Molecular Cell
Article

A Genomic Strategy to Elucidate Modules of Oncogenic Pathway Signaling Networks

Jeffrey T. Chang,[1] Carlos Carvalho,[4] Seiichi Mori,[1] Andrea H. Bild,[5] Michael L. Gatza,[1] Quanli Wang,[1] Joseph E. Lucas,[1] Anil Potti,[1] Phillip G. Febbo,[1] Mike West,[3] and Joseph R. Nevins[1,2,*]
[1]Institute for Genome Sciences and Policy, Duke University Medical Center
[2]Department of Molecular Genetics and Microbiology, Duke University Medical Center
[3]Department of Statistical Science
Duke University, Durham, NC 27708, USA
[4]Department of Econometrics and Statistics, Graduate School of Business, University of Chicago, Chicago, IL 60637, USA
[5]Department of Pharmacology and Toxicology, University of Utah, Salt Lake City, UT 84112, USA
*Correspondence: j.nevins@duke.edu
DOI 10.1016/j.molcel.2009.02.030

SUMMARY

Recent studies have emphasized the importance of pathway-specific interpretations for understanding the functional relevance of gene alterations in human cancers. Although signaling activities are often conceptualized as linear events, in reality, they reflect the activity of complex functional networks assembled from modules that each respond to input signals. To acquire a deeper understanding of this network structure, we developed an approach to deconstruct pathways into modules represented by gene expression signatures. Our studies confirm that they represent units of underlying biological activity linked to known biochemical pathway structures. Importantly, we show that these signaling modules provide tools to dissect the complexity of oncogenic states that define disease outcomes as well as response to pathway-specific therapeutics. We propose that this model of pathway structure constitutes a framework to study the processes by which information propogates through cellular networks and to elucidate the relationships of fundamental modules to cellular and clinical phenotypes.

INTRODUCTION

The phenotypic heterogeneity of human cancers presents major challenges to advancing our understanding of disease mechanisms as well as to developing effective strategies for therapeutic design. This heterogeneity is also reflected at a molecular level in the variations in activity of cell-signaling pathways that control cell growth and determine cell fate, processes critical for driving the cancer phenotype. Recent studies describing in-depth analyses of gene mutations in a number of human cancers have emphasized the importance of placing such data in pathway-specific contexts (Ding et al., 2008; Jones et al., 2008; Network, 2008; Parsons et al., 2008; Wood et al., 2007). Certain biological processes do represent relatively simple series of biochemical events linked in an orderly fashion, such as the known biochemical pathways associated with energy metabolism. However, the extension of this notion of a linear pathway is neither useful nor appropriate as a description of the events associated with complex cellular responses to environmental inputs such as growth stimulation. Rather, the signaling events represent activities in complex networks of multiple signaling modules that each respond to given inputs (Segal et al., 2004). A module is the unit of signaling activity. One example is phosphatidylinositol 3-kinase (PI3K) phosphorylating Akt to activate its kinase activity; another is cyclin D/Cdk4 phosphorylating Rb to eliminate its negative control of E2F. These modules are defined by the biochemical events that they mediate. They are assembled into pathways by virtue of the nature of the signaling processes, but this is fluid, variable, and context dependent. For instance, PI3K can be activated by Ras, but PI3K can also be activated by a variety of other signaling events, so the PI3K module is part of the Ras pathway in one setting but part of another pathway in a different setting. Ultimately, the complex assemblage of these signaling modules constitutes the signaling network that is activated in response to a particular input under a defined set of conditions.

The Ras-signaling network, frequently altered in human cancers, exemplifies modular structure. Ras controls numerous processes related to cell proliferation and fate through interactions with secondary effectors (Shaw and Cantley, 2006). Mutations in Ras can alter its ability to interact with specific effectors, decoupling the downstream activities into discrete modules that contribute complementary activities critical to the initiation and maintenance of tumors (Lim and Counter, 2005; White et al., 1995). Of nearly a dozen effectors identified, the Raf kinase, RalGEF, and PI3K modules are studied most thoroughly (Mitin et al., 2005). Because particular modules are connected to specific characteristics of the tumor phenotype, having an unbiased catalog of the modules that comprise pathways, as well as the means to measure them, will prove valuable in efforts to pinpoint the precise modules that drive a tumor phenotype.

Thus, it is critical to develop methods to assay the activity of individual signaling modules as the basic units of signaling activity. Though measures of protein phosphorylation could be an approach, this is limited by the availability of reagents to carry

out the assay (usually antibodies), the sensitivity of the measurements, and the capacity to do this on a scale sufficient to eventually reconstruct the signaling network. Gene expression data represent one form of data that is an accessible, useful source for these measurements. Ultimately, cell-signaling events lead to changes in gene expression, and thus, regardless of whether or not the module directly involves transcriptional activity, the eventual result of the signaling process will be a change in gene expression. Further, whole-genome measures of gene expression from DNA microarray analysis provide the complexity of data that can discern the subtle distinctions in signaling events.

Genome-scale expression data have a proven ability to characterize the complex biological diversity in tumors or cells lines (Bild et al., 2006b; Segal et al., 2004). Multiple studies have shown that the activity of a pathway, such as amplification of MYC or mutation in RAF, leads to distinctive patterns in the expression of genes—the expression signatures of the pathways (Adler et al., 2006; Solit et al., 2006). Even pathways that operate primarily through posttranslational mechanisms such as phosphorylation cascades leave recognizable gene expression signatures (Bild et al., 2006a; Huang et al., 2003; Sweet-Cordero et al., 2005). For these pathways, the genes in the signatures reflect the downstream transcriptional consequences of protein-level regulation; whereas those genes may not coincide with the ones in the primary cascades, they nevertheless provide measures of upstream pathway activity. This suggests that the complexity of pathway machinery is reflected in the complexity of the expression data; we then need analysis methods to deconvolute this complexity and identify contributions of fundamental pathway modules.

To address this central question of deciphering pathway complexity, we have developed an approach to deconstruct pathways into underlying modules based on structure observed in gene expression profiles (Bild et al., 2006a; Lamb et al., 2006). Our approach builds on statistical factor analysis methods (Brunet et al., 2004; Carvalho et al., 2008; Lucas et al., 2006; Seo et al., 2007). By centering the analysis on the genes in a pathway, this analysis produces a set of pathway-related signatures that we hypothesize represent the activities of the modules of the pathway. To exemplify and test the approach, we deconstruct the Ras-signaling and E2F transcriptional regulatory pathways to reveal a series of module signatures that can predict drug sensitivity and dissect clinical outcomes in practically meaningful ways. This generates a deeper understanding of the complexity of pathway function by elucidating the modules reflected in natural variability of genomic expression structure. The analysis also leads to opportunities for therapeutic advances through the identification and characterization of clinically relevant pathway modules that may now be more specifically targeted with drugs.

RESULTS

Methodology to Deconstruct Pathway Structure

To identify gene expression signatures that represent the activity of pathway modules, we first define an initial set of genes on which to focus the pathway analysis. Because Ras function is mediated through protein interactions, we define the Ras pathway to be the proteins that bind to Ras either directly or with one degree of separation in a protein-protein interaction network (Table S1 available online) (Rual et al., 2005). Then, we apply a strategy based on statistical factor analysis with the Bayesian Factor and Regression Modeling (BFRM) tools (Carvalho et al., 2008; Lucas et al., 2006; Wang et al., 2007). Statistical factor modeling deconvolutes a gene expression data set into a series of underlying signatures with a model of the form $X = A\Lambda + \Psi$, wherein X is an $n \times m$ matrix of the gene expression data (n and m are the numbers of genes and samples in the data set, respectively), A is a sparsely defined $n \times k$ matrix indicating the genes in the signatures (k is the number of signatures) and defining weights between gene signature pairs, Λ is a $k \times m$ matrix of the scores of the signatures across the data set, and Ψ reflects measurement error and residual biological noise in the data. The number of signatures k is estimated statistically. Thus, the analysis can identify underlying components of variation in expression that relate to multiple, intersecting sets of genes, whose signatures reflect subtle, modular aspects of expression variation related to the network under study.

In the context of pathway analysis, BFRM is applied to an initial, selected set of genes identified as core for the pathway. To aid in exploring the structure of an incompletely defined pathway, analysis then iterates through a two-step cycle in which the factor model decomposition is supplemented with an evolutionary search to identify additional genes that, in terms of expression variation, relate to the estimated factors; these gene are candidate contributors to pathway structure (Carvalho et al., 2008; Wang et al., 2007). Hence, the statistical analysis allows for iterative expansion of the initial set of genes to enrich the core pathway gene list; this provides a step toward improved pathway understanding because it now also dissects the contributions of closely related modules to the complex patterns of expression variation across heterogeneous cancer data.

The analysis results in a collection of estimated statistical factors that define signatures; each signature is a set of genes with estimated weights (regression coefficients from the factor analysis). Any further expression sample, whether from tumors or cell lines, can then be scored for the level of activity of a signature by taking the weighted average of the expression levels of its genes. Cells or tumors with high scores for a given signature share similar activation levels, and those with low scores share the opposite levels; the magnitude of the scores differentiate levels of biological activity linked to the pathway module represented by the signature. For example, high scores coincide with high levels of Ras pathway activity measured on each of a number of factor scores related to multiple modules of the Ras pathway.

Here, we use the NCI-60 cancer cell line data set as the source of expression data because this diverse collection of cancer cell lines exhibits widely varying activity in the Ras pathway (Ross et al., 2000).

Dissecting the Ras Pathway into Modules

By using the strategy outlined in Figure 1, we generate a collection of 20 signatures derived from the Ras core pathway (Figure 2A and Table S2). For comparison, we also show the

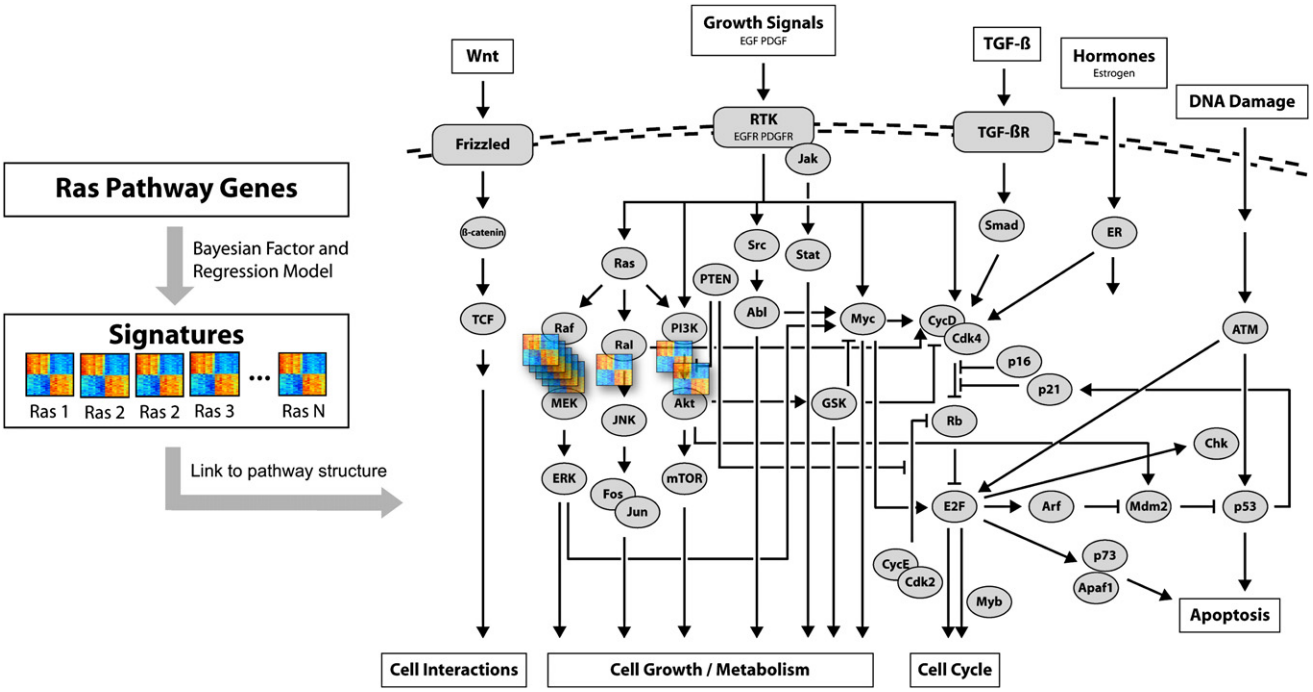


Figure 1. Pathway Module Gene Expression Signatures

(Right) The schematic depicts an interconnected signaling network organized as a collection of modules that define pathway structure. Perturbations of key genes result in changes to the transcriptional profile of the cell. Expression signatures developed from the pathway module analyses represent activity of units of this structure.

(Left) To elucidate pathway modules, we decompose a pathway based on common patterns in the expression of the genes in the pathway. First, we identify a set of core pathway genes and a set of samples with microarray gene expression data evidencing a range of variation in activity in the pathway. We next apply the BFRM software to decompose the pathway into a series of signatures. As part of the modeling process, BFRM expands the pathway genes into a set that more comprehensively describes pathway structure, ultimately yielding a collection of estimated signatures. Finally, we link the signatures with known modules of the pathway. We reason that, if a signature is significantly related to activity of a particular module of the pathway, then that signature is a representative of that pathway module.

Ras pathway activity predicted for each cell line from the expression signature defined in Bild et al. (2006a). We reason that, if these signatures represent units of Ras-related gene expression, a subset of these signatures should correlate with activities of known effectors. We explore this with two distinct indicators of pathway activation measured on the NCI-60 cell line panel: the presence of mutations in the Ras pathway genes and measures of the sensitivity of the cells to drugs that target specific Ras pathway modules (Table S3). We reason that the activation of a particular module of the Ras pathway will create sensitivity to a drug that targets activities within that module.

As shown in Figure 2B, the scores for Ras signature 14 show a significant association with mutations in Raf, providing evidence that this signature quantifies activity in the Raf module of the Ras pathway. Furthermore, this signature is also strongly related to sensitivity to a drug that inhibits ERK; this distinguishes this signature as being related to signaling down the Raf-MEK-ERK module. A similar analysis finds that Ras signature 11 denotes activity in PI3K-Akt signaling. To quantify the statistical significance of these findings, we calculate the probability of obtaining these signatures by chance (see Experimental Procedures) and find the p value of Akt (signature 11) to be 0.00005 and that of Raf (signature 14) to be 0.000006. Obtaining the two Ras pathway signatures by chance in an analysis with 20 signatures is also exceedingly unlikely ($p < 0.00001$). These results show that the approach can identify precise signaling activities through specific downstream pathway modules.

Validation of Ras Pathway Modules

To validate the capacity of the Ras pathway module signatures to predict pathway activity, we have taken advantage of the identification of Ras mutants that selectively activate the downstream effectors Raf, Ral, and PI3K (Lim and Counter, 2005). We generated RNA from cells expressing the mutant proteins and evaluated the gene expression data with each of the 20 previously derived Ras module signatures (GEO accession number GSE14934). As shown in Figure 3A, Ras signature 14, which was linked to the Raf effector arm based on the analysis in the NCI-60 data set, also identified the cells expressing the Ras mutant activating Raf signaling and distinguished them from the cells in which the other two Ras pathway effectors were activated. Conversely, Ras signature 11 which was previously linked to the PI3K arm also identified the cells expressing the Ras activating PI3K effector pathway and distinguished these from the other mutant cells. These findings provide a strong, independent validation of the capacity of these module

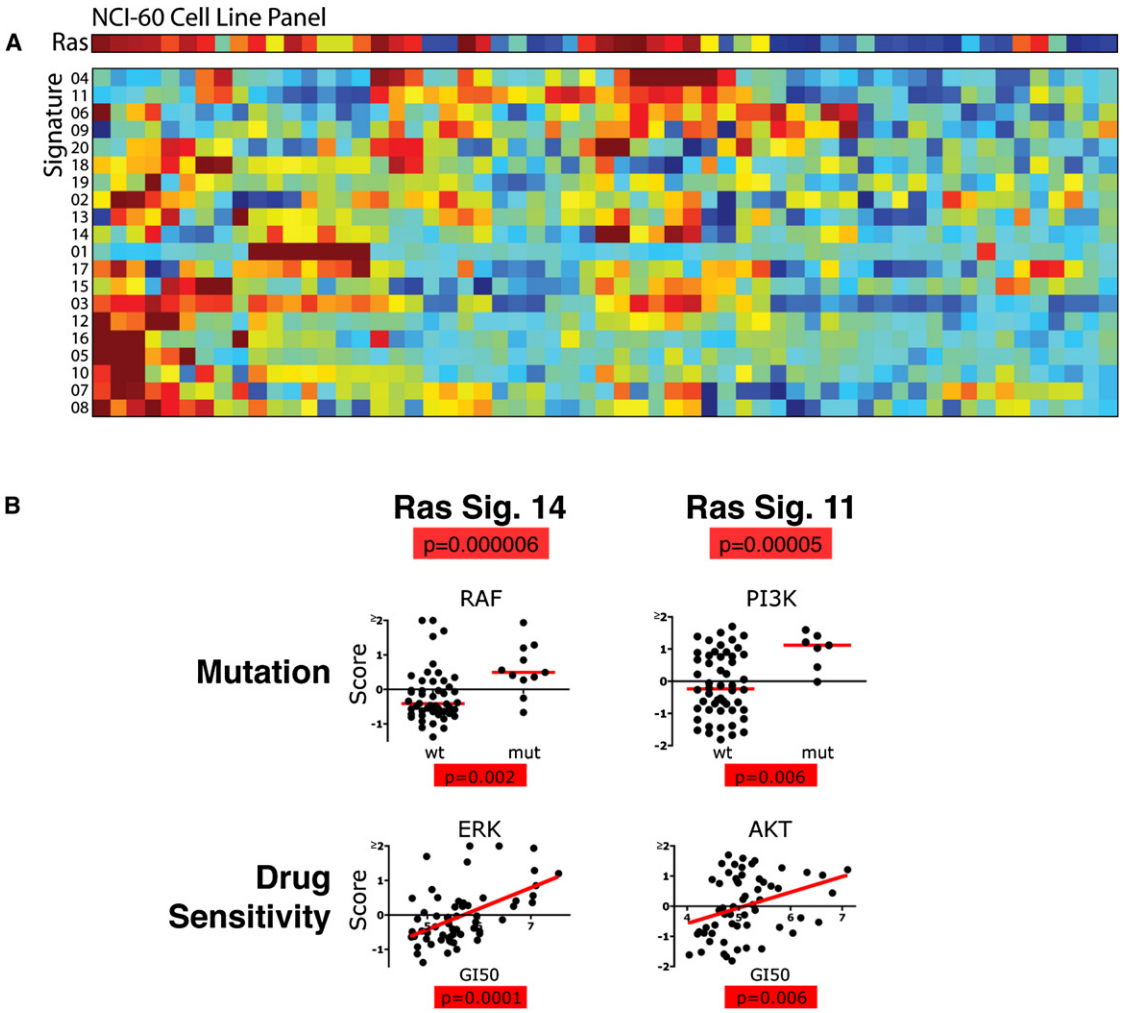


Figure 2. Linking Signatures to Modules of the Ras Pathway

(A) The top row shows the Ras pathway activity as predicted from a Ras signature across the NCI-60 cell lines. On the large heat map, each row shows the profile of the scores for a single signature derived from dissecting the Ras pathway into underlying components of variation. Cell lines with similar scores for a signature (e.g., all deep red or all deep blue) exhibit similar levels of activity. The signatures and cell lines are ordered based on hierarchical clustering.

(B) The scatter plots in the "Mutation" row show the relationship between signatures 14 and 11 and the presence of mutations in the NCI-60 cell lines. In each plot, the signature scores of cells with no mutations are shown on the left, and those with mutations are shown on the right. The scores for signatures 14 and 11 discriminate for mutations in Raf and PI3K, respectively.

The plots in the "Drug Sensitivity" row link signatures 14 and 11 with sensitivity to compounds that inhibit ERK (hypothemycin) and Akt (triciribine), respectively. (Database identifiers for the compounds are provided in Table S3.) The GI50, the negative log concentration of the drug required to inhibit growth by 50%, is plotted against the signature scores. Higher GI50 numbers indicate that the cells are sensitive to the compound and show activation of signaling through the targeted pathway. Consistent with the mutation data, higher scores for signatures 14 and 11 are associated with cells sensitive to ERK and Akt inhibitors, respectively.

signatures to accurately identify cells expressing the relevant Ras effector pathway.

Although the analysis derived from the NCI-60 data set does not provide the opportunity to link a signature with the Ral effector arm given the lack of relevant drug sensitivity or mutation data, we did identify a Ras module signature that shows a link with this effector. Signature 17 is correlated with Ral and distinguishes these cells from those in which the other two Ras effector pathways have been activated (Figure 3B).

To further verify that these signatures recognize activation of Ras pathway modules, we predict that they will be able to distinguish cells in which the relevant module is activated. To assess this, we evaluated gene expression data sampled from prostates of transgenic mice expressing activated Akt (GSE1413). As shown in Figure 3C, the Akt signature (signature 11) accurately discriminates the Akt+ samples from the controls. In contrast, the Raf signature (signature 14) does not discriminate these two samples. Conversely, by using expression data from a breast cancer cell line expressing Raf or its downstream effector MEK (GSE3542), the Raf signature (signature 14) discriminates against the controls, whereas the Akt subsignature does not (Figure 3D). Finally, in a more heterogeneous data set of 90

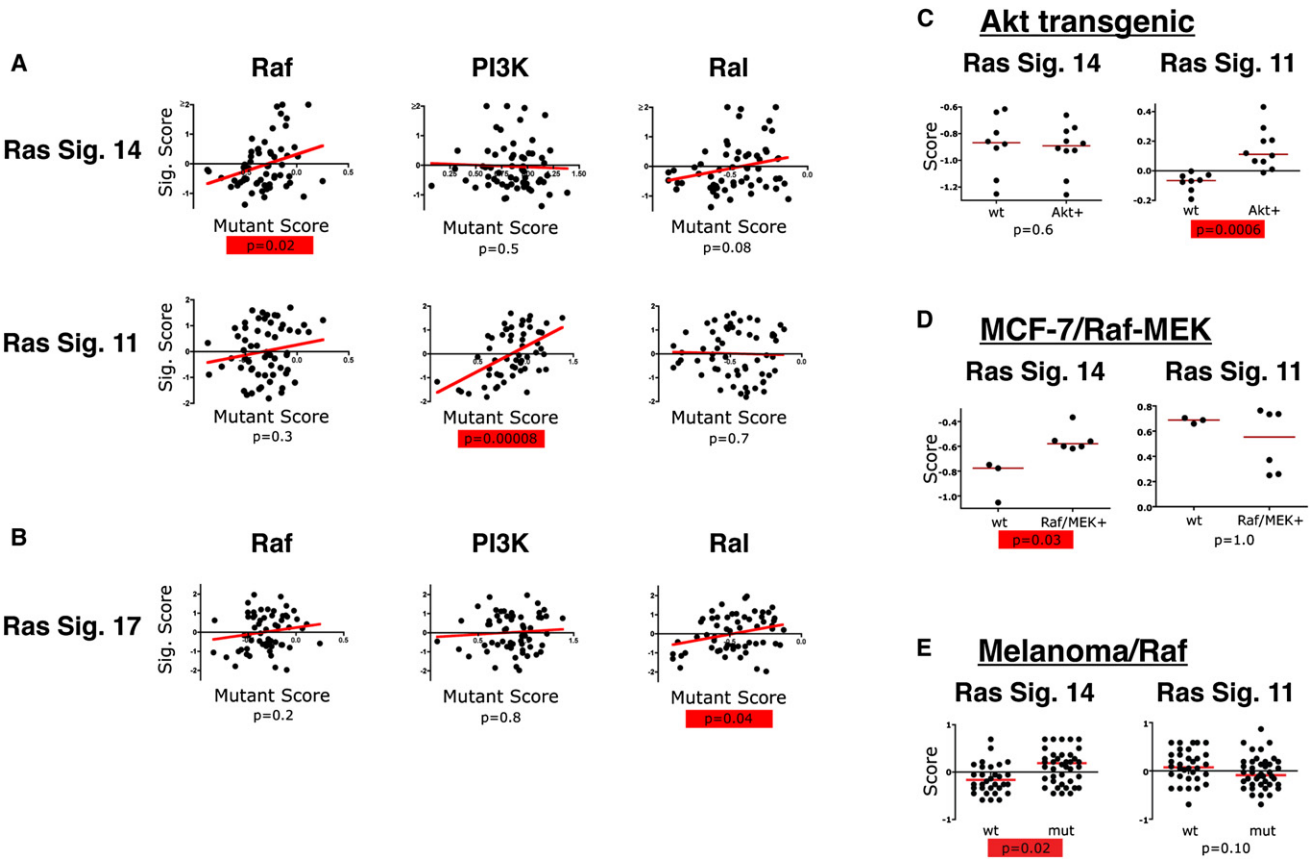


Figure 3. Linking Signatures with Ras Pathway Modules

(A) These plots show the relationships between the predicted activity of the Raf-, PI3K-, and Ral-activating Ras mutants (x axes) and Ras signatures 14 and 11 (y axes) across the NCI-60 cell lines. The scores for signature 14 share a significant relationship with the predicted activity of the Raf-activating mutant across these cells, whereas those for signature 11 are related to the PI3K-activating mutant. Red regression lines illustrate the trend between signatures and drug sensitivity. The p values are calculated from a nonparametric correlation as described in the Experimental Procedures.

(B) These plots are analogous to those in (A) and show that signature 17 is correlated with the predicted activity for Ral.

(C) Ras signatures 14 and 11 are projected onto a data set profiling the prostates of transgenic mice expressing Akt (GSE1413). Signature 11 discriminates the Akt-driven prostates from the wild-type controls, and signature 14 does not.

(D) Signature 14 discriminates MCF-7 cells expressing either Raf or MEK from wild-type cells, and signature 11 does not (GSE3542).

(E) In a data set of 90 melanomas, signature 14, but not signature 11, predicts the samples with Raf mutations (GSE4845).

melanomas that is sequenced for Raf mutations (GSE4845), the score of the Raf, but not the Akt, signature is linked to Raf mutations (Figure 3E) (Hoek et al., 2006).

Further analysis demonstrates that the Ras signatures can also be derived from an independent Ras pathway based on a Ras overexpression microarray experiment rather than from genes with known protein interactions (Table S1 and Figure S1). Furthermore, we have verified through simulations that the signatures cannot be derived from randomly selected initial gene sets (data not shown). Hence, though the approach is specific for the pathway being analyzed, it is not sensitive to a specific definition of the pathway genes used to initialize the analysis.

Taken together, these data provide strong evidence that pathway signatures can be identified with this approach, that they are specific to the pathway module being measured, and that they are robust in their capacity to predict the activation of the pathway related to that signaling module.

Identification of Pathway Module Signatures of the E2F Pathway

The Rb/E2F network provides a second context and several examples of the utility of the approach. Rb regulates the activity of the family of E2F transcription factors that, in turn, control expression of genes critical for the $G_1 \rightarrow S$ and $G_2 \rightarrow M$ transitions (Hernando et al., 2004; Ishida et al., 2001; Muller et al., 2001; Zhu et al., 2004) (Figure 4A). This dichotomy of E2F function provides an opportunity to explore the extent to which signature analysis can reveal pathway module signatures linked to these distinct roles of E2F proteins.

By using the same strategy as in the Ras investigation, we deconstruct the E2F pathway with BFRM analysis applied to the NCI-60 data; this identifies eight signatures (Tables S4–S5). Of these, one is significantly associated with S phase in the cell cycle, and two are significantly associated with mitosis based on their association to drugs that affect either S phase

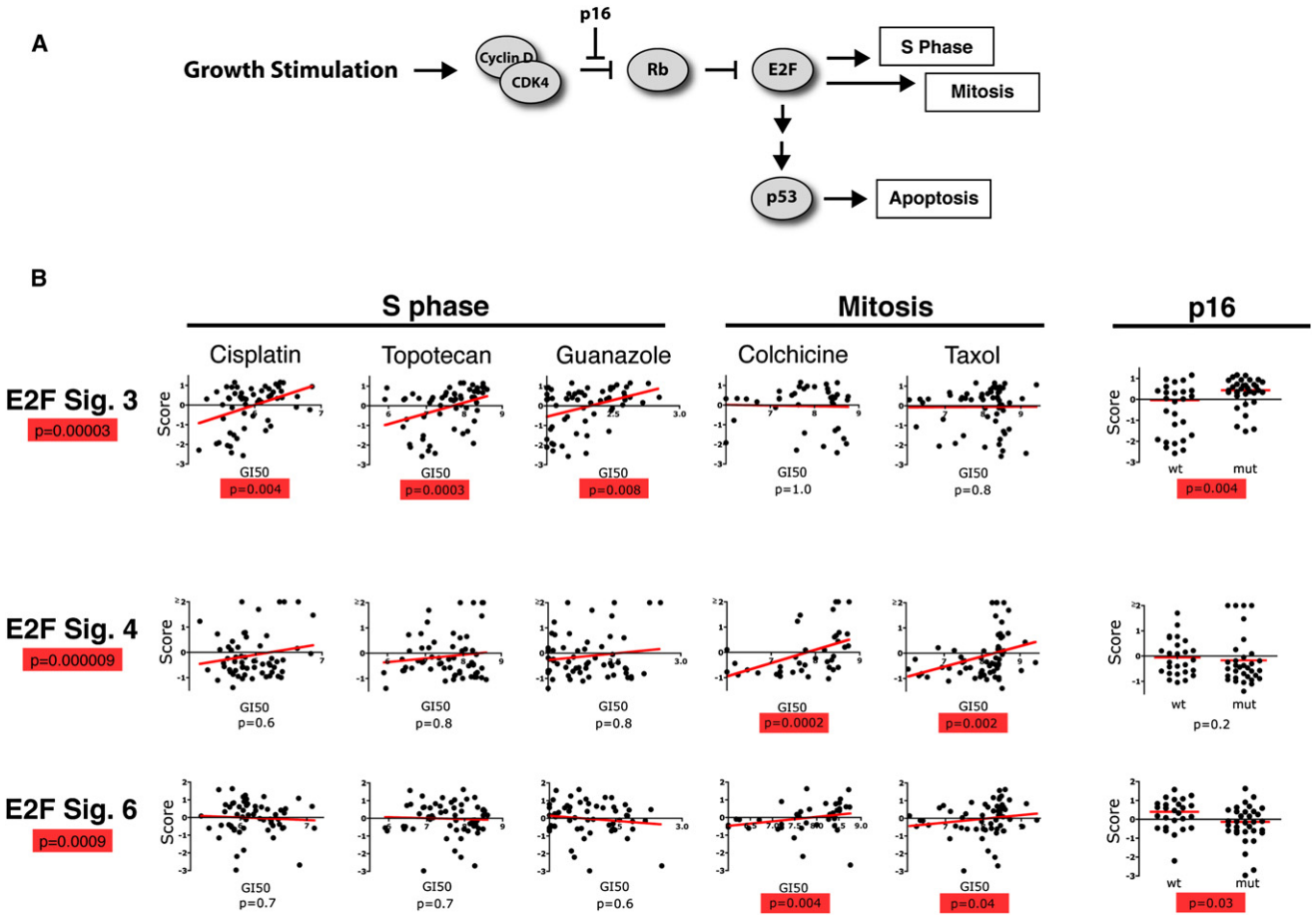


Figure 4. Modules of the Rb/E2F Pathway

(A) The Rb/E2F pathway regulates genes and processes in S phase and mitosis.

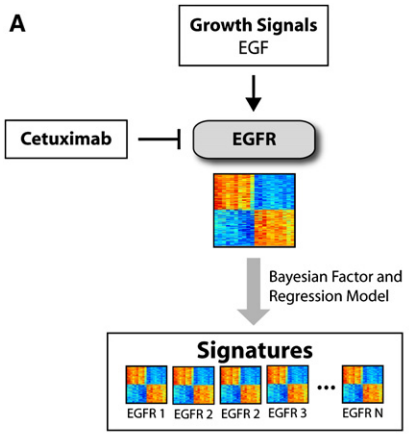
(B) Three signatures derived from the Rb/E2F pathway are significantly associated with specific phases of the cell cycle. In this figure, the signatures are produced from the E2F pathway. The scatter plots show the relationship between the scores of signatures 3, 4, and 6 and drugs that target events in S phase and mitosis of the cell cycle. The GI50 and red regression lines are defined as in Figure 2B. Here, signature 3 is significantly associated with three drugs that target aspects of S phase (Table S3) and is not associated with drugs that target mitosis. Furthermore, high scores for signature 3 are associated with mutations in p16, an inhibitor of S phase activities. Conversely, signatures 4 and 6 are associated with mitosis, but not S phase.

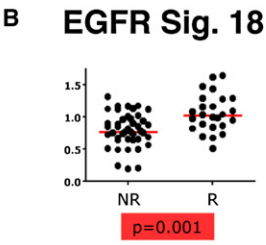
or mitotic events (Figure 4B). Cell lines with high scores on E2F signature 3 are correlated with sensitivity to three drugs that target S phase with distinct mechanisms of action (Koster et al., 2007; Weinstein et al., 1992). High scores for this signature are also correlated with mutations in p16 (CDKN2A), a component of the $G_1 \rightarrow S$ checkpoint.

Next, we find that cell lines with high scores on E2F signatures 4 and 6 are sensitive to drugs that target the mitotic spindle (Weinstein et al., 1992). As expected, the converse is not true; the S phase signature 3 is not associated with sensitivity to mitotic drugs, and the mitotic signatures 4 and 6 are not associated with sensitivity to S phase drugs. By using a similar approach as above, the p value for the S phase signature, signature 3, is 0.00003, and those for the mitosis signatures (signatures 4 and 6) are 0.000009 and 0.0009, respectively; the p value for the entire analysis is $p < 0.00001$. Thus, modular deconstruction of the E2F pathway identifies specific processes known to be related to the pathway. This further supports the value of the strategy, as exhibited in the Ras analysis, and the view that the decomposition approach is general and can be extended to pathways with divergent mechanisms of action.

Clinical Utility of Pathway Module Signatures

Ultimately, the utility of the pathway module signatures lies in the capacity to better understand and dissect the complexities of signaling events underlying clinically relevant phenotypes. To explore and exemplify this, we have analyzed pathway expression modules in relation to the response of colon cancer patients to the epidermal growth factor receptor (EGFR)-specific therapeutic cetuximab (Khambata-Ford et al., 2007; GEO accession number GSE5851) (Figure 5A). The activation status of EGFR, including the use of an EGFR pathway signature, is simply incapable of discriminating responses to cetuximab. Thus, it is of interest to ask whether discrimination can be obtained from a refined understanding of the EGFR network in terms of pathway module signatures. Following the same approach





C

EGFR		Ras		E2F	
Module	*p* value	Module	*p* value	Module	*p* value
18	0.001	11	0.049	2	0.052
1	0.062	3	0.052	5	0.124
4	0.079	2	0.062	7	0.147
9	0.104	19	0.093	4	0.174
13	0.115	18	0.133	6	0.208
2	0.127	13	0.162	3	0.360
11	0.166	20	0.162	1	0.461
5	0.213	1	0.182	8	0.500
3	0.237	7	0.222		
12	0.334	14	0.222		
10	0.347	5	0.285		
14	0.353	8	0.347		
8	0.461	4	0.423		
6	0.558	12	0.492		
19	0.584	10	0.525		
16	0.647	15	0.629		
17	0.656	6	0.770		
15	0.722	9	0.799		
7	0.741	17	0.809		
20	0.929	16	0.869		

Figure 5. Predicting Clinical Response with Pathway Module Signatures

(A) Cetuximab, a monoclonal antibody, targets the EGF receptor. By using a signature for EGFR activation as a starting point, we apply BFRM to decompose the pathway into 20 modules, as represented by gene expression signatures.

(B) Clinical response to cetuximab is assessed in 68 patients with advanced colorectal cancer (CRC). In these patients, we predict the activity of the 20 EGFR modules with the pathway module signatures. The plot shows that EGFR module 18 can significantly distinguish patients with response (R) and no response (NR) to the EGFR inhibitor cetuximab.

(C) Assessment of each of the EGFR, Ras, and E2F pathway module signatures for prediction of cetuximab response. The table shows the ability of each module to identify patients that respond to the EGFR inhibitor cetuximab. The p value is calculated with a nonparametric Wilcoxon test that compares the predicted module scores between responders and nonresponders.

used for the generation of the Ras- and E2F-signaling modules, we created a set of 20 signatures derived from an initial EGFR signature (Tables S6–S7). We then evaluated the collection of pathway module signatures, derived from EGFR, Ras, and E2F, for their capacity to distinguish response to cetuximab. Evaluation of the EGFR modules revealed one (signature 18) that significantly distinguished cetuximab responders from nonresponders (Figure 5B). In contrast, neither the Ras nor the E2F pathway module signatures could effectively distinguish response to cetuximab (Figure 5C).

As a second example of the use of pathway module signatures in a clinical context, we focus on past work that has shown a capacity of pathway signatures to dissect the complex heterogeneity across tumor types (Bild et al., 2006a). An example arises from the analysis of 74 lung tumors comprised of an approximately equal number of squamous cell carcinomas and adenocarcinomas (GEO accession number GSE3141) (Figure 6A). This analysis demonstrates the power of pathway signatures to identify subgroups of tumors, including two large subgroups that are characterized by Ras pathway activation and E2F3 pathway activation. We used the sets of Ras and E2F pathway module signatures to determine whether these broad groups can be further dissected into clinically meaningful subtypes based on the activity of module signatures. As shown in Figure 6B, this analysis reveals that Ras pathway signature 9 significantly distinguishes low- and high-risk survival groups. This has been further extended to a second lung tumor data set involving 89 adenocarcinoma samples (GEO accession number GSE3593-ACOSOG) (Potti et al., 2006) (Figure S2A). The E2F-signaling modules and, in particular, E2F signature 6, which linked to the mitotic component of the E2F pathway, identify a cohort of tumors with poor survival and are again reproducible in an independent data set (GSE5828) (Figure S2B).

Taken together, these analyses demonstrate the capacity of pathway module signatures to identify properties of tumors that relate to significant clinical phenotypes. Such a model that represents the complexity of the underlying biological activities provides the ability to improve tumor classification, to reveal more precise prognosis, and also to predict response to pathway-specific drugs. This suggests that a rational strategy to target therapeutics may be improved by using an approach

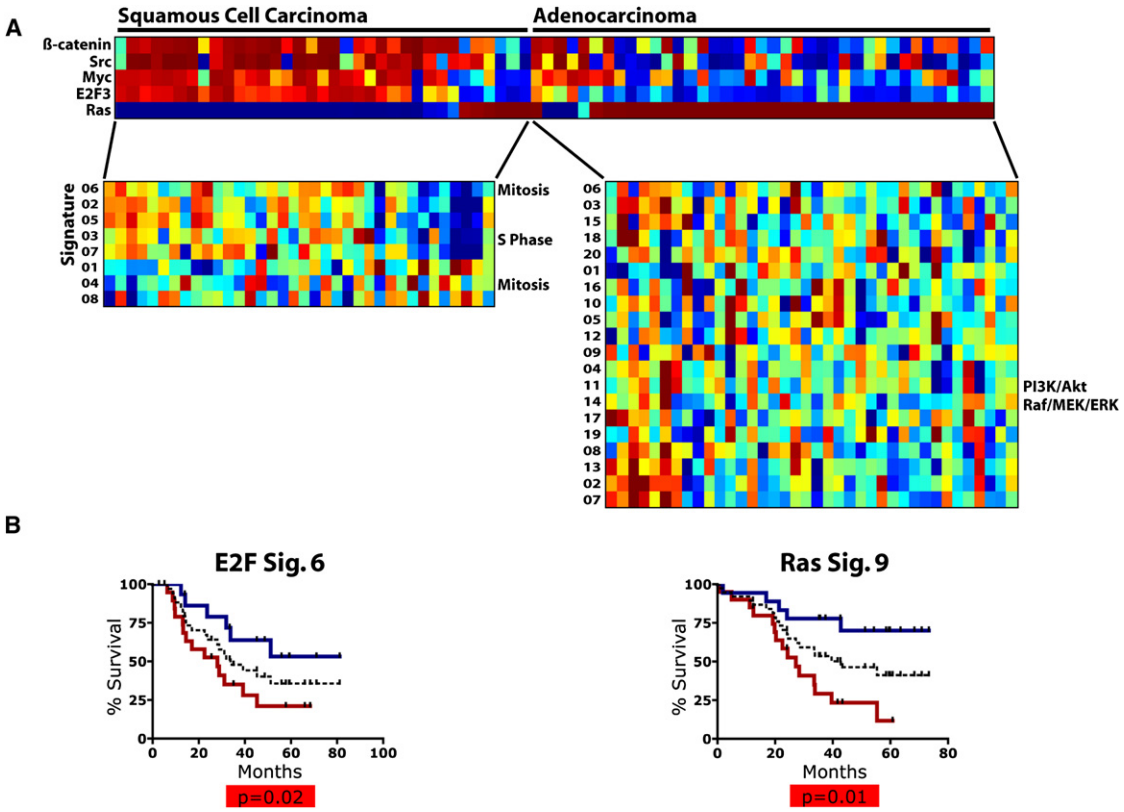


Figure 6. Activity of Pathway Module Signatures in Lung Cancer

(A) Predicted Ras and E2F3 pathway activities across a set of 74 lung tumors. Ras activity is largely absent in the squamous cell carcinomas and is present in nearly all adenocarcinomas. Conversely, E2F3 activity is high in squamous cell carcinoma but low in adenocarcinomas. The main heat map shows the scores of the 8 E2F pathway and 20 Ras pathway signatures (rows) projected onto the subset of 36 squamous cell carcinomas and 38 adenocarcinomas (columns), respectively.

(B) Kaplan-Meier survival curves for E2F signature 6 and Ras signature 9 across their respective diseases. The dotted black line shows the overall survival of all patients in the data set. The red line shows only the patients with high signature scores, and the blue line shows those with low scores.

that takes into account variations in the ways that signals are propagated through pathways.

DISCUSSION

An overarching goal of systems biology is the ability to understand the functioning of cellular signaling pathways, not as isolated units or linear sets of events but as networks of interconnected events. The importance of developing an understanding at this level is emphasized by the recent studies of human cancers detailing the complex array of mutations that arise and the importance of placing this data in a pathway-specific context (Ding et al., 2008; Jones et al., 2008; Network, 2008; Parsons et al., 2008; Wood et al., 2007). A key challenge in addressing this goal is the availability of tools that can measure the variation in activity and output of the pathways in response to diverse inputs and cellular contexts. Multiple studies have shown the capacity of gene expression data to describe such subtle characteristics of biology not achievable through other means of analysis—hence, our interest in taking such studies further to realize some of the potential to address the complexity of cell-signaling events (di Bernardo et al., 2005; Ergün et al., 2007). The real challenge is to dissect the complexity of the gene expression information such that the resulting signatures reveal the discrete modules of the cell-signaling pathways. By so doing, these signatures become tools that can provide a measure of the individual activities that foster pathway complexity.

Our initial investigations deconstruct the Ras, E2F, and EGFR pathways into collections of module signatures that describe refined, discrete, or modular aspects of pathway function. This complements the classical view of pathways as wiring diagrams by providing structure in the form of modules that are measured by discrete gene expression signatures. As clearly demonstrated here, by relating various signatures to either drug sensitivity or presence of mutations, it has been possible to link several of these to characterized modules of Ras and E2F pathway activity. A key strength of this approach is that it can uncover, via an unbiased and automated statistical analysis, signatures of pathway-related activities driven by unknown molecular mechanisms. As growing numbers of molecular activities are characterized with expression signatures, the ability to link these pathway module signatures with underlying mechanisms will increase rapidly. Nevertheless, the ability to anchor the analysis on a pathway, combined with a rigorous methodology for exploring the

surrounding functional landscape, sets each signature in an initial coarse pathway context. This proof of concept provided by our several oncogenic network examples suggests that the study of the additional pathways to provide further measures of modular activity will help to achieve the goal of deciphering cellular signaling on a genomic scale.

It is important to recognize that expression signatures also represent practical tools that can be of value in present day clinical practice, independently of their potential to contribute to improved understanding of pathway structure from a systems viewpoint. In particular, the ability to add value in a prognostic setting, as illustrated in the dissection of the squamous and adenocarcinoma samples, provides a very clear potential use of this information. Similarly, the capacity of module signatures to refine the prediction of therapeutic outcomes, such as in the example of drugs that target the EGFR-signaling pathway, is clearly also relevant. We have previously demonstrated the potential of pathway-specific signatures to guide the use of various targeted therapies in a general sense. The work here takes this an important step further by making use of signatures that quantify the activity of more specific pathway modules for which drugs have been developed to facilitate the development of strategies that can accurately identify those patients most likely to benefit from a given drug. The increased biological resolution and specificity of module signatures becomes critical when the complexity of signaling pathway alterations, resulting from the complex array of mutations and genome alterations in human tumors, otherwise obscures the ability of simple pathway analysis to accurately predict response.

The pathway module strategy offers an ability to unravel complex pathway structures and identify functional modules whose activities may be connected to molecular processes that mediate sensitivity to targeted therapeutics. Within this framework, it is possible to explore the relationship between pathway function and subtle perturbations in the complex array of inputs; to investigate how this is influenced by the action of other signaling pathways and networks; and ultimately, to provide more precise connections between molecular activities and their manifestations in disease.

EXPERIMENTAL PROCEDURES

NCI-60 Data

The Ras and E2F pathway analyses used the gene expression data on the 59 NCI-60 cancer cell lines available on the NCI Developmental Therapeutics Program Web site (Scherf et al., 2000). These cell lines represented nine tissues and included data on their sensitivities to almost 45,000 in terms of GI50 numbers (Shoemaker, 2006), as well as mutational status on key cancer genes (Ikediobi et al., 2006). The gene expression data were generated by Novartis on Affymetrix U95A microarrays in triplicate and were averaged. To select the most important genes, we discarded probe sets that exhibited very low levels and variance of expression (Table S8). We predicted the Ras and E2F pathway activity on the NCI-60 and lung cancer data sets by using procedures described (Bild et al., 2006a).

Pathway Decomposition

To deconstruct the Ras, E2F, and EGFR pathways into modules, we applied the evolutionary statistical factor analysis to the NCI-60 gene expression data set. We centered the analyses on sets of genes known to represent aspects of the core pathways of interest. Reasoning that Ras signaled through phosphorylation events that depended on physical protein interactions, we included in the Ras pathway the Ki-Ras, Ha-Ras, and N-Ras isoforms and all proteins that physically interacted with these Ras proteins directly or indirectly through an intermediate protein. To do so, we used a protein network constructed by combining interactions from the BIND, BioGRID, DIP, HPRD, IntAct, MINT, and MIPS databases and the Vidal genome-wide yeast two-hybrid screen (Rual et al., 2005). This resulted in 589 genes that corresponded to 498 Affymetrix probe sets in the filtered NCI-60 data set. For the E2F pathway, because E2F regulated the transcription of genes directly, we created a gene set by combining multiple gene expression profiles of E2F function described in Chang and Nevins (2006), resulting in 224 genes matched to 216 probe sets. For the EGFR pathway, we generated a signature by comparing the expression profiles of EGF-treated human bronchial epithelial cells infected with adenoviruses expressing EGFR against those expressing a control, as described previously (Bild et al., 2006a). The sparse statistical factor analysis utilized the BFRM software that implements models and methods previously described (Wang et al., 2007); the software and examples are freely available at http://icbp.genome.duke.edu/bfrm.html. Analyses used the default parameters in all cases. We initialized models with one latent factor and iteratively included factors and genes until the analysis terminated after increasing the gene set to 1000 genes. This analysis identified 20 module signatures in the Ras pathway, 8 in E2F, and 20 in EGFR.

Ras Mutant Data

We obtained HEK-HT cells expressing mutant forms of Ras that signaled constitutively down the Raf ($Ras^{G12V,T35S}$), RalGEF ($Ras^{G12V,E37G}$), and PI3K ($Ras^{G12V,Y40C}$) branches of the pathway, as well as one with only wild-type Ras protein (Lim and Counter, 2005). We cultured the cells in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS). Then, from each cell type growing asynchronously, we collected total RNA in five independent replicates by using RNeasy Mini kits from QIAGEN. The Duke Microarray Facilities processed the RNA samples and hybridized them to Affymetrix U133A 2.0 microarrays. We normalized the raw CEL files by using the MAS5 implementation in the Bioconductor toolkit. We made the data available in GEO (accession number GSE14934). As above, we discarded the genes with the lowest levels and variance of expression (Table S8). Then, we merged the Ras mutant expression data set with the NCI-60 data set by matching probe sets based on Entrez Gene IDs. For genes that matched multiple probe sets, we resolved the ambiguity by choosing the one with the most similar correlation structure as described in Shankavaram et al. (2007). To project a Ras mutant onto the NCI-60 data set, we quantile normalized the merged data set, selected the 200 genes most correlated with Ras mutant of interest, and used an SVM with a first-degree polynomial kernel to predict its activity on the NCI-60 cell lines (Chang and Lin, 2001). We calculated the association between the Ras subsignature profiles and Ras mutant profiles with a nonparametric Kendall correlation and rejected all associations with p values > 0.05.

External Data

To project the Ras, E2F, or EGFR pathway module signatures onto data sets of the mouse prostate (GSE1413), Raf/MEK breast cancer cell lines (GSE3542), Raf mutations in melanoma (GSE4840, GSE4841, and GSE4843), lung cancer (GSE3141), lung adenocarcinoma validation (GSE3593-ACOSOG), cetuximab response (GSE5851), and lung squamous cell carcinoma validation (GSE5828), we processed the data as described above for the NCI-60 data set. Specific characteristics of the data sets and parameters used are reported in Table S8. When the CEL files were not available, we used the signal data provided in GEO. The melanoma data set was provided on three platforms that we processed separately, and we then combined the results. For the mouse data set, we converted the target probes from mouse to human by using the HomoloGene database.

Association of Module Signatures with Drug Sensitivity and Mutation Data

Drug sensitivities (quantified as the concentration resulting in 50% growth inhibition) were available for each cell line from the National Cancer Institute Developmental Therapeutics Program Web site (http://dtp.nci.nih.gov/index.

html). We calculated the association between the subsignature profiles and drug sensitivities by using a nonparametric Kendall correlation and rejected all associations with p values > 0.05. Gene mutation data for each cell line were available from the COSMIC database (Bamford et al., 2004). For each gene, we split the cell lines into two groups based on the presence of a mutation and compared, between the groups, the scores for each signature by using a Wilcoxon rank sum test, rejecting all comparisons with p values > 0.05.

To calculate the p value of each signature, we used a sampling strategy, generating using one million randomly generated signatures, calculating the association with drug sensitivities and gene mutations as above, and then computing the bootstrap significance levels for each original signature. To deal with the multiple modules in an analysis (e.g., the two Ras modules in the Ras pathway analysis), we randomly simulated the same number of signatures (20 for Ras and 8 for E2F) and evaluated whether this set of random signatures related to the same pathway modules. We repeated this 100,000 times, and the p value was the portion of sampled analyses that could also identify the pathway modules.

Analysis of Lung Adenocarcinoma Samples

We assessed the association between pathway module signatures and survival time as follows. For each signature, we split samples into two equally sized groups based on signature scores and generated Kaplan-Meier curves using the Prism software.

Analysis of Therapeutic Response

In the colorectal tumor data set, we measured the association between the profiles of the Ras, E2F, and EGFR factors and the response to cetuximab. In each data set, we first discarded the samples in which the response was not able to be determined. We split the remaining samples into two groups, wherein the nonresponders constituted the patients with progressive disease and the responders were patients with stable or regressive disease. To calculate the significance of the difference of the factor scores between the two sets of patients, we used a nonparametric Wilcoxon rank sum test.

ACCESSION NUMBERS

The microarray data set containing gene expression arrays of cells expressing Ras mutants is available in GEO under accession GSE14934.

SUPPLEMENTAL DATA

Supplemental Data include eight tables and two figures and can be found with this article online at http://www.cell.com/molecular-cell/supplemental/S1097-2765(09)00146-4.

ACKNOWLEDGMENTS

We thank Bernard Mathey-Prevot, Ashley Chi, Wencheng Zhu, and Steve Angus for helpful discussions; Chris Counter and Kevin O'Hayer for the Hek-HT cells; Marc Vidal and Wei Chen for the protein-protein interaction networks; and Shirin Khambata-Ford and David Mauro for helpful advice in the analysis of the cetuximab study. All aspects of the research were supported under the NCI Integrative Cancer Biology Program via grant NIH 5-U54-CA112952-05 and 5-RO1-CA106520-05 to J.R.N. Additional aspects related to statistical factor models were partially supported under NSF grants DMS-0102227 and 0342172. J.T.C. is supported by postdoctoral fellowship #PF-05-047-01-GMC from the American Cancer Society and NIH K99-LM009837-01. We are grateful to Kaye Culler for her assistance in the preparation of the manuscript.

Received: July 1, 2008
Revised: December 31, 2008
Accepted: February 25, 2009
Published: April 9, 2009

REFERENCES

Adler, A.S., Lin, M., Horlings, H., Nuyten, D.S., van de Vijver, M.J., and Chang, H.Y. (2006). Genetic regulators of large-scale transcriptional signatures in cancer. Nat. Genet. *38*, 421–430.

Bamford, S., Dawson, E., Forbes, S., Clements, J.B., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., and Wooster, R. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br. J. Cancer *91*, 355–358.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., et al. (2006a). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature *439*, 353–357.

Bild, A.H., Potti, A., and Nevins, J.R. (2006b). Linking oncogenic pathways with therapeutic opportunities. Nat. Rev. Cancer *6*, 735–741.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169.

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Carvalho, C.M., Chang, J., Lucas, J., Nevins, J.R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modelling: Applications in gene expression genomics. J. Am. Stat. Assoc. *103*, 1438–1456.

Chang, C.-C., and Lin, C.-J. (2001). LIBSVM: A library for support of vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chang, J.T., and Nevins, J.R. (2006). GATHER: A systems approach to interpreting genomic signatures. Bioinformatics *22*, 2926–2933.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., and Collins, J.S. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat. Biotechnol. *23*, 377–383.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M., Cibulskis, K., Sougnez, C., and Wilson, R.K. (2008). Somatic mutations effect key pathways in lung adenocarcinoma. Nature *455*, 1069–1075.

Ergün, A., Lawrence, C.A., Kohanski, M.A., Brennan, T.A., and Collins, J.J. (2007). A network biology approach to prostate cancer. Mol. Syst. Biol. *3*, 82.

Hernando, E., Nahle, Z., Juan, G., Diaz-Rodriguez, E., Alaminos, M., Hemann, M., Michel, L., Mittal, V., Gerald, W., Benezra, R., et al. (2004). Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. Nature *430*, 797–802.

Hoek, K.S., Schlegal, N.C., Brafford, P., Sucker, A., Ugurel, S., Kumar, R., Weber, B.L., Nathanson, K.L., Phillips, D.J., Herlyn, M., et al. (2006). Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signatures. Pigment Cell Res. *19*, 290–302.

Huang, E., Ishida, S., Pittman, J., Dressman, H., West, M., and Nevins, J.R. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat. Genet. *34*, 226–230.

Ikediobi, O.N., Davies, H., Bignell, G., Edkins, S., Stevens, C., O'Meara, S., Santarius, T., Avis, T., Barthorpe, S., Brackenbury, et al. (2006). Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol. Cancer Ther. *5*, 2606–2612.

Ishida, S., Huang, E., Zuzan, H., Spang, R., Leone, G., West, M., and Nevins, J.R. (2001). Role for E2F in the control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. Mol. Cell. Biol. *21*, 4684–4699.

Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Maknkoo, P., Carter, H., Kamiyama, H., Jimeno, A., et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science *321*, 1801–1806.

Khambata-Ford, S., Garrett, C.R., Meropol, N.J., Basik, M., Harbison, C.T., Wu, S., Wong, T.W., Huang, X., Takimoto, C.H., Godwin, A.K., et al. (2007). Expression of epiregulin and amphiregulin and K-ras mutation status predict

disease control in metastatic colorectal cancer patients treated with cetuximab. J. Clin. Oncol. *25*, 3230–3237.

Koster, D.A., Palle, K., Bot, E.S., Bjornsti, M.A., and Dekker, N.H. (2007). Antitumour drugs impede DNA uncoiling by topoisomerase 1. Nature *448*, 213–217.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: Using gene expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935.

Lim, K.H., and Counter, C.M. (2005). Reduction in the requirement of oncogenic Ras signaling to activation of P13K/AKT pathway during tumor maintenance. Cancer Cell *8*, 381–392.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J.R., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In Bayesian Inference for Gene Expression and Proteomics, P.M.K.A. Do and M. Vannucci, eds. (New York: Cambridge University Press).

Mitin, N., Rossman, K.L., and Der, C.J. (2005). Signaling interplay in Ras superfamily function. Curr. Biol. *15*, R563–R574.

Muller, H., Bracken, A.P., Vernell, R., Moroni, M.C., Christians, F., Grassilli, E., Prosperini, E., Vigo, E., Oliner, J.D., and Helin, K. (2001). E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. Genes Dev. *15*, 267–285.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. Science *321*, 1807–1812.

Potti, A., Mukherjee, S., Prince, R., Dressman, H.K., Bild, A., Koontz, J., Kratzke, R.A., Watson, M., Kelley, M., Ginsburg, G.S., et al. (2006). A genomic strategy to refine prognosis in non-small cell lung carcinoma. N. Engl. J. Med. *355*, 570–580.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C.A., Spellman, P., Iyer, V., Jeffrey, S.S., van de Rign, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173–1178.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat. Genet. *24*, 236–244.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. *36*, 1090–1098.

Seo, D.M., Goldschmidt-Clermont, P.J., and West, M. (2007). Of mice and men: Sparse statistical modelling in cardiovascular genomics. Annals of Applied Statistics *1*, 152–178.

Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K., Reimers, M.A., Scherf, U., Kahn, A., Dolginow, D., et al. (2007). Transcript and protein expression profiles of the NCI-60 cancer cell panel: An integromic microarray study. Mol. Cancer Ther. *6*, 820–832.

Shaw, R.J., and Cantley, L.C. (2006). Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature *441*, 424–430.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Solit, D.B., Garraway, L.A., Pratilas, C.A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J.M., She, Y., Osman, I., et al. (2006). BRAF mutation predicts sensitivity to MEK inhibition. Nature *439*, 358–362.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J.J., Ladd-Acosta, C., Mesirov, J., Golub, T.R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat. Genet. *37*, 48–55.

Wang, Q., Carvalho, C., Lucas, J.E., and West, M. (2007). BFRM: Bayesian factor regression modelling. Bull Intl Soc Bayesian Anal *14*, 4–5.

Weinstein, J.N., Kohn, K.W., Grever, M.R., Viswanadhan, V.N., Rubinstein, L.V., Monks, A.P., Scudiero, D.A., Welch, L., Koutsoukos, A.D., Chiausa, A.J., et al. (1992). Neural computing in cancer drug development: Predicting mechanism of action. Science *258*, 447–451.

White, M.A., Nicolette, C., Minden, A., Polverino, A., Van Aeist, L., Karin, M., and Wigler, M.H. (1995). Multiple Ras functions can contribute to mammalian cell transformation. Cell *80*, 533–541.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108–1113.

Zhu, W., Giangrande, P., and Nevins, J.R. (2004). E2Fs link the control of G1/S and G2/M. EMBO J. *23*, 4615–4626.