list. The *in vitro* defined cyclin D1 module correlated with cyclin D1 levels in human tumors, but a module of known E2F target genes did not, suggesting that in authentic human tumors, overexpression of cyclin D1 results in predominantly CDK-independent effects. Lamb *et al* cloned a subset of the promoters of genes in the cyclin D1 module, mapped the sites of cyclin D1 responsiveness within these promoters using reporter gene assays, and identified a sequence with strong similarity to the consensus binding site of C/EBP β , which was one of the most highly ranked genes in the cyclin D1 module in human tumors. Cyclin D1 was found to interact with C/EBP β and relieved transcriptional repression by C/EBP β . Indeed, cyclin D1 could not transactivate its target genes in C/EBP β -deficient cells, demonstrating that C/EBP β is an essential transcriptional effector of cyclin D1. Because overexpression of an inhibitory isoform of C/EBP β in transgenic mice was also previously found to be oncogenic (Zahnow *et al*, 2001), the gene module analysis, combined with functional studies, have linked these two pathways and suggests an alternative oncogenic mechanism of cyclin D1 through C/EBP β in human cancers (Lamb *et al*, 2003). In this instance, gene module analysis provided the formal genetic evidence that cyclin D1 functions via the C/EBP pathway in human cancers, and classic molecular experiments provided mechanistic depth to how this connection functions *in vivo*.

Similarly, Chang *et al* (2004) used a gene module approach to test the hypothesis that genes involved in wound healing are reactivated in cancer progression. Based on the idea that processes in the wound response such as cell migration, matrix remodeling, and angiogenesis are well suited to tumor cell invasion and metastasis, Chang *et al* hypothesized that wound response genes may be evident in a subset of cancers that are destined to become metastatic. Because fibroblasts *in vivo* are exposed to serum, the soluble fraction of clotted blood, only in the context of tissue injury, Chang *et al* used detailed *in vitro* experiments to define a module of genes that characterized the response of fibroblasts to serum exposure as a proxy for wound-healing genes. The authors then compared the expression pattern

of the wound response module in approximately 500 published global gene expression profiles of human tumors and corresponding normal tissues, and found that the wound response module was activated in a subset of human cancers from various tissues but not in their orthotopic normal counterparts (Fig 2). In addition, primary tumors expressing the wound response module were more likely to progress to metastasis and death compared to tumors that did not (Fig 2). The best validation of a gene module's prognostic value is to test its ability to predict outcome in large independent sets. Chang *et al* (2005) subsequently examined a database of 295 breast cancer patients and found that the prognostic power of the wound response gene module was reproducible in this independent set of patients. In early breast cancer, the wound response module was the most significant predictor of metastasis compared to established clinical and histologic criteria in multivariate analysis, highlighting the wound response module as a powerful prognostic marker (Chang *et al*, 2005). Thus, both Lamb *et al* and Chang *et al* characterized gene modules of interest by

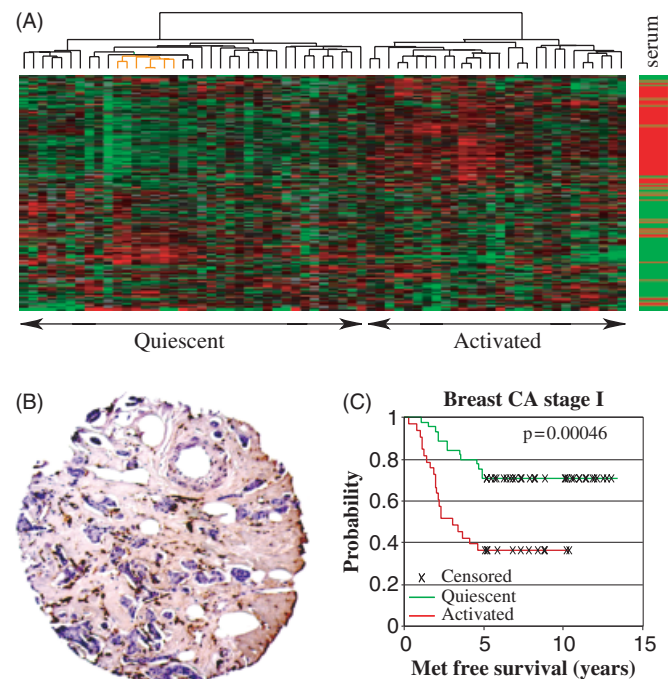


Figure 2

Wound response gene module predicts breast cancer progression.

(A) Expression of the fibroblast serum response genes in breast cancer and benign breast samples reveals a biphasic pattern. The status of serum regulation of each gene is shown on the right bar: red indicates serum-induced; green indicates serum-repressed. Analysis reveals samples with expression of the serum-repressed genes (quiescent group) and those with expression of the serum-induced genes (activated group). The seven benign samples are indicated by orange branches. (B) Validation of gene expression by *in situ* hybridization on tissue microarrays. Expression of LOXL2, a fibroblast serum-induced gene, on a 600 micron core of breast carcinoma is shown. LOXL2 is expressed in stromal fibroblasts (stained brown) but not in tumor cells (counterstained purple by hematoxylin), validating the expression array data of an activated wound response signature. (C) Prognostic value of wound response gene module. Kaplan–Meier survival curves of 78 breast cancer patients prospectively classified as having activated or quiescent stroma. Patients were matched for age, histologic diagnosis, and tumor stage and underwent the same treatment protocol. Tumors with activated phenotype had significantly increased risk of metastasis and mortality.

expression profiling of *in vitro* model systems and then used the resulting gene modules to query microarray data from primary human tissue to test hypotheses about the genetic pathways in human cancers. Both studies demonstrate the ability of gene module analysis to provide novel and valuable insight into complex human diseases that are not amenable to traditional genetic techniques.

Regulatory Networks

In many biological studies, we are interested in identifying causal relationships—i.e., gene A is upstream of gene B and induces B. Several investigators have applied probabilistic graphical models, specifically Bayesian Networks, to identify regulatory relationships from static views of global gene expression patterns (Segal *et al*, 2003; Beer and Tavazoie, 2004). Bayesian Networks are a particularly useful type of model because they organize a set of variables into a hierarchical model of conditional probabilities; the value of the daughter variables are the joint conditional probability of the parent variables. Typically, the model is used to evaluate many combinations of specific variables, and particular models that produce good fit of the data are validated by additional computational and experimental tests (Friedman, 2004).

For example, because microarray data provide a global view of mRNA abundance, the underlying regulatory network could be the set of active transcription factors or the set of promoter and enhancer elements that produced the genome-wide transcriptional pattern. Segal *et al* approached this problem by reasoning that many transcription factors and signal transducers are under transcriptional control themselves; thus, a regulatory model may be constructed by relating the expression pattern of genes that encode transcription factors to that of all other genes (Segal *et al*, 2003). Segal *et al* developed a probabilistic Bayesian algorithm, termed module networks, to identify the correlations between the expression level of a manually curated set of genes encoding transcription factors and signaling proteins, termed regulators, and all other transcribed genes (Segal *et al*, 2003). Transcribed genes were grouped into modules based on the expression changes of the regulators, and regulators were allowed to combine into hierarchical patterns that were conditional, additive, or antagonistic. Thus, unlike hierarchical clustering that only identifies genes with similar patterns of expression, regulatory programs allowed logical operations such as AND, OR, IF, and BUT. An iterative process of regulatory tree building and gene module assignment is performed to optimize both predictions. In each iteration, the procedure learns the best regulation program for each module, and given the inferred regulation programs, reassigns each gene to the module whose program best predicts its behavior. These two steps are iterated until convergence is reached. This method has the advantage of generating testable hypotheses about gene modules and their regulatory programs in a single analysis. This method, however, is limited by current biological knowledge because it relies on compiling a list of candidate regulators. In addition, although this strategy can accommodate heterologous data such as

proteomic or enzyme activity profile data, currently most high throughput data of regulators are transcriptional analyses. Thus, the predicted regulatory trees can be wrong because they fail to take into account post-transcriptional and post-translational regulation.

To demonstrate the power of this strategy, Segal *et al* (2003) used a set of 466 candidate regulators and a set of 173 arrays that measure responses of *S. cerevisiae* to various stresses, which resulted in 50 modules with regulation trees. It should be noted that this type of algorithm will always produce a regulatory tree; the key assessment is the quality of the regulatory trees and gene modules that are produced. A good regulatory tree will encompass transcription factors that are known to act or interact with one another, and the gene modules will have member genes that can be shown to have shared functions. Segal *et al* found that thirty-one of the 50 modules had over 50 percent of its genes with the same functional annotation, thirty of 50 modules included genes previously known to be regulated by at least one of the module's predicted regulators, and fifteen of the 50 modules had a match between the *cis*-regulatory motifs in the upstream regions of the module's genes and the regulator known to bind to that motif. This is a rather remarkable feat given that the only input information was gene expression data; no biochemical, genetic, or sequence data was used to make the predictions. To further validate this strategy, Segal *et al* chose three novel regulatory relationships predicted by the regulatory network model, mutated each regulator, and performed global gene expression analysis. In all three cases, the deletion mutants selectively affected the gene modules that they were predicted to regulate. Thus, module networks is a useful method for generating hypotheses that can accurately predict regulators, the processes that they regulate, and the conditions under which they are active.

Hypotheses of regulatory mechanisms can also be generated from shared *cis*-regulatory DNA elements in the upstream regions of the genes in each module. Beer and Tavazoie (2004) demonstrate a computational modeling method for building regulatory networks based on these regulatory DNA elements on a genome-wide scale. They used a similar Bayesian probabilistic model to identify the upstream DNA motifs that predict the expression pattern of each gene module under different conditions. Beer and Tavazoie demonstrate that prediction with DNA elements requires complex rules because the expression level of a gene is controlled by the occupancy states of multiple upstream binding sites. They validated their regulatory network models by demonstrating that they correctly predicted the expression patterns of 73% of the genes using 20% leave-out analysis. Although this *cis*-regulatory DNA elements method would not reliably predict genes that are regulated by more distant DNA elements or that have alternative regulatory mechanisms such as chromatin modification, they demonstrate that local upstream DNA sequence can predict gene expression patterns of a large portion of genes, at least in *S. cerevisiae*.

The large number of hypotheses generated from regulatory networks analysis or *cis*-regulatory DNA elements analysis can be validated in a high throughput fashion using chromatin-immunoprecipitation followed by microarray

analysis (ChIP-chip) (Fig S1). ChIP-chip, initially described by Ren *et al* (2000) and Iyer *et al* (2001) uses antibodies specific to a candidate regulator for genome-scale chromatin-immunoprecipitation combined with microarrays spotted with intergenic and promoter sequences to identify their bound targets. For example, Odom *et al* (2004) demonstrate the use of antibodies to HNF1 α , HNF4 α , and HNF6 to identify all genes that are bound by these transcription factors in human liver and pancreas tissues, and hybridized the immunoprecipitated chromatin fragments to a custom microarray of over 10,000 human promoter sequences. These results revealed that a surprisingly large fraction of genes transcribed in the liver and pancreas were bound by HNF4 α , providing a molecular explanation for the role of HNF4 α mutations and polymorphisms in hereditary and sporadic forms of diabetes mellitus. In addition, as mentioned above, regulator networks also can be validated by expression profiling of mutants of the regulator to see if its signature recapitulates the effect on the genes in the module that it was predicted to regulate.

The power of these new bioinformatic methods relies on *a priori* biological knowledge to compile gene modules, functional annotations, and/or lists of candidate regulators. Gene expression databases are publicly available that contain many different expression profiles that show how expression is altered by stress, disease, pharmaceuticals, developmental stages, different growth conditions, and genetic perturbations. The growing wealth of contributions to these databases improves our ability to extract meaningful information from microarray data and thus is a growing resource for future discoveries. This emphasizes the importance of unrestricted access to published microarray data to the entire scientific community to reach the full potential of scientific progress.

The post-genome era offers great promise toward our understanding of the molecular mechanisms underlying human disease. As a result of the rapid advancement of bioinformatic techniques, microarray analysis can be used for both exploratory discovery and definitive hypothesis testing to learn biologic pathways and their regulatory networks. These analyses can then be validated with genome-wide techniques such as ChIP-chip as well as traditional genetic and biochemical experiments. Resources for these techniques are readily available on the web (Table I). The integration of experimental and computational biology has the potential to promote rapid advancement in the diagnosis, prognosis, and treatment of human diseases.

We thank E. Segal and J. Rinn for discussions and A. Adler for comments on the manuscript. H. Y. C. is supported by a grant from NIAMS (AR050007).

Supplementary Material

The following supplementary material is available for this article online.
Figure S1 ChIP-chip.

DOI: 10.1111/j.0022-202X.2005.23827.x

Manuscript received September 27, 2004; revised March 29, 2005; accepted for publication April 15, 2005

Address correspondence to: Howard Y. Chang, Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA 94305, USA. Email: howchang@stanford.edu

References

- Alizadeh AA, Eisen MB, Davis RE, *et al*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511, 2000
- Beer MA, Tavazoie S: Predicting gene expression from sequence. *Cell* 117:185–198, 2004
- Bergmann S, Ihmels J, Barkai N: Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* 2:85–93, 2004
- Bittner M, Meltzer P, Chen Y, *et al*: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406:536–540, 2000
- Bowcock AM, Shannon W, Du F, *et al*: Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum Mol Genet* 10:1793–1805, 2001
- Brown PO, Botstein D: Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21:33–37, 1999
- Carroll JM, McElwee KJ, E King L, Byrne MC, Sundberg JP: Gene array profiling and immunomodulation studies define a cell-mediated immune response underlying the pathogenesis of alopecia areata in a mouse model and humans. *J Invest Dermatol* 119:392–402, 2002
- Chang HY, Nuyten DS, Sneddon JB, *et al*: Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci USA* 102:3738–3743, 2005
- Chang HY, Sneddon JB, Alizadeh AA, *et al*: Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biology* 2:206–214, 2004
- Churchill GA: Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 32 (Suppl.):490–495, 2002
- Clark EA, Golub TR, Lander ES, Hynes RO: Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* 406:532–535, 2000
- Friedman N: Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805, 2004
- Golub TR, Slonim DK, Tamayo P, *et al*: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537, 1999
- Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409:533–538, 2001
- Lamb J, Ramaswamy S, Ford HL, *et al*: A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell* 114:323–334, 2003
- Mootha VK, Handschin C, Arlow D, *et al*: Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc Natl Acad Sci USA* 101:6570–6575, 2004
- Mootha VK, Lindgren CM, Eriksson KF, *et al*: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34:267–273, 2003
- Nomura I, Gao B, Boguniewicz M, Darst MA, Travers JB, Leung DY: Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: A gene microarray analysis. *J Allergy Clin Immunol* 112:1195–1202, 2003
- Odom DT, Zizlsperger N, Gordon DB, *et al*: Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303:1378–1381, 2004
- Oestreicher JL, Walters IB, Kikuchi T, *et al*: Molecular classification of psoriasis disease-associated genes through pharmacogenomic expression profiling. *Pharmacogenomics J* 1:272–287, 2001
- Petersen KF, Befroy D, Dufour S, *et al*: Mitochondrial dysfunction in the elderly: Possible role in insulin resistance. *Science* 300:1140–1142, 2003
- Ren B, Robert F, Wyrick JJ, *et al*: Genome-wide location and function of DNA binding proteins. *Science* 290:2306–2309, 2000
- Saito E, Fujimoto M, Hasegawa M, *et al*: CD19-dependent B lymphocyte signaling thresholds influence skin fibrosis and autoimmunity in the tight-skin mouse. *J Clin Invest* 109:1453–1462, 2002
- Segal E, Friedman N, Koller D, Regev A: A module map showing conditional activity of expression modules in cancer. *Nat Genet* 36:1090–1098, 2004
- Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34:166–176, 2003
- Storz MN, van de Rijn M, Kim YH, Mraz-Gernhard S, Hoppe RT, Kohler S: Gene expression profiles of cutaneous B cell lymphoma. *J Invest Dermatol* 120:865–870, 2003

- Stuart JM, Segal E, Koller D, Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249–255, 2003
- Tracey L, Villuendas R, Dotor AM, *et al*: Mycosis fungoides shows concurrent deregulation of multiple genes involved in the TNF signaling pathway: An expression profile study. *Blood* 102:1042–1050, 2003
- Whitfield ML, Finlay DR, Murray JI, *et al*: Systemic and cell type-specific gene expression patterns in scleroderma skin. *Proc Natl Acad Sci USA* 100:12319–12324, 2003
- Zahnow CA, Cardiff RD, Laucirica R, Medina D, Rosen JM: A role for CCAAT/enhancer binding protein beta-liver-enriched inhibitory protein in mammary epithelial cell proliferation. *Cancer Res* 61:261–269, 2001
- Zhou X, Krueger JG, Kao MC, *et al*: Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiol Genomics* 13:69–78, 2003
- Zhou X, Tan FK, Xiong M, *et al*: Systemic sclerosis (scleroderma): Specific autoantigen genes are selectively overexpressed in scleroderma fibroblasts. *J Immunol* 167:7126–7133, 2001