nature genetics

A module map showing conditional activity of expression modules in cancer

Eran Segal[1,4], Nir Friedman[2], Daphne Koller[1] & Aviv Regev[3]

DNA microarrays are widely used to study changes in gene expression in tumors, but such studies are typically system-specific and do not address the commonalities and variations between different types of tumor. Here we present an integrated analysis of 1,975 published microarrays spanning 22 tumor types. We describe expression profiles in different tumors in terms of the behavior of modules, sets of genes that act in concert to carry out a specific function. Using a simple unified analysis, we extract modules and characterize gene-expression profiles in tumors as a combination of activated and deactivated modules. Activation of some modules is specific to particular types of tumor; for example, a growth-inhibitory module is specifically repressed in acute lymphoblastic leukemias and may underlie the deregulated proliferation in these cancers. Other modules are shared across a diverse set of clinical conditions, suggestive of common tumor progression mechanisms. For example, the bone osteoblastic module spans a variety of tumor types and includes both secreted growth factors and their receptors. Our findings suggest that there is a single mechanism for both primary tumor proliferation and metastasis to bone. Our analysis presents multiple research directions for diagnostic, prognostic and therapeutic studies.

Cancer is a multifaceted phenomenon, originating in different tissues and involving disruptions of various cellular processes. Aberrations in regulation of key proliferation and survival pathways are common to all tumors, whereas alterations in other pathways may be specific to certain tumors. Understanding which mechanisms are general and which are specific has important therapeutic implications, but few studies[1–4] address this issue from a genome-wide perspective. Here, we used DNA microarray data in a comprehensive analysis aimed at identifying the shared and unique molecular 'modules' underlying human malignancies. Two recent studies[3,5] demonstrate the utility of similar approaches in the context of a single module. The result of our analysis is a global map showing the modules that are induced or repressed in a wide variety of clinical conditions.

We analyzed a 'cancer compendium' of expression profiles compiled from 26 studies (**Supplementary Table 1** online), measuring the expression of 14,145 genes in 1,975 arrays spanning 17 categories (**Fig. 1a**). First, we organized genes into higher-level modules, and then we identified clinical conditions in which different modules are induced or repressed.

We started by collecting 2,849 biologically meaningful gene sets, including clusters of coexpressed genes, genes expressed in specific tissue types[6] and genes belonging to the same functional category or pathway[7–9] (**Fig. 1b**). We identified the arrays in which each gene set has a prominent expression signature by testing whether the expression of a statistically significant fraction of the genes in the set changed coordinately in the array (**Fig. 1c**,**d**). In our compendium, the change in expression of each gene in a given array is relative to the average expression of the gene across all arrays in the relevant data set.

Gene sets reflect biological modules only approximately. Only a subset of genes in a set may contribute to its expression signature, and different gene sets may have similar signatures across the arrays, owing to either an overlap between the gene sets or coregulation of nonoverlapping gene sets. When several gene sets (a cluster) have similar signatures, we extracted from this cluster a core module, which both refines the gene composition of each gene set and combines several related gene sets. This module more closely reflects the genes that participate in a specific biological process, as it consists of the genes whose expression profile corresponds to the signature of the cluster. Overall, we identified 456 statistically significant modules (**Supplementary Note** and **Supplementary Fig. 1** online) that span various processes and functions, including metabolism, transcription, translation, degradation, cellular and neural signaling, growth, cell cycle, apoptosis and extracellular matrix and cytoskeleton components.

In the second step of our analysis, we used these modules to characterize clinical conditions according to the combination of modules that are activated and deactivated in them. Using information provided in the original studies, we annotated all the arrays with 263 biological and clinical conditions, including tissue and tumor type, diagnostic and prognostic information, and molecular markers.

[1]Computer Science Department, Stanford University, Stanford, California 94305, USA. [2]School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel. [3]Bauer Center for Genomics Research, Harvard University, Cambridge, Massachusetts 02138, USA. [4]Present address: Center for Studies in Physics and Biology, The Rockefeller University, New York, New York 10021, USA. Correspondence should be addressed to D.K. (koller@cs.stanford.edu) or A.R. (aregev@cgr.harvard.edu).

Published online 26 September 2004; doi:10.1038/ng1434

For each module and each condition, we tested whether the module was induced (or repressed) in a significant fraction of the arrays labeled with the condition. We distinguished between 'specific' and 'general' annotations: specific annotations are evaluated within each category, whereas general annotations are evaluated only relative to their lack of association with arrays from the other categories. We compiled the module-condition pairs into a global module map for cancer (**Fig. 2**).

The results must be interpreted with caution, because the biological interpretation of induction (or repression) of a module in a given condition depends on our choice of normalization (**Supplementary Note** online). In addition, interpretation may be confounded by combining diverse data sets, each normalized separately. To address this problem, we used annotations in a way that is strictly local to each category (**Supplementary Note** online) in the final analysis step, in which we paired modules with clinical annotations.

The module map shows that some modules (*e.g.*, cell cycle; **Fig. 3a**) are shared across multiple tumor types and may be related to general tumorigenic processes, whereas others are more specific to the tissue origin or progression of particular tumors. For example, modules related to neural processes (*e.g.*, #274 and #137) are repressed in a subset of brain tumors (relative to other central nervous system tumors), and an intermediate filament module (#357) is induced in squamous cell lung carcinomas and reduced in lung adenocarcinomas (both relative to other lung tumors), consistent with the idea that de-differentiation processes accompany tumorigenesis. Related modules, such as cell cycle modules (**Fig. 3a**), seem to form building blocks that are used together in different conditions. More specialized modules, such as signaling and growth regulatory modules (**Fig. 3b,c**), are used in distinct combinations by various tumors.

Conversely, the module map characterizes each condition by a particular combination of modules. For example, invasive hepatocellular carcinoma (HCC) is characterized by induction of cell cycle modules and repression of modules related to metabolism, detoxification, the extracellular matrix and signaling (relative to hepatitis-infected liver tissue and noninvasive HCC). Estrogen receptor–positive breast cancer is characterized by repression of modules containing keratins and other intermediate filaments (relative to other breast adenocarcinomas and human mammary epithelial cells). The map indicates that related conditions involve related modules, albeit in distinct ways (**Fig. 3d,e**). For example, various tumors of hematologic origin (**Fig. 3d**) involve similar immune, inflammation, growth regulation and signaling modules. The pattern of involvement separates different tumor types and subtypes.

Characterizing conditions in terms of modules provides important insights into the mechanisms underlying specific malignancies. For example, the growth inhibitory module (**Fig. 4**) consists primarily of growth suppressors (11 of 16) whose expression is coordinately repressed in a subset of acute leukemia arrays (relative to the leukemia category; 40 arrays; $P < 4 \times 10^{-29}$). Some of these genes are direct (*DUSP2* (ref. 10), *DUSP4* (ref. 11), *DUSP6* (ref. 12)) or indirect (*RGS3* (ref. 13), *RGS4* (ref. 14)) repressors of ERK1, an activator of cell proliferation (**Fig. 4b**) known to be constitutively active in acute leukemia[10]. Others (*MAP3K7IP1* (also called *TAB1*; ref. 15) and *GADD45G* (ref. 16)) are activators of the apoptosis repressor p38 (**Fig. 4b**). Thus, the concerted downregulation of these growth suppressors may allow ERK1 and p38 to escape regulation, leading to uncontrolled prolifera-

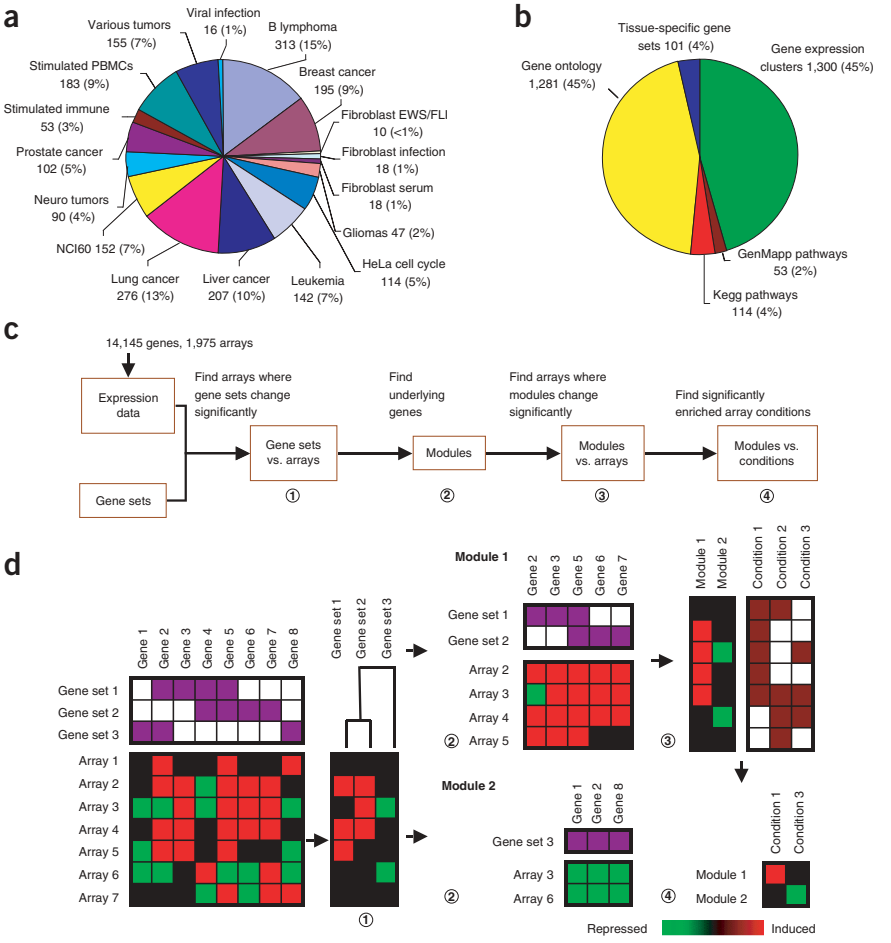

Figure 1 Overview of the analysis procedure. (**a**) Composition of the 1,975 arrays in our compiled cancer compendium according to the conditions they represent. PBMCs, peripheral blood mononuclear cells. (**b**) Composition of the 2,849 gene sets in our analysis according to the source from which they were compiled. (**c**) Flow chart of the different steps in our analysis. (**d**) Example of the analysis on an input expression data of seven arrays, eight genes and three gene sets. Circled numbers correspond to steps in the flow chart. In this example, gene sets 1 and 2 are significantly induced in arrays 2–5 and thus constitute a gene set cluster, whereas gene set 3 is significantly repressed in arrays 3 and 6 and thus constitutes its own gene set cluster. The module resulting from the first gene set cluster includes genes 2, 3, 5, 6 and 7, as these genes contribute to the significant expression of this gene set cluster. Although gene 4 is a member of both gene sets 1 and 2, it is not part of the module, as it did not contribute to their significance (gene 4 is repressed in the arrays where these gene sets are significantly induced). In the final step of the analysis, arrays are annotated with clinical conditions 1–3; for example, array 1 is annotated with conditions 1 and 2. The set of arrays where module 1 is significantly induced (arrays 2–5) is enriched for condition 1, and the set where module 2 is significantly repressed is enriched for condition 3.

Figure 2 The cancer module map: a matrix of modules (rows) versus array clinical conditions (columns), where a red (or green) entry indicates that the arrays in which the corresponding module was significantly induced (or repressed) contained more arrays with the given annotation than would be expected by chance. The intensity of the entries corresponds to the fraction of arrays in the module with the given annotation that were significantly induced (or repressed). White entries indicate that both the induced and repressed arrays were significant for the given annotation. Only significant modules are shown. A subset of significant conditions is shown; redundant conditions were removed for clarity. Only columns (rows) with two or more significant entries are shown. The number of genes in each module and the number of arrays annotated with each condition are shown using gray bars (in log-scale). Each condition annotation is followed by an abbreviated code of the data set in which it was analyzed and by the number of arrays with that annotation. The box (top right) contains details for these abbreviations. Asterisks indicate general annotations. The rows and columns of the matrix were each clustered into distinct clusters[30], and the resulting clusters are indicated by vertical and horizontal lines. We manually assigned, whenever possible, a concise label to module clusters (right; colored bars) or condition clusters (bottom; colored bars). 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. CNS, central nervous system; ECM, extracellular matrix; MMPs, matrix metalloproteinases.

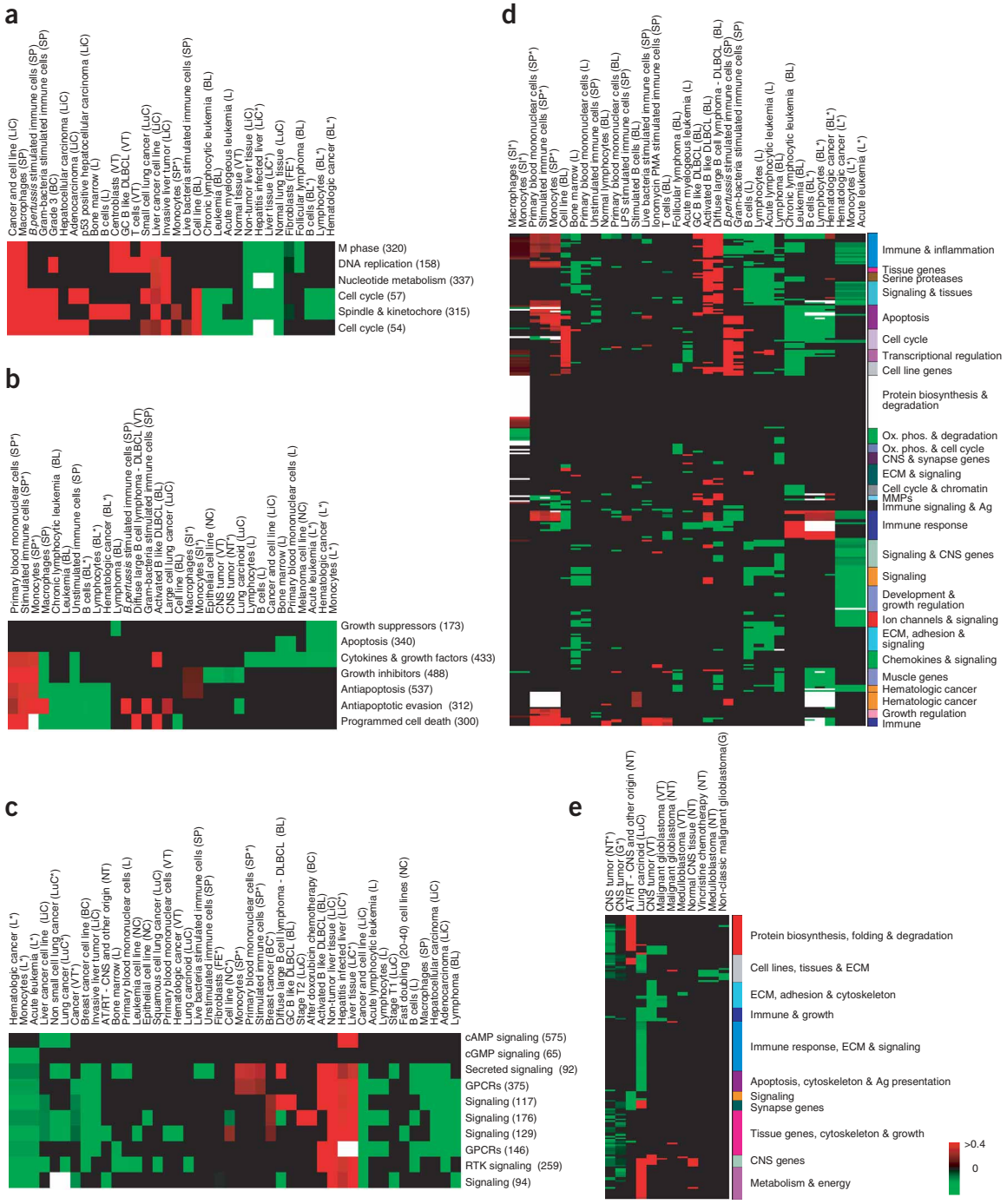

Figure 3 Combinatorial signatures in the cancer module map. Five submatrices of the full map (**Fig. 2**) showing rows of numbered modules organized by conditions that show similarities (**a–c**) and module clusters arranged by related conditions (**d,e**). Each column heading is followed by the code (**Fig. 2**) of the data set on which the condition was analyzed. The box at the top right of **Figure 2** contains details for these abbreviations. (**a**) Cell cycle modules induced in HCC, small cell lung cancer and grade-three breast cancer, repressed in several normal tissues, in chronic lymphocytic leukemia (CLL) and acute myeloid leukemia (AML). (**b**) Growth regulatory 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. (**c**) Signal transduction modules representing a variety of pathways are coregulated in various tumors. Most modules are repressed in HCC and ALL. A subset is induced in activated B-like diffuse large B-cell lymphoma (DLBCL), and another subset is reduced in stage T1 lung adenocarcinoma. White elements indicate modules that are both induced and repressed in the same condition, either because some module genes were induced and others repressed or because the modules were induced in certain arrays and repressed in others from the same condition. GPCR: G protein–coupled receptors; RTK: receptor tyrosine kinase. (**d**) Immune system conditions use similar modules in distinct ways. Many modules are shared across tumor types, cell types and data sets, including DLBCL, ALL, AML, CLL and follicular lymphoma. But each condition has a unique module signature. CNS: central nervous system; ECM: extracellular matrix. (**e**) CNS tumors are characterized by a combination of CNS–specific genes, immune response modules, ECM and cytoskeletal proteins, and neural signaling modules. Lung carcinoid tumors, of neurological origin, use similar modules.

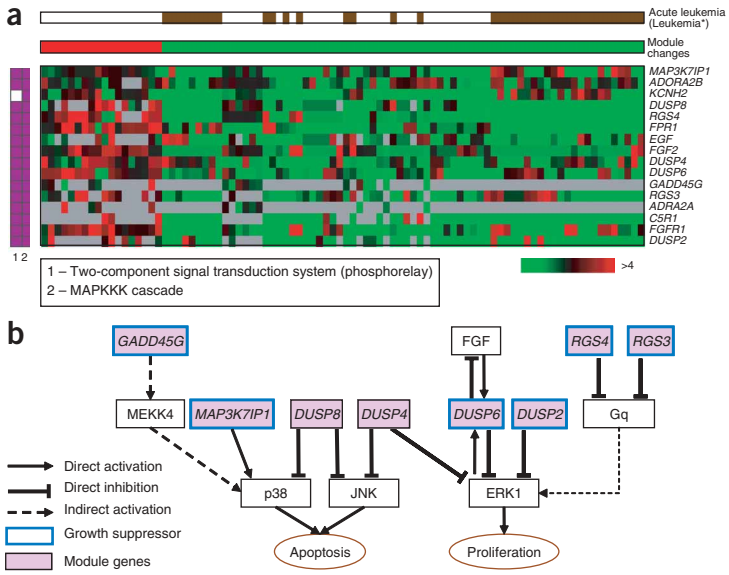

Figure 4 Growth inhibitory module (#173), a module that responds significantly to one specific condition: acute leukemia. (**a**) Expression profile of genes in the growth inhibitory module. Shown are all arrays in which expression of the module's genes changed significantly, and the direction of change (induction or repression) in each such array (red or green, respectively). Gray pixels represent missing values. The arrays corresponding to acute leukemia are indicated by brown pixels in the top row, followed by an abbreviated code of the data set in which they were analyzed. Asterisks denote general annotations. The membership of the module genes in the two gene sets from which the module was generated is shown (left, purple pixels). (**b**) Module genes (purple) in the context of the MAPK pathways of proliferation and apoptosis. The pathway was compiled from known interactions in the literature. All of the module genes were significantly repressed in acute leukemia, and most are known to inhibit cell growth (bold blue border). Only *DUSP2* was previously implicated in acute leukemia; other module genes are new potential targets.

tion and reduced cell death. *DUSP2* has been implicated in acute leukemia[10]; the other genes may offer new therapeutic targets.

The steroid catabolism module (**Fig. 5**) primarily contains steroid hormone enzymes (8 of 13) whose expression is repressed in a subset of HCC and hepatic cell lines (relative to hepatitis-infected liver tissue and HCC; 31 arrays; $P < 4 \times 10^{-8}$). This may indicate more than a general reduction in metabolic processes. Expression of an additional module (#404), consisting of steroid hormone receptors (6 of 25 module genes) and binding proteins (15 of 25), is repressed in a subset of HCC and hepatic cell lines (relative to hepatitis-infected liver tissue and HCC; 24 arrays; $P < 2.5 \times 10^{-6}$). This reduction of steroid hormone catabolism in HCC is consistent with the fact that HCC is significantly more prevalent in men and postmenopausal women[17] and that elevated levels of serum testosterone predict an increased HCC risk. Overall, these results suggest that an imbalance in the generation of steroid hormones and in receiving steroid hormone signals may have a role in hepatitis and HCC.

Other modules provide insight into a variety of tumors. For example, the bone osteoblastic module (**Fig. 6**) consists of genes associated with proliferation and differentiation of bone-building cells. These genes are induced in 172 arrays, including a subset of breast cancer samples (relative to other breast cancer and human mammary epithelial cells; 37 arrays; $P < 5.6 \times 10^{-14}$) and a subset of nontumor hepatitis-infected liver (relative to other hepatitis-infected liver tissue

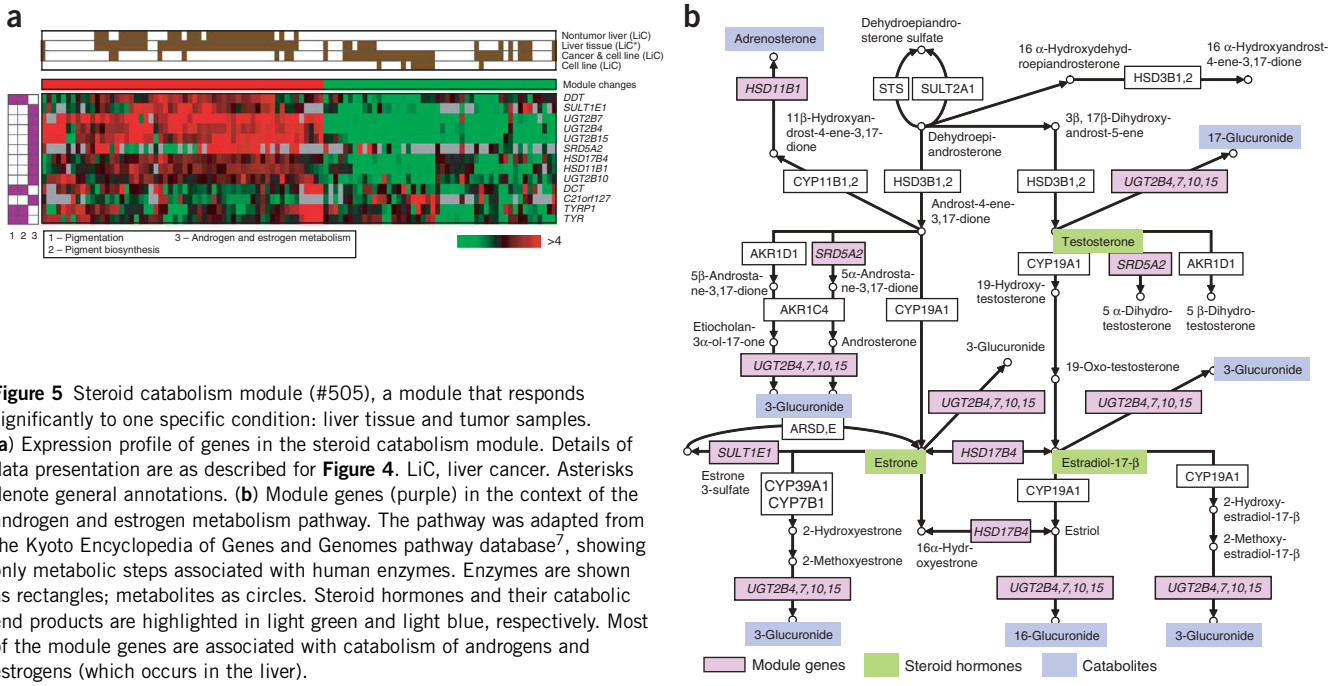

Figure 5 Steroid catabolism module (#505), a module that responds significantly to one specific condition: liver tissue and tumor samples. (**a**) Expression profile of genes in the steroid catabolism module. Details of data presentation are as described for **Figure 4**. LiC, liver cancer. Asterisks denote general annotations. (**b**) Module genes (purple) in the context of the androgen and estrogen metabolism pathway. The pathway was adapted from the Kyoto Encyclopedia of Genes and Genomes pathway database[7], showing only metabolic steps associated with human enzymes. Enzymes are shown as rectangles; metabolites as circles. Steroid hormones and their catabolic end products are highlighted in light green and light blue, respectively. Most of the module genes are associated with catabolism of androgens and estrogens (which occurs in the liver).

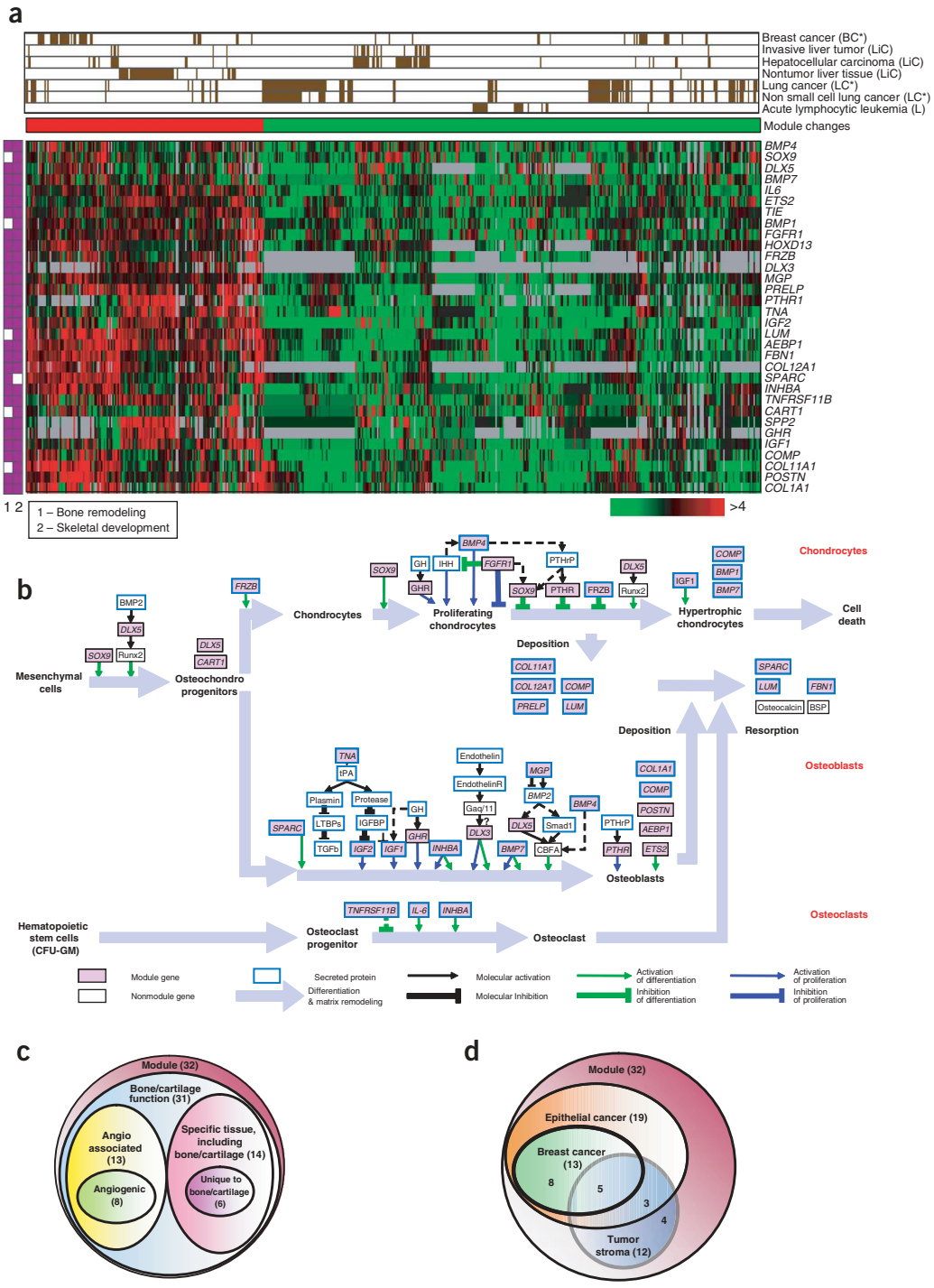

Figure 6 Bone osteoblastic module (#234), a module that responds significantly to multiple conditions, including breast cancer, lung cancer, HCC and ALL. (**a**) Expression profile of genes in the bone osteoblastic module. Details of data presentation are as described for **Figure 4**. LiC, liver cancer; BC, breast cancer; LC, lung cancer; L, leukemia. Asterisks denote general annotations. (**b**) Module genes in the context of the molecular pathways underlying bone remodeling. The pathways are shown for the differentiation and matrix remodeling events (light blue arrows) of the three main cell types in bone and cartilage: chondrocytes (top), osteoblasts (middle) and osteoclasts (bottom). The coordination and balance among the three processes results in either bone building or resorption. The module genes (purple) are primarily associated with proliferation and differentiation of chondrocytes and osteoblasts. Even those module genes that are related to osteoclast induction encode proteins that are typically secreted by osteoblasts. The genes include both intracellular or membrane proteins (thin black border) and extracellular secreted ones (bold blue border), thus forming a coherent and self-sufficient autocrine module. (**c**) The expression and function of 32 module genes in normal tissues based on previous immunohistochemical and *in situ* hybridization experiments. Almost all (31 of 32) of the genes function in bone or cartilage (blue), and 14 are expressed primarily (pink) or uniquely (purple) in bone or cartilage. In contrast, only 8 of the genes are angiogenic (green), and another 5 genes are partly associated with blood vessels or antiangiogenic function (yellow). (**d**) The expression of 23 of the 32 module genes in epithelial tumors and their surrounding stroma based on previous immunohistochemical and *in situ* hybridization experiments. Whereas 19 of the genes are associated with breast cancer (green) or other epithelial tumors (orange), only 4 are expressed solely in stroma (blue).

and HCC; 47 arrays; $P < 10^{-10}$). Expression of these genes is repressed in 361 arrays, including subsets of HCC (relative to other hepatitis-infected liver tissue and HCC; 48 arrays; $P < 2 \times 10^{-9}$), a subset of ALL1 acute lymphoblastic leukemia (relative to other acute lymphoblastic leukemia and acute myeloid leukemia; 10 arrays; $P < 9 \times 10^{-6}$) and a subset of lung cancer samples (relative to other lung cancers; 120 arrays; $P < 10^{-33}$).

Bone-related clinical conditions have been associated with all of these malignancies. In particular, bone metastasis is a key phenomenon in breast cancer, and some breast metastases are known to be osteoblastic[18]. Not all primary breast tumors activate the osteoblastic module, consistent with the fact that many breast metastases to bone are not osteoblastic[18] and probably use different mechanisms[19]. Bone metastasis is also common in lung cancer[18] and was recently implicated in HCC[20]. Finally, ALL has been associated with reduced bone-mass density in a subpopulation of individuals[21]. The bone osteoblastic module reflects these diverse phenomena and may partially explain them. Although osteoblastic metastasis is also common in prostate cancer[18], the module was not substantially expressed in the prostate cancer samples in our compendium. As several genes in the module that are known to be transcriptionally induced in prostate cancer (*MGP*, *IGF2*, *IL6* and *GHR*) are not induced in this data set, we suspect that these arrays are uninformative about osteoblastic metastasis.

The induction of the bone osteoblastic module in breast cancer is particularly interesting. Previous studies suggested that breast tumors preferentially metastasize to bone owing to a cycle of positive feedback through reciprocal secretion of growth factors between the tumor and bone cells[18]. It was previously unclear, however, whether the molecular mechanisms necessary to initiate this cycle are present in the primary tumor[19]. We found that both the secreted growth factors and the intracellular proteins required to receive their signal were induced in primary breast cancer tumors, suggesting that the primary tumor uses the osteoblastic mechanism for its own paracrine proliferation. One might suspect that the module is induced in the surrounding stroma rather than in the tumor itself. Previous immunohistochemical and *in situ* hybridization experiments (**Fig. 6d**) indicate that 19 of the 32 module genes are expressed in epithelial cells in tumors and some also in metastasis of breast cancer to bone (*e.g.*, *IGF2* (ref. 18), *BMP4* (ref. 18), *IL6* (ref. 18), *FRZB*[22] and activin A[23]). Only 4 of 32 genes, all of which encode secreted proteins, are expressed solely in the stroma, indicative of possible paracrine signaling between tumor and breast stroma. This process may be subsequently substituted by signaling between the metastasized tumor and bone stroma. Thus, this borrowed module may both be innately useful to the primary tumor and provide a mechanism for effective osteoblastic bone metastasis. This hypothesis is consistent with recent findings on the metastatic potential of primary tumors[24,25] and identifies several new targets for further research.

The downregulation of the bone osteoblastic module in HCC, ALL and lung cancer is also notable. There is no clear explanation for this downregulation in lung and HCC tumors, but repression of this growth-inducing module in the ALL bone marrow samples provides a potential explanation for the reduced bone mass density in ALL. Dlx3 and Dlx5, two ALL-1 targets that are crucial to osteoblast proliferation and differentiation[26], are part of the module.

In conclusion, our method provides a global view of cancer and shows that tumors can be characterized by combinations of a relatively small number of modules. Several other methods have been proposed for global analysis of microarray data[27–29]. Notably, our work, which is the first to apply such global analysis to human data, uses existing biological knowledge directly, in the form of gene sets and clinical annotations. Furthermore, unlike recent meta-analysis[4] of a large compendium of cancer expression profiles, our approach focuses on identifying modules of genes and is independent of predefined queries (**Supplementary Note** online).

The results of our analysis are publicly available on a data-mining website; the automated tool that we used to generate the analysis is also available. This tool allows researchers to construct a module map from any collection of gene sets and expression data in any organism and to study new data in the context of a large compendium. Although the quality of current annotations and normalization procedures may limit the map's accuracy, our examples indicate that many phenomena are sufficiently robust to be detected using our approach. Thus, our approach provides a valuable tool for understanding the molecular basis of cancer, both for specific tumors and for tumorigenic processes in general.

METHODS

DNA microarray data set. We downloaded data available for 1,975 human DNA microarrays from the Stanford Microarray Database and the Center for Genomic Research at the Whitehead Institute (**Supplementary Table 1** online). We normalized the expression of each gene g in every data set separately. For data sets generated using Affymetrix chips, we first determined the log (base 2) of the expression value of gene g in each array (truncating to 10 expression values that are below 10). For data sets generated using spotted cDNA chips, we used the log-ratio (base 2) between the measured sample and the control sample. In both types of data sets, we then normalized the (log-space) expression value of gene g in each array relative to its average expression in all the arrays in the same data set, by subtracting its average in that data set from each of its expression measurements. After this normalization, the mean value of a gene, in each data set, is zero.

Gene sets. We compiled 2,849 gene sets, obtained as follows: 1,281 from the Gene Ontology[8] hierarchy (downloaded on July 2003, version 1.320); 114 from the Kyoto Encyclopedia of Genes and Genomes[7] (downloaded on May 2003); 53 from the Gene MicroArray Pathway Profiler[9] (downloaded on July 2003); 101 tissue-specific expressed gene sets[6] (one gene set was defined for each array by taking all genes above absolute expression of 400; we removed genes whose absolute expression was >400 in >50 of the 101 arrays); and 1,300 gene sets obtained by clustering each of the data sets of **Supplementary Table 1** online using a published clustering method (the P-cluster algorithm[27]) and taking clusters of coexpressed genes.

Identifying arrays in which the expression of gene sets changes significantly. To identify the arrays in which each gene set was significantly induced (or repressed), we defined the induced (or repressed) genes in each array to be those genes whose change in expression was greater (or less) than twofold. For each gene set and each array, we calculated the fraction of genes from that gene set that were induced (or repressed) in that array and used the hypergeometric distribution to calculate a P value for this fraction (compared with the null hypothesis of choosing the same number genes at random). We corrected for multiple tests using the false discovery rate correction with 5% false rate.

Statistical significance of array–gene set pairs. We evaluated the number of array–gene set pairs in which the gene set was significantly induced (or repressed) in the array (as described above). Overall, we found 299,233 such pairs; only 14,962 would be expected by chance ($P < 0.05$), suggesting that the selected gene sets are informative for the cancer compendium (**Supplementary Fig. 2** online).

Automatic identification of gene set clusters. We carried out (bottom-up) hierarchical clustering of the gene sets in the matrix of all significant array–gene set pairs[30]. This resulted in a tree in which each leaf node, corresponding to some gene set G, is associated with a vector (indexed by arrays) that is zero everywhere except for entries that correspond to arrays in which set G was significantly induced (or repressed), in which case the entry contains the

fraction (or negative fraction) of genes from set G that are induced (or repressed) in an array a. Each internal node is associated with a vector representing the average of all of the gene set vectors at its descendant leaves. We annotated each interior node with the Pearson correlation between the vectors associated with its two children in the hierarchy. We defined as a cluster each interior node whose Pearson correlation differed by more than 0.05 from the Pearson correlation of its parent node in the hierarchy, resulting in 577 clusters of gene sets. Such interior nodes represent points in the tree with a large gap between the similarities in expression of the node's children and the similarity in expression of the node and its sibling.

Testing consistency of a gene with expression of a gene set. Given a gene set G and a gene g, we tested whether the expression of g was consistent with the significant changes in the expression of G. We first identified the subsets of arrays I and R in which G was significantly induced and repressed, respectively. We then measured the extent to which the expression of g changed by more (or less) than twofold in arrays in I (or R) with the score

$$\text{Score}(g) = \sum_{\{a \in I | g \text{ is induced in } a\}} -\log(p_a) + \sum_{\{a \in R | g \text{ is repressed in } a\}} -\log(p_a),$$

where p_a is the fraction of genes in array a that are induced (or repressed) by more than twofold for arrays in I (or in R). This score assigns more weight to induction in arrays where there are fewer induced genes (and respectively for repression).

We evaluated the significance of the score for gene g with respect to the null hypothesis where the genes in each array are randomly permuted. Under this null hypothesis, the score for gene g is the sum of independent binary random variables, one for each array in I and R. The random variable corresponding to array a attains the value $-\log(p_a)$ with probability p_a and the value of 0 with probability $1 - p_a$. Because the score for gene g in this model is a sum of independent random variables, its mean μ and variance σ^2 are the sum of the means and variances, respectively, of the these variables and can be computed analytically:

$$\mu = \sum_{a \in I \cup R} -p_a \log p_a$$

$$\sigma^2 = \sum_{a \in I \cup R} p_a (1 - p_a) \log^2 p_a.$$

Moreover, by the central limit theorem, the distribution of the score for gene g under the null hypothesis can be closely approximated by a Gaussian distribution with mean μ and variance σ^2. We used standard methods for computing the tail probability of a Gaussian distribution to compute the probability of attaining a score as large as the observed score under the null hypothesis.

Deriving modules from clusters of gene sets. For each cluster of gene sets, we defined G to be the union of the gene sets in the cluster. We then tested each gene in G for consistency (as described above). The resulting module consists of genes whose expression is significantly consistent with the expression of the gene set (after false discovery rate correction for multiple hypotheses using 5% false rate). Leave-one-out cross-validation analysis (**Supplementary Note** and **Supplementary Fig. 1** online) showed that 456 of the 577 gene-set clusters were significant at $P < 0.01$. All further analysis was carried out only for the 456 modules derived from these 456 gene set clusters.

Enrichment of clinical annotations. To characterize conditions as a combination of activated and deactivated modules, we associated each array with the annotations it represents, from a total of 263 clinical annotations that we compiled based on published studies (see our project website for the complete set of clinical annotations). We distinguished between 185 specific annotations (present in <70% of the arrays in a given category; **Fig. 1a** and project website) and 78 general annotations (present in 70% or more of the arrays in a category). For example, 'Stage T2' is a specific annotation in the 'lung cancer' category (12.6% of samples in this category), whereas 'lung cancer' is a general annotation (86% of the samples in the 'lung cancer' category). For each module and each annotation, we calculated the fraction of arrays associated with that annotation of the total number of arrays in which the module is significantly induced (or repressed) and used the hypergeometric distribution to calculate a P value for this fraction. For specific annotations, we only considered arrays in the same category when computing the P value. For general annotations, we considered all other arrays in the compendium as background (*i.e.*, the other arrays were marked as not having the general annotation). In both cases, all annotations were strictly local (*e.g.*, the lung cancer annotation in the lung cancer category is distinct from the lung cancer annotation in the 'various tumors' category and is reported separately). We carried out a false discovery rate correction for multiple hypotheses and took $P < 0.05$ to be significant in **Figure 2**.

GeneXPress. We carried out all analysis and visualizations in GeneXPress. This tool can identify the arrays in which gene sets are significantly expressed, and the clinical annotations enriched in these significant arrays, and can be used for any input expression data and gene sets in any organism. GeneXPress is freely available for academic use.

URLs. More detailed results, including the expression compendium, clinical annotations that we compiled and all the significant gene set–array pairs, viewable in GeneXPress, can be found on our project website (http://dags.stanford.edu/cancer). The website also contains detailed views of all 456 modules in the format of **Figures 4–6**, which can be searched and browsed in various ways. GeneXPress is freely available for academic use at http://GeneXPress.stanford.edu/. All expression data used is available from the Stanford Microarray Database (http://genome-www5.stanford.edu/Microarray/SMD/) and the Center for Genomic Research at the Whitehead Institute (http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi).

ACKNOWLEDGMENTS

We thank J. Effrat, T. Fojo, Y. Friedman, A. Kaushal, W. Lu, T. Pham, M. Tong, and R. Yelensky for technical help with software and visualization and I. Ben-Porath, Y. Dor, L. Garwin, N. Kaminski, D. Pe'er, O. Rando and T. Raveh for comments on previous versions of this manuscript. E.S., N.F. and D.K. were supported by a National Science Foundation grant under the Information Technology Research program. E.S. was also supported by a Stanford Graduate Fellowship. N.F. was also supported by an Alon Fellowship, by the Harry & Abe Sherman Senior Lectureship in Computer Science and by the United States-Israel Bi-National Science Foundation grant. N.F. and A.R. were supported by a Center of Excellence Grant from the National Institute of General Medical Sciences. A.R. was also supported by the Bauer Center for Genomics Research.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 16 March; accepted 25 August 2004
Published online at http://www.nature.com/naturegenetics/

1. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
2. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **98**, 15149–15154 (2001).
3. 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).
4. 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).
5. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–723 (2003).
6. Su, A.I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465–4470 (2002).
7. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
8. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
9. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31**, 19–20 (2002).
10. Kim, S.C. *et al.* Constitutive activation of extracellular signal-regulated kinase in human acute leukemias: combined role of activation of MEK, hyperexpression of extracellular signal-regulated kinase, and downregulation of a phosphatase, PAC1. *Blood* **93**, 3893–3899 (1999).
11. Chu, Y., Solski, P.A., Khosravi-Far, R., Der, C.J. & Kelly, K. The mitogen-activated protein kinase phosphatases PAC1, MKP-1, and MKP-2 have unique substrate specificities and reduced activity in vivo toward the ERK2 sevenmaker mutation. *J. Biol. Chem.* **271**, 6497–6501 (1996).

12. Furukawa, T., Sunamura, M., Motoi, F., Matsuno, S. & Horii, A. Potential tumor suppressive pathway involving DUSP6/MKP-3 in pancreatic cancer. *Am. J. Pathol.* **162**, 1807–1815 (2003).
13. Leone, A.M., Errico, M., Lin, S.L. & Cowen, D.S. Activation of extracellular signal-regulated kinase (ERK) and Akt by human serotonin 5-HT(1B) receptors in transfected BE(2)-C neuroblastoma cells is inhibited by RGS4. *J. Neurochem.* **75**, 934–938 (2000).
14. Shi, C.S. *et al.* Regulator of G-protein signaling 3 (RGS3) inhibits Gbeta1gamma 2-induced inositol phosphate production, mitogen-activated protein kinase activation, and Akt activation. *J. Biol. Chem.* **276**, 24293–24300 (2001).
15. Ge, B. *et al.* TAB1beta (transforming growth factor-beta-activated protein kinase 1-binding protein 1beta), a novel splicing variant of TAB1 that interacts with p38alpha but not TAK1. *J. Biol. Chem.* **278**, 2286–2293 (2003).
16. Mita, H., Tsutsui, J., Takekawa, M., Witten, E.A. & Saito, H. Regulation of MTK1/MEKK4 kinase activity by its N-terminal autoinhibitory domain and GADD45 binding. *Mol. Cell. Biol.* **22**, 4544–4555 (2002).
17. Granata, O.M. *et al.* Altered androgen metabolism eventually leads hepatocellular carcinoma to an impaired hormone responsiveness. *Mol. Cell. Endocrinol.* **193**, 51–58 (2002).
18. Mundy, G.R. Metastasis to bone: causes, consequences and therapeutic opportunities. *Nat. Rev. Cancer* **2**, 584–593 (2002).
19. Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537–459 (2003).
20. Iguchi, H. *et al.* A possible role of VEGF in osteolytic bone metastasis of hepatocellular carcinoma. *J. Exp. Clin. Cancer Res.* **21**, 309–313 (2002).
21. Boot, A.M., van den Heuvel-Eibrink, M.M., Hahlen, K., Krenning, E.P. & de Muinck Keizer-Schrama, S.M. Bone mineral density in children with acute lymphoblastic leukaemia. *Eur. J. Cancer* **35**, 1693–1697 (1999).
22. Ugolini, F. *et al.* Differential expression assay of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and Fibroblast Growth Factor Receptor 1 (FGFR1) as candidate breast cancer genes. *Oncogene* **18**, 1903–1910 (1999).
23. Reinholz, M.M., Iturria, S.J., Ingle, J.N. & Roche, P.C. Differential gene expression of TGF-beta family members and osteopontin in breast tumor tissue: analysis by real-time quantitative PCR. *Breast Cancer Res. Treat.* **74**, 255–269 (2002).
24. Bernards, R. & Weinberg, R.A. A progression puzzle. *Nature* **418**, 823 (2002).
25. Hynes, R.O. Metastatic potential: generic predisposition of the primary tumor or rare, metastatic variants-or both? *Cell* **113**, 821–823 (2003).
26. Ferrari, N. *et al.* DLX genes as targets of ALL-1: DLX 2,3,4 down-regulation in t(4;11) acute lymphoblastic leukemias. *J. Leukoc. Biol.* **74**, 302–305 (2003).
27. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
28. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377 (2002).
29. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986 (2004).
30. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).