nature genetics

From signatures to models: understanding cancer using microarrays

Eran Segal, Nir Friedman, Naftali Kaminski, Aviv Regev & Daphne Koller

Genomics has the potential to revolutionize the diagnosis and management of cancer by offering an unprecedented comprehensive view of the molecular underpinnings of pathology. Computational analysis is essential to transform the masses of generated data into a mechanistic understanding of disease. Here we review current research aimed at uncovering the modular organization and function of transcriptional networks and responses in cancer. We first describe how methods that analyze biological processes in terms of higher-level modules can identify robust signatures of disease mechanisms. We then discuss methods that aim to identify the regulatory mechanisms underlying these modules and processes. Finally, we show how comparative analysis, combining human data with model organisms, can lead to more robust findings. We conclude by discussing the challenges of generalizing these methods from cells to tissues and the opportunities they offer to improve cancer diagnosis and management.

Genomics provides powerful tools with which to probe the components and behavior of biological systems. Microarrays, high-throughput chromatin immunoprecipitation[1,2] (ChIP) and tissue microarrays[3] inform us on different perspectives of the molecular mechanisms underlying cellular functions. The staggering volume of molecular data resulting from the rapid adoption of such techniques has underscored the importance of computational analysis as a key link between data generation and the formulation of new hypotheses. It is widely believed that genomics will transform our understanding of the mechanisms underlying the function of cells and organisms, and revolutionize the diagnosis and management of disease by offering an unprecedented comprehensive view of the molecular underpinnings of pathology[4,5]. Gene-expression profiling has been applied extensively in cancer research. Gene-expression microarrays have been analyzed using clustering algorithms that group genes and samples on the basis of expression profiles, and statistical methods that score genes on the basis of their relevance to various clinical attributes (**Supplementary Note** online). Using these methods, investigators have identified new classes of hematological malignancies, predicted prognosis in lung cancer and breast cancer and made many mechanistic observations (**Supplementary Fig. 1** online). Despite the natural caution associated with the implementation of new technologies in the clinical arena, the utility of the results of microarray analysis as an effective diagnostic tool at the point of care is already being assessed[6].

Approaches such as clustering and identification of gene signatures, though successful, tend to ignore much of the signal in the data, both in genes whose activity changes but does not pass the threshold for differential expression and in genes that are differentially expressed but unfamiliar to the researcher analyzing the list. Furthermore, because these analyses are done at the gene level, they are prone to the inherent noise that exists both in the sample population and in different stages of assaying gene expression. Moreover, simply listing genes associated with a certain tumor type is far from identifying the biological processes in which these genes are involved. Finally, clustering genes with similar expression patterns does not identify the causal molecular mechanisms that regulate them. Therefore, developing analysis methods that can extract a more biologically meaningful understanding of the processes giving rise to cancer is a key challenge. Here, we focus on ongoing research that attempts to achieve this goal, discuss challenges in its application to complex multicellular tissues and conclude with some opportunities for using these methods to improve cancer diagnosis and treatment.

A module-level view

To transcend from individual genes to biological processes, several recent methods[7–10] use gene modules as the basic building blocks for analysis. These methods aim to distill a higher-order and more interpretable characterization of transcriptional changes. Moreover, by considering coherent changes in expression in larger modules, we can identify patterns that are too subtle to discern when considering expression profiles of individual genes in isolation.

Mootha *et al.*[8] (**Fig. 1a**) tested biologically coherent sets of genes (*e.g.*, pathways) for association with disease phenotypes. They applied their method to a data set of human diabetic muscle, with the goal of identifying processes that were systematically altered in diabetic muscle. Their analysis showed that, by examining the joint behavior of a set of genes, they could detect significant changes even in cases where the expression of individual genes was not significantly different. It was only in the coherent signal associated with a higher-level entity that the pattern was evident.

Eran Segal is at the Center for Studies in Physics and Biology, Rockefeller University, New York, USA. Nir Friedman is in the School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel. Naftali Kaminski is at the Dorothy P. and Richard P. Simmons for Interstitial Lung Diseases, Pulmonary Allergy and Critical Care Medicine, University of Pittsburgh, USA. Aviv Regev is at the Bauer Center for Genomics Research, Harvard University, Cambridge, Massachusetts, USA. Daphne Koller is in the Computer Science Department, Stanford University, Stanford, California, USA. e-mail: koller@cs.stanford.edu

Published online 26 May 2005; doi:10.1038/ng1561

Segal *et al.*[7] applied a module-level analysis to obtain a global view of the shared and unique molecular modules underlying human cancer. They compiled a 'cancer compendium' from multiple studies and a large collection of biologically meaningful gene sets from experimental studies and human-curated annotations. They identified gene sets with similar behavior across arrays, combined them into modules and used these modules to characterize a variety of clinical conditions (*e.g.*, tumor stage and type) by the combination of activated and deactivated modules. In the resulting 'cancer module map'[7] (**Fig. 1b**), the activation or repression of some modules (*e.g.*, cell cycle) was shared across multiple tumor types and could be related to general tumorigenic processes, whereas others (*e.g.*, growth-regulatory modules; **Fig. 1c**) were more specific to the tissue origin or progression of particular tumors. Conversely, the module map characterized each condition by

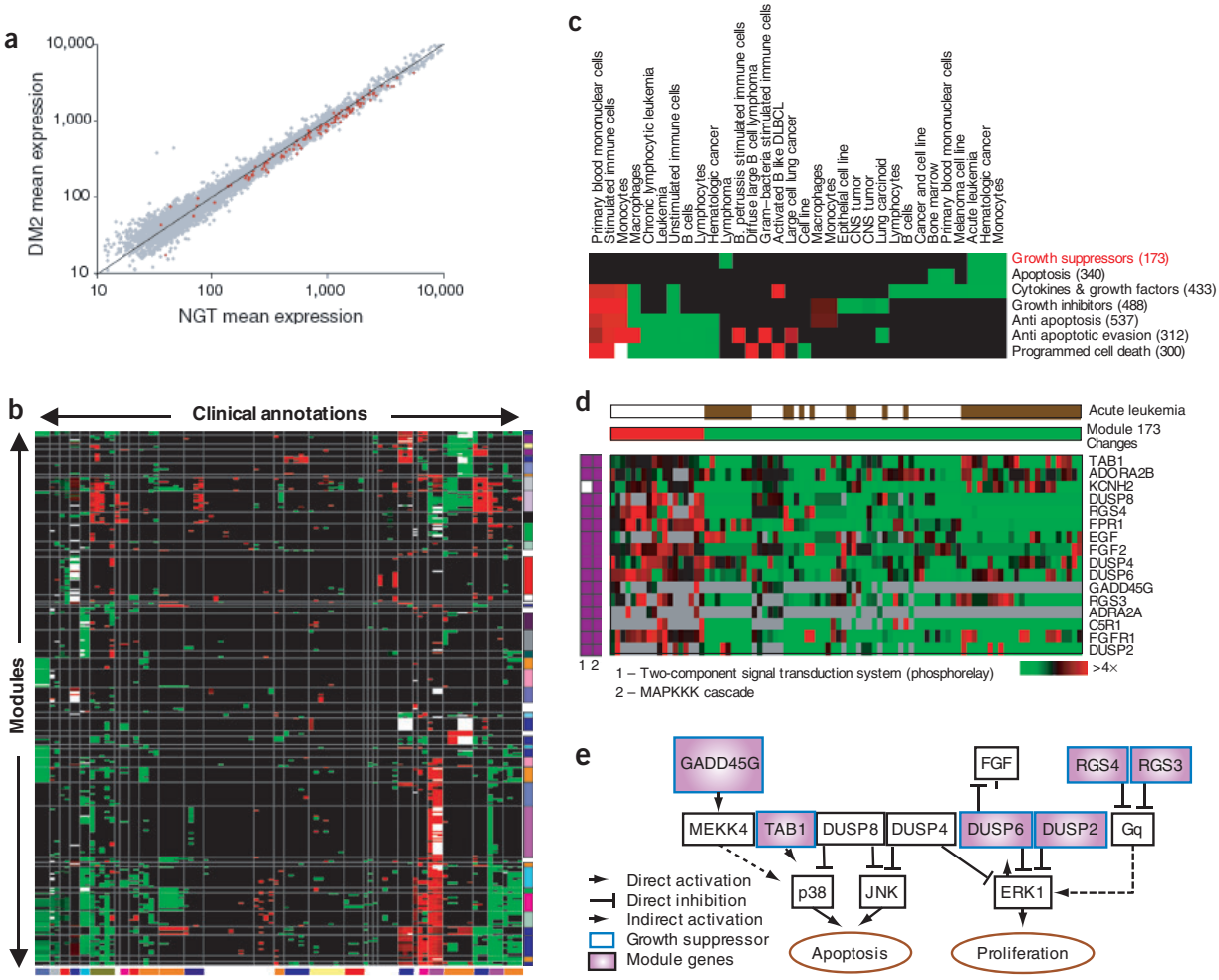

Figure 1 Module-level analysis. (**a**) Example from the gene-set enrichment analysis method of Mootha *et al.*[8] showing that the expression of oxidative phosphorylation genes is reduced in diabetic muscle. The mean expression of all genes (gray) and of the oxidative phosphorylation genes (red) is plotted for individuals with type 2 diabetes mellitus (DM2) versus those with normal glucose tolerance (NGT). The individual genes in the set changed only modestly and could not be identified using standard analyses for differentially expressed genes. But the pattern over the set as a whole is statistically significant. (**b**–**e**) These panels illustrate how module maps[7] suggest new functional roles for specific proliferation and apoptosis genes in acute leukemia. (**b**) The cancer module map of Segal *et al.*[7], shown as a matrix of modules (rows) versus array clinical conditions (columns), in which a red (or green) entry for module *m* and condition *c* indicates that the arrays in which module *m* was significantly induced (or repressed) contained more arrays from condition *c* than would be expected by chance. The intensity of the entries corresponds to the fraction of arrays in the module from condition *c* that were significantly induced (or repressed). White entries indicate that both the induced and repressed arrays were significant for the given annotation. The rows and columns of the matrix were each clustered into distinct clusters, and the resulting clusters are indicated by vertical and horizontal lines. Related conditions (or modules) are often clustered together in the module map. But many modules are shared across conditions, indicating that tumors are characterized by combinations of a small number of shared and unique modules. (**c**) Submatrix of the full map in **b** for related growth-regulatory processes. These modules are mostly used by hematologic malignancies. In most cases, a particular condition shows either uniform induction or repression of most growth-modulating modules, both apoptotic and antiapoptotic, indicating a complex response. (**d**) The Growth Inhibitory Module (highlighted in red in **c**). Shown are all arrays in which the module's genes change significantly, and the direction of change (induction or repression) in each such array is indicated (middle; red or green, respectively). Gray pixels represent missing values. The arrays corresponding to acute leukemia are indicated by brown pixels in the top row. The membership of the module genes in the two gene sets from which the module was generated is shown (left, purple pixels). (**e**) Growth Inhibitory Module genes (purple) in the context of the MAPK pathways of proliferation and apoptosis (as compiled from known interactions in the literature). Most module genes are known to inhibit cell growth (bold blue border). Some are known to directly or indirectly repress ERK1, an activator of cell proliferation known to be constitutively active in acute leukemia. Others are known to activate the apoptosis repressor p38. Thus, the concerted downregulation of these growth suppressors may allow ERK1 and p38 to escape regulation, leading to uncontrolled proliferation and reduced cell death. Only DUSP2 was previously implicated in acute leukemia; other module genes are new potential targets.

a particular combination of module activity, providing insight into the mechanisms underlying specific malignancies. For example, the Growth Inhibitory Module (**Fig. 1d**) consisted primarily of growth suppressors coordinately repressed in a subset of acute leukemia arrays and suggested a possible explanation for the uncontrolled proliferation and reduced cell death in these tumors (**Fig. 1e**). Other modules were shared across a diverse set of clinical conditions, suggestive of common tumor-progression mechanisms. For example, a bone osteoblastic module, spanning various tumor types, included both secreted growth factors and their receptors, suggesting a single mechanism for both primary tumor proliferation and metastasis to bone.

These results and others[8,9,11–13] illustrate the value of analyzing complex processes such as tumorigenesis in terms of higher-level gene modules and biological processes. This type of analysis increases our ability to identify the signal in microarray data and provides results that are more interpretable than gene lists. In particular, when grouped together into a coherent module, the functional and clinical effects of pleiotropic genes might become more apparent, as would the complexity of the mechanism that has to be addressed therapeutically (**Fig. 1e**). Finally, a modular approach can be applied uniformly to multiple data sources from different tumor types, thereby uncovering the commonalities and differences of multiple clinical conditions.

From modules to regulatory mechanisms

The characterization of cancer processes in terms of transcriptional changes in genes or modules is only a step towards the goal of obtaining a detailed mechanistic model of the processes leading to malignancy. Recent work attempts to use gene expression and other genomic data to understand regulatory interactions between genes and how these might result in tumorigenesis.

Cellular processes are regulated by a variety of mechanisms, occurring at every step in the process of going from DNA to functional proteins. Transcriptional regulation, directly observed in gene-expression data, controls the production of mRNA transcripts. Important components in this process are *cis*-regulatory elements in a target gene's promoter region, *trans*-acting factors that bind to these DNA motifs and signaling molecules that modulate this process based on exogenous and endogenous signals. Genomic data sets offer (noisy) views of different facets of this process. Protein-DNA binding events are directly observed in ChIP-chip assays[14,15]. We can computationally detect *cis* elements in

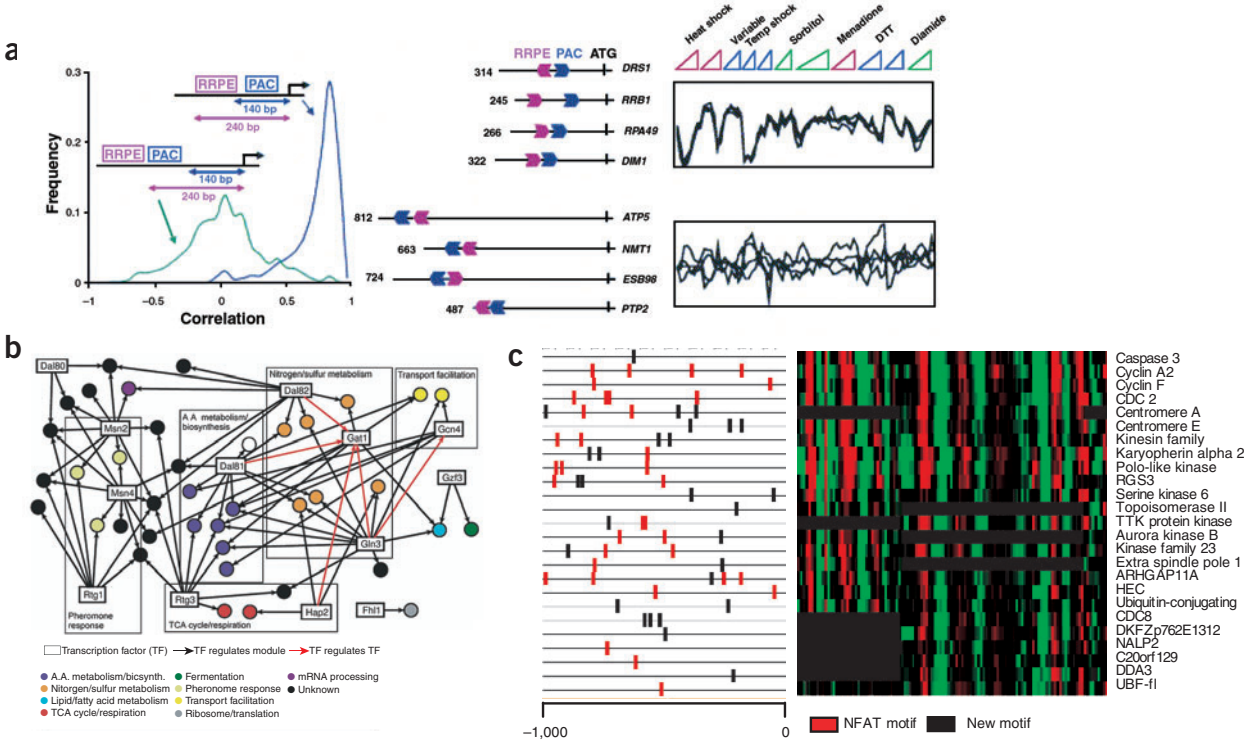

Figure 2 Computational prediction of *cis*-regulatory networks. (**a**) One of the *cis*-regulatory modules produced by the analysis of Beer and Tavazoie[27]. The module is defined in terms of a coherent expression signature and is enriched for genes involved in ribosomal RNA transcription and processing. Its *cis*-regulatory profile is defined by the presence of two computationally discovered sequence elements, PAC and RRPE, in a particular positional configuration on the chromosome. Genes containing both elements in the correct configuration are tightly coregulated, whereas in genes containing only one of the two elements, or containing both elements in a different positional configuration, the distribution of correlations is close to random. The distribution of correlations (left), as well as examples of genes that do (top right) and do not (bottom right) satisfy the positional constraint, along with their expression patterns, are shown. (**b**) Transcriptional gene-regulation network under response to rapamycin, produced by the analysis of Bar-Joseph *et al.*[28]. The network was derived from both gene expression and ChIP data under rapamycin. The analysis resulted in 39 modules (circles) regulated by 13 transcription factors (black arrows). Red arrows between transcriptional regulators indicate that the source transcription factor binds at least one module containing the target transcription factor. The analysis resulted in new predictions regarding transcriptional regulation of the response to rapamycin, for example the regulation of nitrogen metabolism modules by Hap2. (**c**) One of the modules resulting from applying the method of Segal *et al.*[26] to combined human promoters (1,000 bp upstream of predicted transcription start sites) and measured expression of human cell cycle in HeLa cells[62] (E.S. & D.K., unpublished result). Shown is the expression (right) and promoter region (left) of each gene in the module. The genes assigned to this module are known to be involved in mitosis (10 of 25 genes, $P < 10^{-9}$), and one of the motifs automatically identified by the method is the known binding site for the nuclear factor of activated T-cells (NFAT; red motifs), previously reported to have a role in cell-cycle progression[63,64].

promoter sequences, on the basis of experimentally determined sites[16], *de novo* identification or evolutionary conservation[17,18]. Finally, similar expression profiles allow us to identify target genes that are controlled by a shared regulatory mechanism.

Most attempts to identify regulatory relationships from genomic data have focused on the unicellular yeast *Saccharomyces cerevisiae*. One focus aims at reconstructing *cis*-regulatory circuits[19], including identifying new *cis* elements, detecting their targets and identifying combinations of elements that modulate expression of a target gene. Because signal at the level of individual genes is often hard to detect, most approaches focus on regulatory modules, whose member genes are expected to be controlled by similar regulators in a similar way[20,21]. Early approaches identified new individual *cis* elements that are enriched in clusters of coexpressed genes[19], or pairs of elements that act in synergy under specific conditions[22]. Recent extensions increase accuracy by using other sources of data, such as regulator binding (from ChIP-chip assays)[21] or evolutionary conservation[23,24].

More recently, several studies[25–28] have attempted to identify how the set of *cis*-regulatory elements in a gene's promoter governs its behavior and explains the observed expression pattern. Segal *et al.*[26] proposed a model of regulatory modules in which module genes shared both a similar expression profile and a similar profile of *cis* elements. Thus, a gene's *cis* element profile determined its module assignment and hence its expression profile. Beer and Tavazoie[27] subsequently proposed a similar approach, which also included a finer-grained model of promoter configuration. Both groups showed that a substantial fraction of the signal in gene-expression data could be explained in terms of *cis*-element profiles, and that these profiles exhibited an interesting combinatorial organization of elements into various logic gates (OR, AND) and spatial configurations (**Fig. 2a**). This general framework can also accommodate transcription factor–binding data instead of (or in addition to) *cis* elements. For example, Bar-Joseph *et al.*[28] identified gene modules whose expression could be explained by a shared transcription factor–binding profile (**Fig. 2b**), and Segal *et al.*[25] combined expression, sequence and transcription factor–binding to identify combinations of transcription factors, their target modules and the *cis* elements that mediated this regulation.

Despite these successes in model organisms, this approach has yet to be broadly applied in multicellular organisms. In particular, most current methods for detecting *cis* elements are not well suited to the large, complex genomes and long intergenic regions typical of mammals. Nevertheless, several researchers have identified regulatory circuits in expression data from synchronized HeLa cells[26,29], both finding known cell cycle–regulatory elements and targets, and suggesting new ones (**Fig. 2c**). Some of the more successful approaches rely on additional signals, such as evolutionary conservation[30], spatial clustering of *cis* elements in the DNA sequence[30–34] or a global model of *cis* regulation and gene expression[26], to improve the detection of reliable biological signals.

A complementary approach focuses on the transcription factors and signaling molecules that modulate gene expression either directly or indirectly. Although regulator activity is not observed directly, if a regulator is itself transcriptionally regulated, its expression level can serve as a proxy for its activity, allowing us to infer regulatory interactions correctly from expression profiles. Motivated by this insight, several studies[35–37] propose algorithms that construct a Bayesian network describing the probabilistic dependencies between the expression levels of genes. These methods can detect both direct and indirect regulatory relations (*e.g.*, between a MAP kinase and its downstream targets). Recent work[38,39] extends this approach by using more realistic models of binding affinity between transcription factors and binding sites, in accordance with biochemical principles[40].

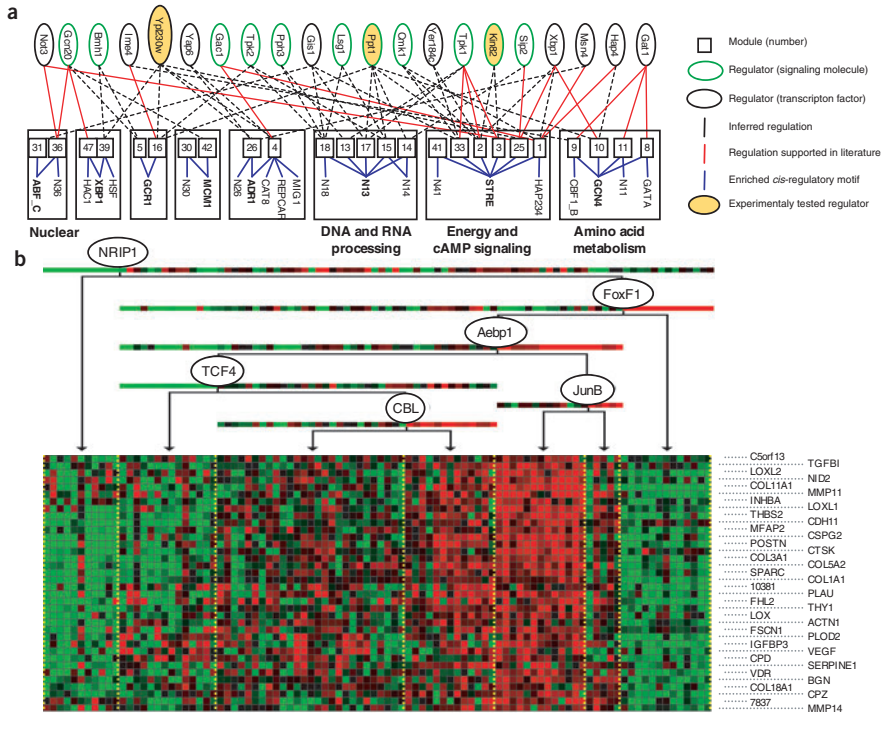

Figure 3 Computational prediction of *trans*-regulatory networks. (**a**) Global module network for yeast stress data, derived by Segal *et al.*[41]. The graph depicts inferred modules (middle, numbered squares), their significantly enriched *cis*-regulatory motifs (bottom) and their associated regulators (top, black-bordered ovals for transcription factors, green-bordered ovals for signal-transduction molecules). Modules are connected to their significantly enriched motifs by solid blue lines. Module groups, consisting of sets of modules that share a common motif, and their associated motifs are enclosed in bold boxes; module groups whose modules are functionally related are labeled (right). Red edges between a module and a predicted regulator are supported in the literature. Three regulators, marked in yellow, correspond to previously uncharacterized regulators whose predicted role was validated experimentally[41]. Modules belonging to the same module group seem to share regulators and motifs, with individual modules having different combinations of these regulatory elements. (**b**) One of the modules identified by the module network analysis of the lung cancer data[43] (N.K., E.S., N.F., A.R., & D.K., unpublished result). According to known databases, 17 of 36 of the genes in this module belong to extracellular matrix–related annotations ($P < 9 \times 10^{-10}$); when we manually curated the genes in this module, we found at least nine additional genes associated with fibrosis and TGFβ signaling. The genes in this module are characterized by having the lowest expression levels in normal lung, with higher expression in individuals who died with the disease. The module network procedure predicts that this module is regulated by Jun-B, a TGFβ-regulated transcription factor, and TCF4, the WNT–β-catenin target transcription factor[65]. The genes are overexpressed when Jun-B is underexpressed, consistent with reports suggesting that loss of Jun-B activity enabled epithelial mesenchymal transition in tumors[66].

A recent extension is based on the observation that many regulatory interactions are shared by all members of a gene module[20,39]. Segal *et al.*[41] proposed the module-network approach for identifying modules of coregulated genes and their shared regulation program, which specified the expression profile of a module's genes as a function of the expression of the module's regulators. As with the identification of cancer modules and *cis* elements, this higher-level analysis improved both statistical robustness and biological interpretability. This approach was successfully applied to a yeast expression data set, identifying functionally coherent modules and known regulatory relations (**Fig. 3a**). It also suggested testable hypotheses regarding the role of transcription factors and signaling molecules, three of which were tested and validated experimentally.

A key limitation of such approaches is that many regulators are regulated post-transcriptionally, and their activity is undetectable in gene-expression data. Nevertheless, in the context of tumorigenic processes, there is reason for optimism. Tumorigenesis often arises from some change to a cell's DNA, which in turn results in a perturbation in expression of certain key regulators. For example, the *Myc* oncogene is amplified in many tumors, resulting in a concomitant change in the expression of its targets[42]. Thus, even regulators that, under normal conditions, are regulated post-transcriptionally may undergo transcriptional regulation in tumor cells, making the regulatory processes more apparent in expression data.

Encouraged by this observation, we applied the module network procedure to a data set of lung cancer arrays[43], focusing on regulation by transcription factors. In addition to the usual cancer-related functional categories (cell cycle, DNA and RNA repair, and metabolism), we found multiple modules enriched for genes associated with extracellular inflammation, immunity and extracellular matrix, processes that are increasingly recognized to be important in tumor generation and progression[44,45]. An in-depth analysis of one of the modules (**Fig. 3b**) suggested that extracellular matrix–related genes, whose expression is often increased in tumors, were not mere representatives of stromal activity but were related to tumor clinical biology and were tightly regulated by cancer-relevant transcription factors. This example illustrates the potential of this approach for identifying transcriptional regulation in complex tissues; it also shows how an unbiased discovery approach can lead the observer to unexpected conclusions (such as the possible role of fibrotic and inflammatory modules in cancer).

Comparative analysis

Taking a more global view, we can extend our analysis to encompass multiple studies across diverse organisms and conditions. In such comparative analysis, conserved patterns can help to identify true biological signals and key mechanisms, and highlight commonalities and differences. This approach is particularly compelling when applied to the

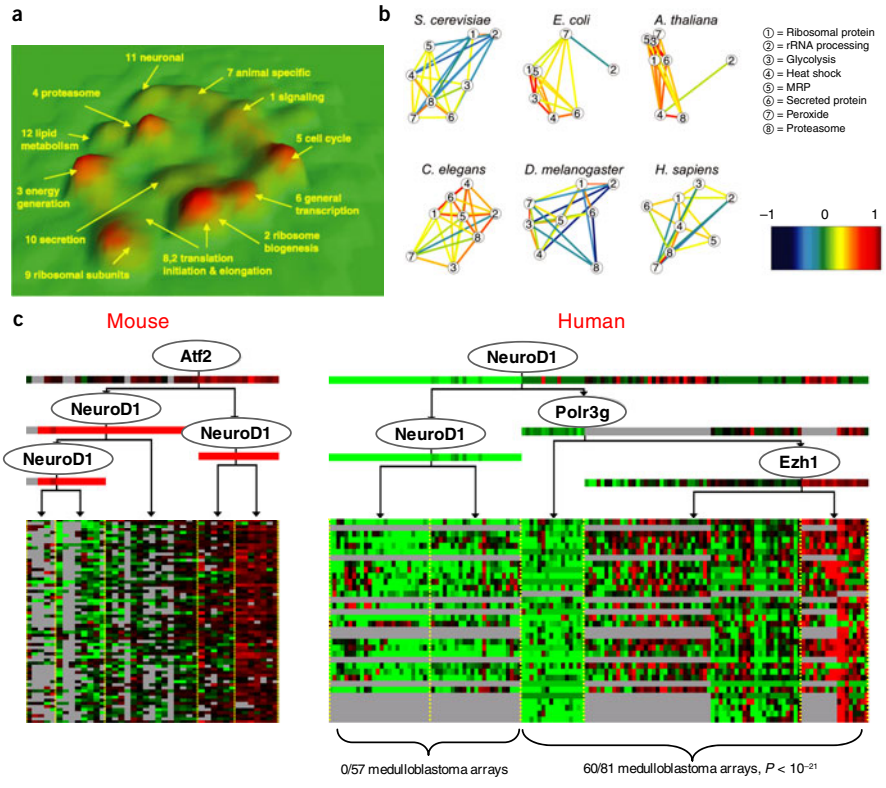

Figure 4 Multispecies analysis of gene expression data. (**a**) Three-dimensional representation of a multispecies coexpression network produced by Stuart *et al.*[46]. The network links metagenes (sets of orthologous genes) across four diverse organisms (human, fly, worm and yeast) that are coexpressed in at least two organisms. The network is visualized as a terrain map, in which highly correlated metagenes are placed in proximity in the *x*-*y* plane, and the density of genes in a region is shown by the altitude in the *z* direction. The visualization uncovers 12 components of highly interconnected metagenes, which were enriched for metagenes involved in similar biological processes. Links in the network suggest potential interactions between genes that have been conserved across evolution and are therefore likely to correspond to functional relationships. Stuart *et al.*[46] showed that the network can be used to predict a gene's function, by labeling it with the annotations of its neighboring genes in the network. They show that, for most functional categories, the multispecies network performed significantly better in making such predictions than any single-species coexpression network. (**b**) Analysis of interaction between eight transcriptional modules across six organisms, as derived by Bergmann *et al.*[47]. Eight modules, whose function is known in yeast, were used to generate corresponding homolog modules in five other organisms. Correlations between the module-expression profiles were computed for all pairs of modules. Modules are shown as circles; significant correlations between them, by colored lines and by distances between the modules. Most of the relations between modules differ among organisms. For example, heat shock and protein synthesis are anticorrelated in yeast and fly and correlated in the other four organisms. (**c**) One of the modules learned in a joint model over expression profiles from normal mouse samples and human brain tumors[67] using a multispecies extension to the module network framework of Segal *et al.*[41] (E.S. & D.K., unpublished result). The module was found in both human and mouse and included 34 orthologous genes from the two organisms. Most of the module genes were previously shown to be expressed in brain[68] (18 of 34, $P < 10^{-12}$), supporting their combination into a single module. Furthermore, a conserved regulatory role was predicted for the neurogenic differentiation factor NeuroD in both human and mouse (ovals in regulation programs). The expression of NeuroD splits the human arrays into two groups; all 60 medulloblastoma arrays in the data set are in the group in which NeuroD is overexpressed, a finding supported by the suggested role of NeuroD in medulloblastoma tumors[69].

available data from an increasing number of mammalian species and of animal models of cancer.

Several works have explored the conservation of coexpression relationships and gene modules across a diverse range of organisms[46–48]. These works showed that conserved coexpression relationships were more likely to correspond to true functional interactions (**Fig. 4a**) and allowed us to study the change in the role of functional modules over evolution (**Fig. 4b**). This analysis can highlight functional modules that have a key role in a process of interest. For example, McCarroll *et al.*[48] identified a common expression signature in aging between flies and worms, which included genes involved in mitochondrial metabolism, DNA repair and cellular transport.

Applying a similar approach to cancer data from mouse and human can shed light on the mechanisms underlying tumorigenesis. For example, Sweet-Cordero *et al.*[49] used three different mouse models of lung cancer to identify signatures of specific genetic alterations that lead to tumorigenesis. They projected the genes in each signature to their human orthologs and used a gene-set–based method[8] to test for activity of these signatures in different human lung tumors. This design used changes observed in controlled manipulations in mouse disease models to draw insights about disease manifestations in humans.

This approach transfers results of an analysis done in mouse to inform a subsequent analysis in human; we can also carry out a joint analysis that explicitly searches for patterns conserved across multiple species. Along these lines, we analyzed a human-mouse data set of normal and tumor brain tissue using an extension[50] of the module network approach[41]. This analysis suggested regulatory modules that were conserved across human and mouse, and proposed new hypotheses regarding regulation in medulloblastoma (**Fig. 4c**).

Challenges and opportunities

The reconstruction of the molecular mechanisms that underlie a complex process, such as tumorigenesis, is a formidable challenge. This challenge arises in part from difficulties associated with microarray assays, including noise in the data and limited reproducibility across platforms and researchers[51,52]. Moreover, most analyses implicitly treat mRNA expression as a surrogate for protein activity level, an assumption that does not account for processes such as mRNA stability, protein degradation and post-translational modifications. In addition, when we attempt to find complex patterns in data, we invariably encounter multiple alternative explanations of the data (*e.g.*, clusters, regulatory modules, etc.). Therefore, the results of such analysis are sensitive to the choice of model and parameters and the specific data used, and must be interpreted with care.

Nevertheless, the successes obtained by combining genomic techniques and computational algorithms to reconstruct networks (albeit primarily in model organisms) are encouraging. Three recurring themes form the basis for this success. The first is the analysis of data at the level of biological modules, rather than individual genes, an approach that produces results that are biologically interpretable and statistically robust. The second is the use of biological knowledge in developing analytic techniques, either directly (*e.g.*, to define functionally coherent gene sets) or indirectly (*e.g.*, to construct biologically realistic models). As we create more realistic biological models, we can hope for better biological understanding and more focused predictions to inform further experiments. The third theme is the integration of multiple sources of data in the analysis. By putting together different partial (and noisy) views of a single complex process (gene expression, promoter sequences, protein-DNA binding, protein-protein interactions and more), we can often obtain a much more accurate and complete picture. In addition, by considering data from different conditions or cell types, we can obtain a more global understanding of the function of the same set of building blocks in different contexts. Finally, the integration of data across organisms allows us to identify functional components based on their conservation and, conversely, to recognize the mechanisms that are the basis for biological diversity.

Although genomic approaches are prevalent in cancer research, we are still far from reconstructing molecular mechanisms in human cancer. In fact, the methods we describe do not always scale easily to mammalian systems. Unlike yeast genomes, mammalian genomes are less compact, and enhancers are more dispersed and remote. Both regulatory and signaling networks are larger and more elaborate, and the control of many genes and processes involves undefined epigenetic mechanisms, a higher degree of combinatorial regulation and multiple signaling pathways. Furthermore, many interactions are context-specific, as different components of the molecular network are active in different cellular states and phenotypes.

Much of the added complexity in applying genomic analysis to cancer is related to multicellularity, which can confound the analysis of data from tissue samples that contain heterogeneous population of cells. Most genomic techniques measure an average signal in a sample from a cell population. This is a concern even when studying unicellular organisms or cell cultures as the averaging process tends to obscure variations between cells[53–55]. When analyzing a heterogeneous tissue, this problem is more pronounced because the signal for different cell types is obfuscated; differential regulation of genes associated to changes in cell state can be hard to distinguish or can even disappear entirely. Moreover, the averaging effect introduces an additional source of noise as the proportions of the different cell types are typically different across samples. This variability may swamp the variability resulting from other, perhaps more relevant, differences between the samples. Another, more challenging issue is raised by intercellular signaling in tissues. Interactions between cells often lead to complex behaviors, which are hard to distinguish from the regulatory processes in the cell itself and cannot be emulated in *in vitro* cell-culture assays. Finally, this epigenetic variability is further confounded in tumor samples, where considerable genetic variability occurs between and within samples.

In light of these challenges, is there hope for systematic mechanistic insights from genomic and computational studies? We believe that a positive answer lies in the combination of computational and experimental insights. Computational methods should be developed to tackle cell and tissue heterogeneity[56,57]. For example, Stuart *et al.*[56] used histological evaluation of tissue heterogeneity to deconvolve expression profiles and identify cell type–specific expression responses. Experimentally, most cancer genomic studies have focused on tumor samples from the human population and have therefore suffered from inevitable confounding genetic and environmental factors, tissue heterogeneity, lack of time courses for disease progression and unavailability of perturbations instrumental in identifying regulatory events. Recent studies[9,49,58] suggest that careful design can greatly improve the utility of such studies, by combining the study of human tissue samples with tissue culture and animal models, to obtain a more controlled and comprehensive view. For example, Lamb *et al.*[9] used expression-profiling in a cell-culture model with genetic perturbations to identify a 'cyclin D signature', followed by computational analysis of a compendium of human tumor expression profiles to find the transcription factor that mediated this response in tumors. Similarly, Kang *et al.*[59] found a set of genes involved in osteoclastic metastasis by combining expression profiling on human cell cultures with phenotypic effects in animal models. Finally, new biological assays, such as *in situ* gene-expression signatures using laser capture microdissection[60,61] or fluorescence microscopy data, can provide more refined observations about gene activation in individual cancer cells[54,55].

Successful identification of mechanisms from genomic data will also require more sophisticated computational methods. Much progress can be made along the themes of modularity, incorporation of biological knowledge and data integration across techniques, conditions or organisms. It is important to develop methods that combine data across experimental systems that address the same phenomenon (*e.g.*, different cancers in humans or the same cancer type in human and mouse) and isolate key mechanisms and root causes of the disease. The development of such computational methods should go hand in hand with that of multipronged experiments combining cell culture, animal models and human tumor samples.

A major challenge for analysis is the identification of the correct context and functional importance of different events and mechanisms. This issue is particularly pronounced in cancer, in which aberrant and normal processes are intertwined. A cancer cell has a mixture of different processes: processes that are the source for tumorigenesis (*e.g.*, a constitutively active Ras mutant); processes that are normal on their own but are suborned by tumors and support their proliferation (*e.g.*, cell division or angiogenesis); processes that may represent the normal host response to the tumor and may even be protective (*e.g.*, immune response and inflammatory-cell infiltration); and perhaps processes that are simply a by-product of cancer and have no functional role. Although some of the modular approaches outlined above enhance our ability to analyze disease process–relevant signatures, we are still far from understanding the role that these signatures have in cancer. We may be able to derive a more comprehensive perspective on cancer processes by integrating existing assays with histopathologic, clinical and environmental information on the one hand, and with measurements of genetic variation, such as SNPs or DNA copy-number changes, on the other.

Finally, when considering the analysis of cancer data, we must keep in mind that our ultimate goal is to improve diagnosis and treatment of the disease. How can the methods we described above help in achieving this goal? Understanding cancer processes and identifying new drug targets is one contribution, but many of the key regulators and basic pathways of carcinogenesis were identified long before the introduction of high-throughput methods, through the careful hypothesis-based work of molecular and cell biologists. Modular analysis can place the complex interactions of these pathways in the biological context of the tumor microenvironment. Previous analyses may tell us that abnormal WNT–β-catenin pathway activation is important in certain solid tumors and increased activation of EGF receptors is important in others. The results of modular analyses can uncover a certain tumor's use of bone-survival machinery (that promotes bone metastasis) or information about its ability to create a proangiogenic microenvironment or evade immune surveillance; any one of these characteristics is potentially crucial to the disease mechanism and the final outcome for an affected individual.

An understanding of the complexity of the pathways that create and sustain tumors can enable a better use of available therapies by using rational combinations in accordance with the pathways that characterize a certain cancer. Furthermore, a detailed view of the tumor's microenvironment could lead to better design of therapeutic interventions that would help to reverse or contain the carcinogenic process. The availability of multiple secreted and membrane proteins that characterize tumors should allow the identification of combinatorial markers for early detection and noninvasive disease classification, whereas the functional and regulatory characterization of tumors should allow personalized treatment of cancer that is based not on histological appearance but on a global and detailed mechanistic understanding of an individual's disease.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

All authors contributed equally to this work. We thank M. Scott and T. Raveh for making available to us their mouse brain microarrays for the multispecies module network analysis. E.S. was supported by a Fellowship from the Center for Studies in Physics and Biology at Rockefeller University. N.F. was supported by the Harry & Abe Sherman Senior Lectureship in Computer Science, by the United States-Israel Bi-National Science Foundation grant and by grants from the US National Institutes of Health. N.K. was partly supported by grants from the US National Institutes of Health, by the Tel-Aviv Chapter of the Israeli Lung Association and by a donation from the Simmons family. A.R. was supported by a grant from the US National Institutes of Health and by the Bauer Center. D.K. was supported by a grant from the US National Science Foundation and by a BioX Center grant.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics/

1. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
2. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
3. Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).
4. Lander, E.S. Array of hope. *Nat. Genet.* **21**, 3–4 (1999).
5. Khan, J. *et al.* Expression profiling in cancer using cDNA microarrays. *Electrophoresis* **20**, 223–239 (1999).
6. Garber, K. Genomic medicine. Gene expression tests foretell breast cancer's future. *Science* **303**, 1754–1755 (2004).
7. 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).
8. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
9. Lamb, J. *et al.* A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**, 323–334 (2003).
10. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* **34**, 226–230 (2003).
11. Rhodes, D.R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* **101**, 9309–9314 (2004).
12. Chang, C.F., Wai, K.M. & Patterton, H.G. Calculating the statistical significance of physical clusters of co-regulated genes in the genome: the role of chromatin in domain-wide gene regulation. *Nucleic Acids Res.* **32**, 1798–1807 (2004).
13. Desai, K.V. *et al.* Initiating oncogenic event determines gene-expression patterns of human breast cancer models. *Proc. Natl. Acad. Sci. USA* **99**, 6967–6972 (2002).
14. Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
15. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**, 8164–8169 (2003).
16. Wingender, E. *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29**, 281–283 (2001).
17. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
18. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
19. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
20. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
21. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
22. Pilpel, Y., Sudarsanam, P. & Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29**, 153–159 (2001).
23. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
24. Pritsker, M., Liu, Y.C., Beer, M.A. & Tavazoie, S. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.* **14**, 99–108 (2004).
25. Segal, E., Barash Y., Simon I., Friedman N. & Koller D. From promoter sequence to expression: a probabilistic framework. *Proceedings of the 6th International Conference on Research in Computational Molecular Biology* 263–272 (ACM Press, Washington, DC, 2002).
26. Segal, E., Yelensky, R. & Koller, D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19** Suppl. 1, i273–i282 (2003).
27. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).

28. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
29. Elkon, R., Linhart, C., Sharan, R., Shamir, R. & Shiloh, Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.* **13**, 773–780 (2003).
30. Sharan, R., Ben-Hur, A., Loots, G.G. & Ovcharenko, I. CREME: cis-regulatory module explorer for the human genome. *Nucleic Acids Res.* **32**, W253–W256 (2004).
31. Schroeder, M.D. *et al.* Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2**, E271 (2004).
32. Segal, E. & Sharan, R. A discriminative model for identifying spatial cis-regulatory modules. *Research in Computational Molecular Biology* 141–149 (ACM Press, San Diego, 2004).
33. Sinha, S., van Nimwegen, E. & Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics* **19** Suppl. 1, i292–i301 (2003).
34. Berman, B.P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **99**, 757–762 (2002).
35. Pe'er, D., Regev, A. & Tanay, A. Minreg: Inferring an active regulator set. *Bioinformatics* **18** Suppl. 1, S258–S267 (2002).
36. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
37. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Combining location and expression data for principled discovery of genetic regulatory networks. *Pacific Symposium on Biocomputing,* 437–439 (World Scientific, Lihue, Hawaii, 2002).
38. Nachman, I., Regev, A. & Friedman, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* **20** Suppl. 1, I248–I256 (2004).
39. Kalir, S. & Alon, U. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117**, 713–720 (2004).
40. Ronen, M., Rosenberg, R., Shraiman, B.I. & Alon, U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* **99**, 10555–10560 (2002).
41. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
42. Lossos, I.S. *et al.* Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. *Proc. Natl. Acad. Sci. USA* **99**, 8886–8891 (2002).
43. Beer, D.G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8**, 816–824 (2002).
44. Wiseman, B.S. & Werb, Z. Stromal effects on mammary gland development and breast cancer. *Science* **296**, 1046–1049 (2002).
45. Chang, H.Y. *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).
46. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
47. Bergmann, S., Ihmels, J. & Barkai, N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* **2**, E9 (2004).
48. McCarroll, S.A. *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* **36**, 197–204 (2004).
49. Sweet-Cordero, A. *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* **37**, 48–55 (2005).
50. Segal, E. *Rich Probabilistic Models for Genomic Data* PhD thesis, Stanford Univ. (2004).
51. Mecham, B.H. *et al.* Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* **18**, 308–315 (2004).
52. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–492 (2005).
53. Cluzel, P., Surette, M. & Leibler, S. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* **287**, 1652–1655 (2000).
54. Lahav, G. *et al.* Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nat. Genet.* **36**, 147–150 (2004).
55. Irish, J.M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
56. Stuart, R.O. *et al.* In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl. Acad. Sci. USA* **101**, 615–620 (2004).
57. Lu, P., Nakorchevskiy, A. & Marcotte, E.M. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. USA* **100**, 10370–10375 (2003).
58. Chang, H.Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* **2**, E7 (2004).
59. Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537–549 (2003).
60. Fuller, A.P., Palmer-Toy, D., Erlander, M.G. & Sgroi, D.C. Laser capture microdissection and advanced molecular analysis of human breast cancer. *J. Mammary Gland Biol. Neoplasia* **8**, 335–345 (2003).
61. Kobayashi, K. *et al.* Identification of genes whose expression is upregulated in lung adenocarcinoma cells in comparison with type II alveolar cells and bronchiolar epithelial cells in vivo. *Oncogene* **23**, 3089–3096 (2004).
62. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.* **13**, 1977–2000 (2002).
63. Caetano, M.S. *et al.* NFATC2 transcription factor regulates cell cycle progression during lymphocyte activation: evidence of its involvement in the control of cyclin gene expression. *FASEB J.* **16**, 1940–1942 (2002).
64. Baksh, S. *et al.* NFATc2-mediated repression of cyclin-dependent kinase 4 expression. *Mol. Cell.* **10**, 1071–1081 (2002).
65. Behrens, J. & Lustig, B. The Wnt connection to tumorigenesis. *Int. J. Dev. Biol.* **48**, 477–487 (2004).
66. Hulboy, D.L., Matrisian, L.M. & Crawford, H.C. Loss of JunB activity enhances stromelysin 1 expression in a model of the epithelial-to-mesenchymal transition of mouse skin tumors. *Mol. Cell. Biol.* **21**, 5478–5487 (2001).
67. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
68. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
69. Rostomily, R.C. *et al.* Expression of neurogenic basic helix-loop-helix genes in primitive neuroectodermal tumors. *Cancer Res.* **57**, 3526–3531 (1997).