

An oncogenic *KRAS2* expression signature identified by cross-species gene-expression analysis

Alejandro Sweet-Cordero¹, Sayan Mukherjee^{3,4}, Aravind Subramanian³, Han You¹, Jeffrey J Roix¹, Christine Ladd-Acosta³, Jill Mesirov³, Todd R Golub^{2,5} & Tyler Jacks^{1,5}

Using advanced gene targeting methods, generating mouse models of cancer that accurately reproduce the genetic alterations present in human tumors is now relatively straightforward. The challenge is to determine to what extent such models faithfully mimic human disease with respect to the underlying molecular mechanisms that accompany tumor progression. Here we describe a method for comparing mouse models of cancer with human tumors using gene-expression profiling. We applied this method to the analysis of a model of *Kras2*-mediated lung cancer and found a good relationship to human lung adenocarcinoma, thereby validating the model. Furthermore, we found that whereas a gene-expression signature of *KRAS2* activation was not identifiable when analyzing human tumors with known *KRAS2* mutation status alone, integrating mouse and human data uncovered a gene-expression signature of *KRAS2* mutation in human lung cancer. We confirmed the importance of this signature by gene-expression analysis of short hairpin RNA-mediated inhibition of oncogenic *Kras2*. These experiments identified both a pattern of gene expression indicative of *KRAS2* mutation and potential effectors of oncogenic *KRAS2* activity in human cancer. This approach provides a strategy for using genomic analysis of animal models to probe human disease.

Gene-expression profiling is a powerful technique to identify tumor subtypes, prognostic signatures and potential therapeutic targets in human cancer^{1–4}. Over the last decade, many mouse models of human cancer that closely mimic specific oncogenic events have been developed. Therefore, it is now possible to compare gene-expression profiles of human and mouse cancer. Results from such an approach could validate the molecular similarity of the animal model to its presumed human counterpart. In addition, the comparison of gene-expression profiles from the mouse model and the human disease could uncover patterns or pathways relevant to human cancer that are obscured in the human data. This approach is likely to be fruitful, because whereas most human cancers are genetically heterogeneous, mouse models are often initiated by a single event that is well characterized. Previous work has identified a close correlation between the expression profile and the presumed initiating event in human leukemia and in mouse breast tumors^{5–7}. Additionally, gene expression microarrays of mouse embryonic fibroblasts transformed by either MYC or HRAS have been used to identify linear combinations of genes ('metagenes') that can accurately discriminate mouse tumors in which expression of either MYC or HRAS is the initiating oncogenic event. These results indicate that identification of gene-expression signatures may be possible for a broad range of oncogenes.

Accumulating evidence suggests that endogenous activation of oncogenes *in vivo* has biological consequences that are distinct from

the effects of oncogene overexpression *in vitro*. For example, *Kras2* overexpression *in vitro* leads to senescence in the absence of cooperating genetic events, whereas expression of oncogenic *Kras2* from the endogenous promoter leads to immortalization and partial transformation⁹. Biochemical evidence suggests that the transforming properties of oncogenic Ras proteins are due to aberrant signaling caused by constitutive activation in the GTP-bound form. Several downstream effector pathways that mediate the physiological function of Ras family proteins have been described. The most well-studied of these are the Raf-MAPK-ERK and PI3K pathways, but other effectors such as RalGDS, Tiam1 and RASSF1A probably have roles in mediating Ras-induced tumorigenesis¹⁰. In addition, effector pathways downstream of Ras are probably cell type-dependent. Many effects of Ras activation are likely to be obscured by the analysis of gene expression. Nevertheless, comparison of gene-expression profiles in cells carrying wild-type Ras and mutant Ras could lead to important insights into the cell type-specific pathways that lead to Ras-mediated oncogenesis. Previous reports identified gene-expression correlates of Ras transformation *in vitro* in rat embryonic fibroblasts¹¹ and in rat ovarian epithelial cells¹². By comparing gene-expression profiles of mouse and human lung cancer, we uncovered a gene-expression profile that is common to lung adenocarcinoma in mice and humans, as well as a specific expression signature of *KRAS2* activation *in vivo*.

¹MIT Center for Cancer Research, Building E17-517, 40 Ames Street, Cambridge, Massachusetts 02139, USA. ²Dana-Farber Cancer Institute, 44 Binney Street, Dana Building, Room 604C, Boston, Massachusetts 02115, USA. ³Eli and Edyth L Broad Institute of MIT and Harvard University, 320 Charles Street, Cambridge, Massachusetts 02141-2033, USA. ⁴Present address: Institute for Genome Sciences and Policy, Institute for Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA. ⁵These authors contributed equally to this work. Correspondence should be addressed to T.J. (tjacks@mit.edu).

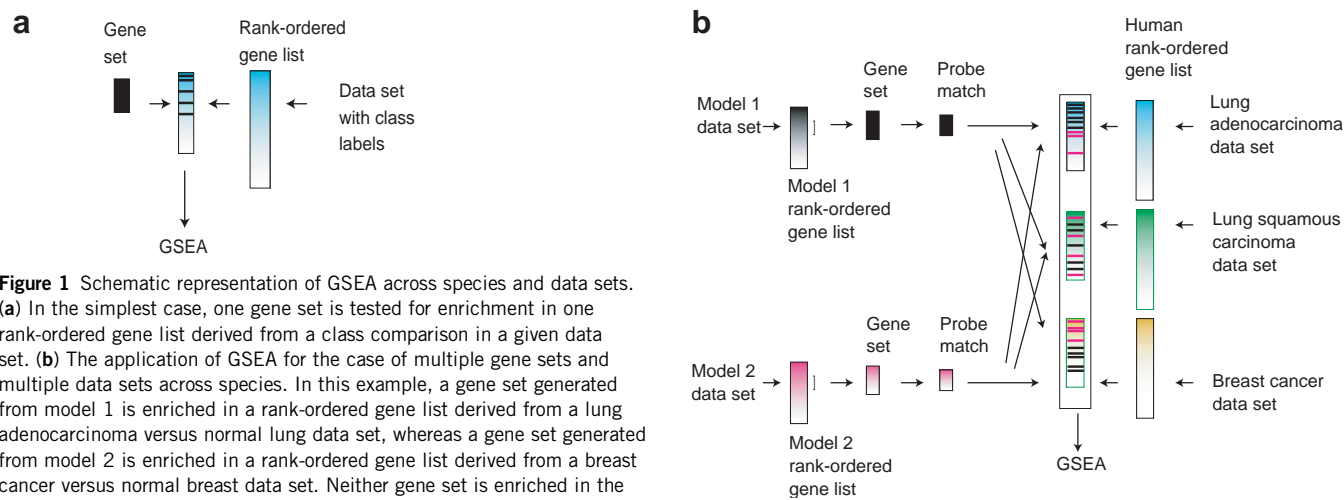


Figure 1 Schematic representation of GSEA across species and data sets.

(a) In the simplest case, one gene set is tested for enrichment in one rank-ordered gene list derived from a class comparison in a given data set. (b) The application of GSEA for the case of multiple gene sets and multiple data sets across species. In this example, a gene set generated from model 1 is enriched in a rank-ordered gene list derived from a lung adenocarcinoma versus normal lung data set, whereas a gene set generated from model 2 is enriched in a rank-ordered gene list derived from a breast cancer versus normal breast data set. Neither gene set is enriched in the lung squamous versus normal lung.

RESULTS

Comparison of gene expression in mouse and human lung cancer

We generated mouse lung cancers using the KrasLA model, in which a latent mutated *Kras2* allele (resulting in the amino acid substitution G12D) is sporadically activated through spontaneous homologous recombination¹³. These mice develop lung adenomas with full pene-

trance; over time, the tumors acquire morphologic characteristics reminiscent of those of human adenocarcinoma, such as nuclear atypia and a high mitotic index. The extent to which these morphologic features reflect an underlying molecular similarity of the mouse and human tumors is not known. To compare across these species and to explore the mouse model more deeply, we carried out

Table 1 GSEA of gene sets upregulated and downregulated in KrasLA in human data sets

Human cancer phenotype data set	KrasLA model gene set			NNK carcinoma model gene set			NNK adenoma model gene set		
	ES	NES	FWER <i>P</i>	ES	NES	FWER <i>P</i>	ES	NES	FWER <i>P</i>
Upregulated									
Lung adenocarcinoma	0.102	1.880	0.041	0.128	1.531	0.421	0.042	-0.774	0.555
Pancreatic adenocarcinoma	0.127	1.574	0.367	0.088	1.750	0.226	0.052	1.090	0.445
Lung squamous cell carcinoma	0.073	1.560	0.373	0.186	1.790	0.129	-0.072	-0.913	0.555
Glioblastoma	0.127	1.330	0.443	0.090	0.910	0.445	0.046	0.575	0.445
Medulloblastoma	0.109	0.856	0.445	0.093	1.300	0.443	0.078	1.760	0.211
Renal cell carcinoma	0.053	1.210	0.445	0.065	1.190	0.445	-0.072	-0.994	0.555
Ovarian adenocarcinoma	0.072	0.701	0.445	0.065	-0.720	0.555	0.064	0.633	0.445
Lung carcinoid	-0.115	-1.310	0.554	0.117	1.150	0.445	-0.087	-1.250	0.555
Lung small-cell carcinoma	-0.104	-1.120	0.555	0.086	1.110	0.445	-0.054	-0.775	0.555
Breast adenocarcinoma	-0.076	-0.920	0.555	-0.089	-1.170	0.555	0.032	0.170	0.445
Prostate adenocarcinoma	0.052	-0.186	0.555	0.062	0.784	0.445	0.101	0.604	0.445
Bladder adenocarcinoma	-0.064	-0.811	0.555	-0.085	-1.310	0.554	0.086	0.923	0.445
Downregulated									
Lung adenocarcinoma	-0.270	-2.090	0.045	-0.239	-1.890	0.204	0.061	0.720	0.335
Lung squamous cell carcinoma	-0.250	-1.680	0.343	-0.259	-1.580	0.406	0.053	-0.610	0.499
Prostate adenocarcinoma	-0.190	-1.610	0.396	-0.136	-1.940	0.162	0.072	0.817	0.501
Lung carcinoid	-0.230	-1.370	0.493	-0.270	-1.440	0.483	-0.066	-0.863	0.499
Lung small-cell carcinoma	-0.199	-1.370	0.493	-0.226	-1.480	0.469	0.036	0.377	0.501
Bladder adenocarcinoma	-0.043	-1.380	0.493	-0.073	-0.944	0.499	0.038	0.704	0.501
Renal cell carcinoma	0.017	1.344	0.496	0.058	1.370	0.492	-0.049	1.000	0.501
Breast adenocarcinoma	-0.616	-1.330	0.499	-0.050	-0.891	0.499	-0.053	-0.674	0.499
Medulloblastoma	-0.070	-0.138	0.499	0.111	0.527	0.501	0.079	-0.568	0.499
Ovarian adenocarcinoma	0.112	-1.120	0.499	-0.136	-1.180	0.499	-0.061	-0.816	0.499
Pancreatic adenocarcinoma	-0.055	1.120	0.501	0.068	1.190	0.501	0.076	-0.552	0.499
Glioblastoma	-0.055	0.954	0.501	0.124	0.922	0.501	0.122	-1.470	0.472

Enrichment of genes that were upregulated or downregulated in the KrasLA mouse model and in the two NNK-induced models was analyzed by GSEA in the indicated human data sets. Positive ES scores indicate enrichment in the tumor class; negative ES scores indicate enrichment in the normal class ('anti-enrichment'). Results in bold are statistically significant. Only the upregulated gene set in the KrasLA model showed enrichment in human adenocarcinoma (positive ES score, significant FWER *P* value). Only the downregulated gene set in the KrasLA model showed anti-enrichment in human adenocarcinoma (negative ES score, significant FWER *P* value).

Table 2 NM analysis of KrasLA gene set in human cancer subtypes

Cancer subtype	NM correlation
Lung adenocarcinoma	0.18 ± 0.0040
Lung squamous cell carcinoma	0.15 ± 0.0060
Lung carcinoid	0.10880
Lung small-cell carcinoma	0.100840
Medulloblastoma	0.018480
Glioblastoma	-0.002730
Prostate adenocarcinoma	-0.006401
Renal cell carcinoma	-0.006401
Breast adenocarcinoma	-0.010303
Bladder adenocarcinoma	-0.010304
Pancreatic adenocarcinoma	-0.011027
Ovarian adenocarcinoma	-0.014214

NM correlations for the upregulated gene set in the KrasLA mouse model compared with human data sets. The highest correlation was for the comparison with human lung adenocarcinoma. Error ranges for the first two comparisons were calculated (Supplementary Methods online).

gene-expression profiling on mouse lung tumors ($n = 31$) and normal mouse lung samples ($n = 19$).

Gene-expression analysis using Affymetrix arrays showed a significant distinction between mouse lung tumors and normal lung. We used significance analysis of microarrays (SAM)¹⁴ to select two gene sets, one containing genes that were upregulated and the other containing genes that were downregulated in tumors relative to normal lung (617 and 510 genes, respectively; Supplementary Tables 1 and 2 online). We assessed the extent to which these mouse-derived gene sets were diagnostic of human lung cancer using a previously published data set of human lung cancer gene expression (the Boston data set¹⁵; details about data sets are available in Supplementary Methods online). To compare mouse and human expression profiles, we matched probes between microarrays. Using either the upregulated or the downregulated mouse gene set, we could classify 196 of 202 human samples (97% accuracy) in the Boston data set as either tumor or normal. Leave-one-out cross-validation on the human data alone was comparably accurate (200 of 202; ref. 16).

The accuracy of the cross-species comparisons of tumor and normal tissue suggested that the mouse model faithfully represented human lung cancer. But these expression changes might reflect nonspecific differences between tumors and normal lung (e.g., differences due to the expression of proliferation-associated genes). Therefore, we assessed the similarity of the sets of genes upregulated or downregulated in mouse lung tumor not only to human lung adenocarci-

noma but also to the other lung cancer subtypes in the Boston data set (small-cell lung cancer, lung carcinoid and squamous lung cancer). We also assessed the similarity of these mouse gene sets to eight other data sets from tissues other than lung that were extracted from the Global Cancer Map, a compendium of human cancer gene-expression profiles¹⁷. The simplest approach to this analysis would be to look for overlap in the corresponding data sets using a Venn diagram. But this approach is limited by the arbitrary cut-off of genes defined as differentially expressed and by the loss of information about the relative position of genes when comparing the two lists (Supplementary Methods and Supplementary Table 3 online). Therefore, we used Gene Set Enrichment Analysis (GSEA), a method for determining whether a rank-ordered list of genes for a particular comparison of interest (e.g., human lung adenocarcinoma versus normal lung) is enriched in genes derived from an independently generated gene set (e.g., mouse model-derived marker genes)^{18,19}. An application of the GSEA procedure is shown in Figure 1a. For a single gene set and data set with two phenotypes, GSEA provides an enrichment score (ES) that measures the degree of enrichment of the gene set at the top (highly correlated with class 1) or bottom (highly correlated with class 2) of a rank-ordered gene list derived from the data set. A nominal P value is used to assess the significance of the ES. The advantage of GSEA is that it takes into account the relative location of genes from the gene set of interest in a rank-ordered gene list. GSEA also allows for the comparative analysis of enrichment of several gene sets over multiple data sets (Fig. 1b). In this case, the normalized enrichment score (NES) is used and multiple comparison testing is accounted for by a family-wise error rate (FWER) P value. We used GSEA to compare mouse-generated gene sets with rank-ordered gene lists from human tumor and normal tissue, as well as rank-ordered gene lists from wild-type *KRAS2* and mutated *KRAS2* human lung adenocarcinoma, and extract meaningful information that was not readily apparent in either the mouse or the human data alone.

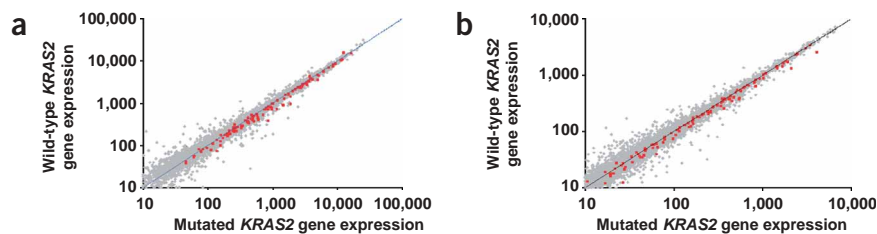
To assess the relative similarity of the KrasLA mouse model versus other published models of mouse lung cancer to human cancer, we used a gene-expression data set from mouse lung tumors induced by the tobacco-associated chemical mutagen 4-(*N*-nitrosomethylamino)-1-(3-pyridyl)-1-butanone (NNK)²⁰. In the NNK mouse model study, both early lesions (NNK adenomas) and late lesions (NNK carcinomas) were compared with normal lung. Using GSEA, we found that the upregulated and downregulated gene sets derived from the KrasLA mouse model were significantly enriched in human lung adenocarcinoma versus normal lung but not in the other lung cancer subtypes (carcinoid, small-cell lung cancer or squamous lung cancer) or any of the other nonlung cancer types. These findings indicate that the

Table 3 GSEA of three mouse models in human mutated *KRAS2* and wild-type *KRAS2* data sets

Human cancer phenotype data set	KrasLA model gene set			NNK carcinoma model gene set			NNK adenoma model gene set		
	ES	NES	FWER P	ES	NES	FWER P	ES	NES	FWER P
Upregulated									
Boston <i>KRAS2</i> data set	0.123	2.110	0.009	-0.034	-0.392	0.488	0.091	1.340	0.488
Ann Arbor <i>KRAS2</i> data set	0.183	1.760	0.057	0.159	1.611	0.231	0.076	1.097	0.353
Downregulated									
Boston <i>KRAS2</i> data set	0.105	1.072	0.295	0.063	0.594	0.481	0.040	0.539	0.485
Ann Arbor <i>KRAS2</i> data set	0.057	0.613	0.507	0.049	0.489	0.510	0.061	1.018	0.348

GSEA of the mouse upregulated and downregulated gene sets in a rank-ordered gene list derived from the human lung adenocarcinoma data sets using mutated *KRAS2* versus wild-type *KRAS2* as the phenotype of interest. The upregulated mouse gene set is enriched in human lung adenocarcinomas with *KRAS2* mutations. The downregulated mouse gene set shows no statistically significant enrichment. Results shown in bold are statistically significant.

Figure 2 *KRAS2* signature in two human data sets. Comparison of gene expression of all genes (gray) to *KRAS2* signature genes (red) in the Ann Arbor (a) and Boston (b) human lung cancer gene-expression profiles. Each data point represents the median value of that gene across all lung cancer samples that have either mutated *KRAS2* (x axis) or wild-type *KRAS2* (y axis). Plots are scaled to reflect expression value differences for the Affymetrix chips used in each data set.



mouse marker gene sets were not a nonspecific signature of cancer or rapidly proliferating cells (Table 1). In contrast, we did not detect consistently significant enrichment of the upregulated and down-regulated gene sets derived from either the NNK adenoma or the NNK carcinoma mouse models in the human data sets. Notably, the mouse lung tumors from the KrasLA mouse model and the carcinogen-induced models were not easily distinguishable on histologic criteria alone (Supplementary Figs. 1 and 2 online). Therefore, our analysis suggests that mouse tumor models that seem to be similar may be quite different at the level of gene expression. In the case described here, the KrasLA model is more similar to human lung adenocarcinoma than are the NNK-induced models. The extent to which this difference is due to different mechanisms of tumor induction (carcinogen exposure versus genetic activation of a specific oncogene) remains to be determined.

Another measure of similarity is the correlation of 'gene neighborhoods' between two data sets. A gene neighborhood is defined as the pairwise correlation matrix of a gene set for a given data set. In this case, the gene set is determined by the genes most differentially expressed in the mouse model tumor compared with normal lung. We computed the correlation of this pairwise matrix with the pairwise matrix of the same gene set in the human data sets (the Neighborhood Mantel (NM) procedure). Using this procedure, we again found that the upregulated gene set derived from the KrasLA mouse model was more correlated with human lung adenocarcinoma than any other data set (Table 2). Taken together, the results of GSEA and NM analysis indicate that the KrasLA model is more closely related to human lung adenocarcinoma than to any other tumor type evaluated.

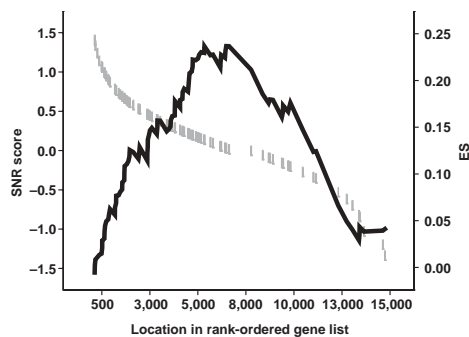


Figure 3 The *KRAS2* signature is enriched in pancreatic adenocarcinoma. The left y axis shows SNR scores (gray bars) for the *KRAS2* signature gene set in a rank-ordered gene list derived from a pancreatic adenocarcinoma versus normal pancreas data set. The right y axis shows ES scores (black line) for the *KRAS2* signature gene set on the rank-ordered gene list. The x axis indicates the location of the *KRAS2* signature genes in the rank-ordered gene list derived from a pancreatic adenocarcinoma versus normal pancreas data set.

The mouse-human gene-expression comparison analysis allowed us to identify a subset of genes that are upregulated in both human and mouse adenocarcinoma (Supplementary Table 4 online). Some of these commonalities may reflect similarity in the cell of origin, but many are probably specifically associated with adenocarcinoma oncogenesis. For example, both the mouse and the human adenocarcinomas had high levels of expression of *RAP1GAI* (RAP1 GTPase activating protein 1), suggesting that downregulation of RAP1 is beneficial for tumorigenesis. In addition, both mouse and human adenocarcinomas showed upregulation of *TGIF*, a corepressor of Smad proteins that is phosphorylated through the MAPK pathway²¹. This protein is known to have a role in the severe developmental abnormality holoprosencephaly; our results suggest that it also has a role in tumorigenesis, possibly by modulation of TGF- β signaling in the lung epithelium.

Identification of an oncogenic *KRAS2* signature

Beyond validating a model, comparing mouse and human gene-expression data sets can also identify expression signatures of specific oncogenes or tumor suppressors that may be obscured by direct analysis of either a human or a mouse cancer data set in isolation. Among the 94 human adenocarcinomas for which *KRAS2* status was obtained in the Boston data set, 34 carried oncogenic mutations in *KRAS2*. Of the 12,588 genes on the human Affymetrix arrays used to analyze those tumors, the subset associated with the wild-type versus mutant phenotype is no larger than expected by chance and, therefore, is not statistically significant (Supplementary Methods online). Similarly, no statistically significant gene-expression correlate of *KRAS2* mutation status was identifiable in the Ann Arbor series of wild-type 46 *Kras2* and 39 mutated *KRAS2* tumors analyzed²². These results indicate that the human data, examined alone, do not support the existence of a gene-expression signature that reflects *KRAS2* mutation. It is possible that some adenocarcinomas carry *KRAS2* mutations

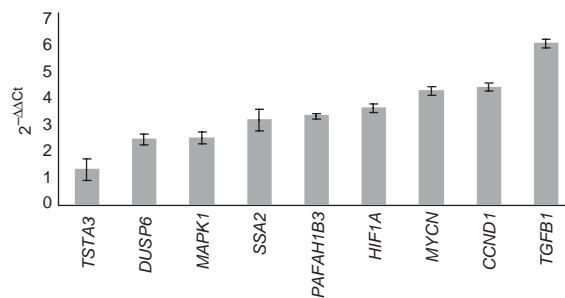


Figure 4 Real-time PCR analysis of expression of selected *KRAS2* signature genes. Comparison between human lung cancer cell lines with mutated *KRAS2* (A549) or wild-type *KRAS2* (H1650). $2^{-\Delta\Delta Ct}$ values were calculated using expression of *TBP* as a reference.

whereas others have mutations elsewhere in the Ras pathway (e.g., in *BRAF*²³ or *EGFR*²⁴), thereby yielding an indistinguishable program of gene expression that represents a common final pathway of transformation. Alternatively, a gene-expression signature of *KRAS2* activation may exist, but the analytical tools applied in the earlier studies were insufficient to detect it.

To explore this question, we examined whether the pronounced gene-expression profile of the *KrasLA* mouse lung tumors compared with that of normal lung tissue contained a set of genes whose expression reflected the *Kras2* mutation status of these tumors and that might, therefore, be selectively affected in human lung cancer specimens with such a mutation. We used GSEA to examine the location of the upregulated and downregulated gene sets derived from the *KrasLA* mouse model in rank-ordered gene lists constructed from the human mutated *KRAS2* versus wild-type *KRAS2* adenocarcinoma distinction in both the Ann Arbor and Boston data sets. Notably, the mouse-derived upregulated gene set was enriched in the mutated *KRAS2* adenocarcinomas of both of these human data sets (Ann Arbor: NES = 1.76, FWER $P = 0.057$; Boston: NES = 2.11, FWER $P = 0.009$; **Table 3** and **Supplementary Table 5** and **Supplementary Figs. 3** and **4** online). The downregulated mouse gene set did not achieve statistical significance in either the Ann Arbor or Boston data sets (**Table 3**). This difference may mean that *Kras2* has stronger influence on upregulation than downregulation of gene expression. Alternatively, it may be due to the fact that the downregulated mouse gene set contains many genes highly expressed in normal tissue that are not expressed at substantial levels in the human tumors and are therefore not prominent in a rank-ordered list comparing mutated *Kras2* versus wild-type *Kras2* tumors. Nevertheless, the enrichment of the upregulated mouse gene set in human lung adenocarcinomas with *KRAS2* mutation argues that a *KRAS2* signature does exist in the human tumors but the relatively homogeneous samples derived from a genetically driven mouse model were needed to act as a filter with which to extract it.

We defined the *KRAS2* signature as those genes that contributed maximally to the GSEA score in both human data sets (89 genes; **Supplementary Table 6** online). The final *KRAS2* signature is shown in **Figure 2**, which shows that each of the *KRAS2* signature genes is, on average, only modestly differentially expressed in the human tumors for both the Ann Arbor and Boston data sets. As such, on a gene-by-gene basis, none of the individual genes meets statistical significance. Considered as a collection of genes (a 'program'), however, the

coordinate regulation of the entire set achieves statistical significance ($P < 0.05$) and is expected to be biologically important.

KRAS2 signature in other gene-expression data sets

The *KRAS2* signature identified by comparing mouse and human can be validated at two levels. First, as a 'pattern' associated with *KRAS2* mutation, the ability of this signature to identify biologically or clinically relevant characteristics of other human tumors could be tested. Second, a subset of the genes in the *KRAS2* signature could be directly linked to the transcriptional consequences of oncogenic *KRAS2* signaling.

If the signature we identified reflects *KRAS2* biology and is not specific to lung cancer, its enrichment might be expected in other cancers with frequent *KRAS2* mutation. To explore this possibility, we next asked whether the *KRAS2* signature was enriched across a range of cancer subclasses. These data sets represented independent samples not used in any of the analysis thus far and included many tumors in which *KRAS2* is not known to be frequently mutated (**Supplementary Methods** online). GSEA was significant using the *KRAS2* signature as a gene set and a rank-ordered gene list constructed from a human pancreatic cancer versus normal pancreas data set²⁵ (NES = 3.01, nominal $P < 0.001$; **Fig. 3** and **Supplementary Table 7** online). Among all human cancers, *KRAS2* mutation is most common in pancreatic adenocarcinomas, with a prevalence of >90% (ref. 26). Therefore, this result further supports the link between the level of expression of *KRAS2* signature genes and *KRAS2* mutation status. Notably, the Global Cancer Map pancreatic cancer data set did not have a significant FWER P value using only the mouse *KrasLA* gene sets (**Table 1**). Using the human and mouse combined *KRAS2* signature, however, a significant P value was obtained in both pancreatic cancer data sets examined (**Supplementary Table 7** online), underscoring the need for the mouse-human comparison to identify the *KRAS2* signature.

To determine whether the *KRAS2* signature included genes whose expression would be affected by the level of signaling downstream of *KRAS2*, we analyzed human lung cancer cell lines whose *KRAS2* mutation status was known. First, we used quantitative RT-PCR to measure the expression level of selected genes chosen from the signature list in the A549 (mutated *KRAS2*) cell line compared with the H1650 (wild-type *KRAS2*) cell line. Consistent with our previous analysis, expression of eight of nine genes from the *KRAS2* signature was more than two times higher in the A549 cells than in the H1650

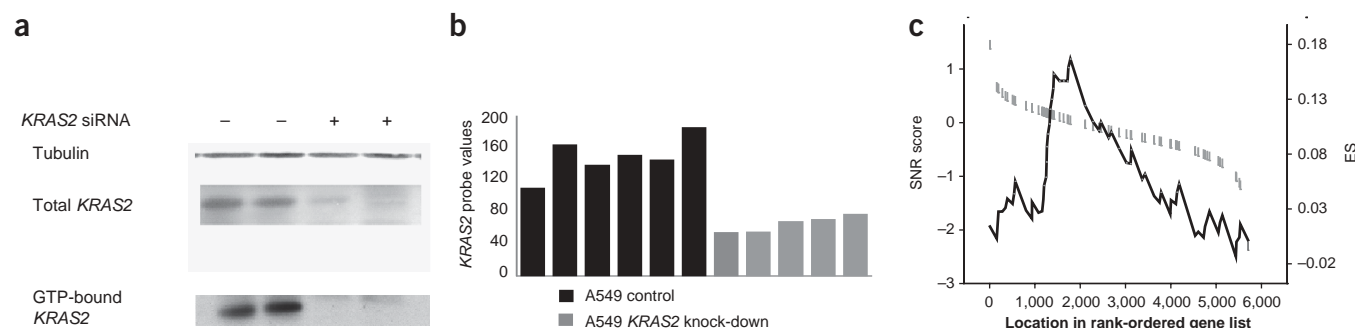


Figure 5 Knock-down of *KRAS2* in the human lung cancer cell line A549. **(a)** Western blot of levels of total *KRAS2* and GTP-bound *KRAS2* in A549 parental and 2 A549 knock-down cell lines. **(b)** Average expression level for the two *KRAS2* probes present in the U133A Affymetrix array for the 11 samples analyzed. **(c)** The left y axis shows SNR scores (gray bars) for the *KRAS2* signature gene set in a rank-ordered gene list derived from the A549 versus A549 knock-down data set. The right y axis shows ES scores (black line) for the *KRAS2* signature gene set on the rank-ordered gene list. The x axis indicates the location of the *KRAS2* signature genes in the rank-ordered gene list derived from the A549 versus A549 knock-down data set.

cells (Fig. 4). To verify whether *KRAS2* signature genes were influenced by the level of signaling downstream of *KRAS2*, we used a short hairpin RNA (shRNA) vector to knock-down expression of *KRAS2* in A549 cells. *KRAS2* knock-down was evident in independent cell populations in terms of total cellular *KRAS2* and *KRAS2*-GTP levels (Fig. 5a). To determine whether the expression of *KRAS2* signature genes was significantly affected by the inhibition of oncogenic *KRAS2*, we carried out expression profiling and then GSEA analysis on six independent samples from parental A549 cells and five independent A549 cells infected with *KRAS2* knock-down shRNA. We also verified substantial knock-down of *KRAS2* by extracting the expression levels measured on the corresponding probes on the microarray (Fig. 5b). The *KRAS2* signature as a whole was enriched in A549 control relative to A549 knock-down cells as determined by the GSEA procedure ($ES = 0.109$, nominal $P = 0.025$; Fig. 5c). These data confirm that the expression levels of a substantial number of genes in the *KRAS2* signature are influenced by signaling downstream of *KRAS2*. Other genes whose expression was not decreased (Supplementary Table 8 online) in the knock-down may nevertheless still be important for *KRAS2*-induced oncogenesis. For example, increased expression of these genes may be generally beneficial for *KRAS2*-transformed cells, despite the fact *KRAS2* mutation does not directly influence their expression.

DISCUSSION

The continued association of a substantial number of the *KRAS2* signature genes with oncogenic *KRAS2* status using these independent assays and experimental conditions lends further credence to the idea that these genes have a role in controlling aspects of *KRAS2*-mediated transformation. Examination of the *KRAS2* signature uncovers several connections to *KRAS2* biology. For example, levels of cyclin D1 are increased in cells overexpressing oncogenic *KRAS2* (ref. 27). Other genes on the list were known to be regulated by the MAPK-ERK pathway but had not been specifically linked to oncogenic *KRAS2*-induced transformation. For example, MKP3 (*DUSP6*) is a dual-specific phosphatase that specifically inhibits ERK signaling^{28,29}. MKP3 was recently shown to be a critical regulator of an FGF8-mediated signaling cascade in the developing limb bud by decreasing ERK signaling that would otherwise lead to apoptosis³⁰. The importance of limiting ERK signaling in Ras-induced tumorigenesis is suggested by the facts that overexpression of Ras *in vitro* causes cellular senescence and that this effect is mediated by the MAPK pathway³¹. Activation of single-copy oncogenic *Kras2* in mouse embryonic fibroblasts, however, does not lead to senescence but rather to hyperproliferation without evidence for increased MAPK signaling⁹. The expression of MKP3 may be involved in modulating ERK activity in *KRAS2*-initiated oncogenesis and could explain why activation of the MAPK pathway is not observed in *KRAS2*-induced lung tumors. The gene *PHLDA1* (also called *TDAG51*) is implicated as a mediator of prosurvival signals downstream of IGF-1 (ref. 32). The A549 knock-down experiment showed that the expression level of this gene was considerably influenced by signaling downstream of *KRAS2*.

In this report, we describe a general approach to the integrative genomic analysis of mouse models of cancer and the human tumors they are intended to represent. We show that in the case of lung cancer, such data integration can both validate the fidelity of the mouse model and extract from the human samples evidence of an oncogene-specific gene-expression program that was not otherwise apparent. The pattern of highly expressed genes in a mouse model of prostate cancer driven by *Myc* overlaps that in human prostate cancer³³. The approach described here provides a framework for the comparison of several models against several human cancer types.

Additionally, our approach allowed us to identify many potentially new effectors of *KRAS2*-induced transformation and underscores the value of mouse modeling in dissecting the role of oncogenes in tumorigenesis.

Lung adenocarcinoma strikes 170,000 individuals per year in the US, most of whom die of their disease. Although *KRAS2* has been known for many years to be mutated in these individuals, the full range of molecular consequences of *KRAS2* activation *in vivo* is not known. The availability of a validated mouse model, together with the identification of a *KRAS2* activation signature in human tumors, should help in the development of anti-Ras pathway therapeutic strategies. The approaches described here represent a powerful genomics-based strategy that can simultaneously assess the similarities between mouse models of human disease and the disease itself and provide a means to elucidate crucial genes affected in the disease process. With the anticipated increase of mouse and other cancer model gene-expression data sets, the methods described here could be applied more broadly. In the case of the KrasLA model, it will now be possible to undertake proof-of-principle experiments to test the role of individual components of the *KRAS2* signature in *KRAS2*-induced oncogenesis using gene-targeting or shRNA-based approaches in the mouse.

METHODS

Mouse genotyping and tumor isolation. We crossed KrasLA2 mice on a 129svJae background with wild-type C57B6 mice to obtain F₁ progeny, which we genotyped as previously described¹³. (Here, these mice are called KrasLA mice for simplicity. See ref. 13 for details of the characteristics of KrasLA1 and KrasLA2 mice.) Mice bearing the *Kras2* latent allele were allowed to develop to 5–6 months of age and then killed by cervical dislocation. We removed their lungs and placed them in RNAlater solution (Ambion). We removed individual tumors large enough to be easily dissected (3–8 mm) and cut them into two pieces. We placed one piece in formalin to be used for histological analysis and stored the other piece at -80°C for extraction of RNA and DNA. All experiments described in this report were approved by the Committee on Animal Care at the Massachusetts Institute of Technology and by the Animal Care and Use Committee at the Dana Farber Cancer Institute.

RNA isolation and preparation for microarray analysis. We placed mouse tumor fragments in Trizol reagent and homogenized them by using first a Kontes disposable pestle and then a polytron homogenizer. We extracted RNA and DNA from Trizol using the manufacturer's instructions. We further purified RNA by using a Qiagen RNA column. We assessed quality of RNA by gel electrophoresis. We then used samples with high-quality RNA to prepare cRNA for hybridization to Affymetrix MG_U74Av2 oligonucleotide arrays as previously described³⁴.

PCR analysis of tumor DNA. We designed PCR primers to amplify the *Kras2*-neomycin cassette boundary found in germline DNA of the KrasLA mice so that the 5' primer was located inside the *Kras2* intronic DNA whereas the 3' primer was located at the 5' end of the neomycin cassette (primer sequences available on request). We quantified PCR amplification using SYBR green detection on an ABI 7000 sequence detection apparatus. We used a TaqMan probe for 18S ribosomal RNA as an internal control. We determined the percent of germline *Kras2* latent DNA (nontumor DNA) by comparing the ΔCt value ($\text{Ct}_{\text{Kras2-neo}} - \text{Ct}_{18\text{S}}$) to a standard curve of ΔCt values in which wild-type DNA (equivalent to tumor DNA in that there should be no *Kras2-neo* amplicon) was mixed with known amounts of spleen DNA from *Kras2* latent mice. We used only those samples in which the Trizol-extracted tumor DNA could be shown to have less than a 30% contribution of nontumor cells for microarray analysis.

Cell lines. We obtained A549 and H1650 cell lines from the American Type Culture Collection. We grew cell lines in accordance with the manufacturer's recommendations. For the validation of the *Kras2* signature, we selected a subset of genes and designed primers to span exons using primer express

software. We carried out real-time PCR analysis on an ABI 7000 sequence detection apparatus using SYBR green. Primer sequences are available on request. We grew cell lines to 70% confluency and extracted RNA as described above for tumor RNA, except that we added an on-column DNase step (Qiagen) to ensure that there was no carry-over of DNA into the PCR reaction. We then reverse-transcribed RNA into cDNA using random hexamers and used the cDNA as template for the PCR reaction.

We cloned an shRNA directed against the serine mutation at codon 12 of human *KRAS2* into the lentilox 3.7 vector³⁵ (sequence available on request). We transfected a host strain with this plasmid for virus production using 293FT cells as previously described³⁵. We selected cells carrying the shRNA or an empty vector control by fluorescence-activated cell sorting for green fluorescent protein 5 d after infection. We replated cells at a density of 2×10^6 cells per 10-cm plate. Forty-eight hours after replating, we extracted RNA as described above for tumor tissue. We then used samples with high-quality RNA as determined by capillary electrophoresis on an Agilent BioAnalyzer to prepare cRNA for hybridization to U133A human Affymetrix GeneChip oligonucleotide arrays.

Western blotting. We lysed cells with 1% SDS boiling lysis buffer, clarified the lysates, normalized them for protein levels and analyzed them by western blotting in 5% bovine serum albumin using a Kras2-specific antibody (Santa Cruz, F-234). We recovered total Ras-GTP from A549 cells using agarose-conjugated Raf-GST in accordance with the manufacturer's recommendations (Upstate). We detected Kras2 from the total Ras-GTP by western blotting as above.

Human tumor gene-expression databases. We used gene-expression data from the Global Cancer Map (190 specimens; 16,063 genes; Affymetrix GeneChip Hu6800 and Hu35KsubA) and two human lung cancer data sets (Boston: 15 normal lung samples, 144 adenocarcinoma samples, 5 small-cell lung cancer cells, 20 carcinoid cells and 20 squamous cells; 12,588 genes; Affymetrix GeneChip HG_U95Av2; and Ann Arbor; 86 primary lung adenocarcinomas and 10 normal lung samples; Affymetrix GeneChip Hu6800). We also used other published and unpublished data sets to evaluate the significance of the *KRAS2* signature beyond lung cancer. See **Supplementary Methods** online for details of these data sets and URLs for the published data.

Matching of human and mouse probes. To compare expression data from the mouse and human data sets, a correspondence must be made between probes on the mouse arrays and probes on the human arrays. We obtained mapped probe sets from Affymetrix. See **Supplementary Methods** online for details of this mapping.

Statistical and computational tools. We used the following statistical and computational tools: SAM¹⁴, signal-to-noise ratio (SNR), GSEA¹⁹, NM and support vector machines¹⁶. A detailed description of these algorithms, except SAM, is provided in **Supplementary Methods** online.

We used SAM for marker selection in constructing gene sets from the KrasLA mouse data set. To ensure that the gene sets selected were representative and did not yield optimistic *P* values when used in GSEA, we selected alternative parameters in SAM for the gene-set selection by varying the tuning (Δ) and relative change thresholds to cover a range of false discovery rates of 1–25%. Our analysis uses a relative change of 1.5, which corresponds to a conservative false discovery rate of <5%. At this threshold level, 617 genes were upregulated and 510 genes were downregulated in KrasLA. See **Supplementary Table 5** online for results of GSEA using alternative cut-offs for the gene set in the mutated *KRAS2* versus wild-type *KRAS2* distinction.

We used the SNR metric to produce rank-ordered gene lists in the human data sets (rank-ordering a data set is important for GSEA, as described below). Given a data set, we identified genes correlated with a particular class distinction (e.g., tumor and normal) by sorting all genes on the array according to the SNR statistic

$$S_i = \frac{\mu_{i,\text{class } 0} - \mu_{i,\text{class } 1}}{\sigma_{i,\text{class } 0} - \sigma_{i,\text{class } 1}},$$

where $\mu_{i,\text{class}}$ and $\sigma_{i,\text{class}}$ represent the mean (or median) and standard deviation of expression, respectively, for each class for the *i*th gene.

GSEA provides a general statistical method to test for the enrichment of sets of genes in expression data sets. The procedure takes as input data sets $\Delta = \{D_1, \dots, D_K\}$, gene sets $\Gamma = \{G_1, \dots, G_M\}$ and a ranking procedure (the SNR score is used as a ranking procedure in our analysis). The procedure outputs an ES and a *P* value for each data set–gene set pair. To correct for multiple hypothesis testing, in addition to nominal *P* values, we output a *P* value that controls the FWER. We first describe GSEA when there is one data set *D* and gene set *G*. The genes in the data set are rank-ordered to form a gene list $L = \{g_{(1)}, \dots, g_{(N)}\}$. The ES is the maximum deviation of the empirical distribution functions $P_{\text{hit}}(G, D, R, i)$ and $P_{\text{miss}}(G, D, R, i)$ (equation 1 given below) to represent the fraction of genes in that are present (hits) or absent (misses) in the ordered gene list up to a given position *i*.

In these functions,

$$P_{\text{hit}}(G, D, R, i) = \frac{\text{Card}[g_{(j \leq i)} \in G]}{N_H}, \quad P_{\text{miss}}(G, D, R, i) = \frac{\text{Card}[g_{(j \leq i)} \notin G]}{N - N_H} \quad (1)$$

where N_H is the number of genes in the gene set, $N - N_H$ is the number of genes not in the gene set, $\text{Card}[g_{(j \leq i)} \in G]$ is the number of genes (cardinality) ranked above the *i*th gene that are in the gene set and $\text{Card}[g_{(j \leq i)} \notin G]$ is the number of genes ranked above the *i*th gene that are not in the gene set. We next define a 'running' ES as the difference between the empirical distribution functions:

$$RES(G, D, R, i) = P_{\text{hit}}(G, D, R, i) - P_{\text{miss}}(G, D, R, i). \quad (2)$$

The enrichment score ES (*G, D, R*) is the maximum deviation from zero of the above difference:

$$ES(G, D, R) = RES(G, D, R, \arg \max_{i=1, \dots, N} |RES(G, D, R, i)|). \quad (3)$$

The significance of the above enrichment score is assessed by permutation testing. For the case of multiple gene sets and data sets, we control the FWER of the multiple comparisons (each gene set–data set pair). The distribution under the null hypothesis for each comparison has great variation, and so the ES of each comparison must be normalized to common scale (in the same way a *z* score serves as a normalization for the *t* statistic). This normalized score is called the NES (**Supplementary Methods** online). For the case of multiple comparisons, we report the ES, NES and corresponding FWER *P* value. Otherwise, the ES and nominal *P* value are reported. We carried out 1,000 random permutations to build the null distribution.

We derived the mouse gene sets used in our analysis using a defined cut-off from a rank-ordered list of genes. Because the cut-off is arbitrary, we analyzed the significance of the ES obtained using the mouse gene set compared with other gene sets that were constructed by random sampling or permutation. In the sampling procedure, we constructed random gene sets by sampling a subset of the mouse upregulated tumor markers. We could then compute the ESs for these random gene sets in the human data sets (**Supplementary Methods** and **Supplementary Fig. 3** online). In the permutation procedure, we constructed random gene sets by permuting the phenotypic labels in the mouse and then recomputing the rank-ordered gene list. We then extracted gene sets from these rank-ordered lists using the same number of genes as was in the original gene set. For each gene set, we computed the ES using the wild-type *KRAS2* versus mutated *KRAS2* phenotype in the Ann Arbor and Boston data sets. We could then compute the significance of the ES obtained from the original gene set from the empirical distribution of the enrichment scores from the random gene sets (**Supplementary Methods** and **Supplementary Fig. 4** online).

The NM score is a measure of the correlation of gene neighborhoods between a reference data set and a comparison data set. For the genes most correlated with a phenotypic label of interest in the reference data set, we computed a matrix of pairwise gene-expression correlations (Pearson correlation coefficients):

$$M_{ij}^R = \frac{\langle g_i, g_j \rangle}{\|g_i\| \|g_j\|},$$

where g_i is the expression profile of the *i*th gene over the samples in the reference data set and $i, j \in \{1, \dots, L\}$. The matrix of pair-wise correlations, M^C , of the same genes is computed in the comparison data set. The classical

Mantel test statistic³⁶ is used to compute the NM score:

$$\frac{1}{L^2 - L - 1} \sum_{\substack{i,j=1 \\ i \neq j}}^L \frac{(M_{ij}^R - \mu_R)(M_{ij}^C - \mu_C)}{\sigma_R \sigma_C}, \quad (4)$$

where μ_R , μ_C , σ_R and σ_C are the means and standard deviations, respectively, of the matrices of pairwise correlations in the reference and comparison data sets, respectively (**Supplementary Methods** online).

We used support vector machines to classify samples (**Supplementary Methods** online).

URL. All data sets described in this report can be accessed at http://web.mit.edu/ccr/labs/jacks/sweetcordero_et_al/.

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

ACKNOWLEDGMENTS

We thank P. Tamayo and K. Haigis for comments and critical review of the manuscript and M. You for providing access to the gene expression data and histology slides for the NNK mouse models. This work was supported in part by the National Institutes of Health and the National Cancer Institute. T.J. and T.R.G. are investigators of the Howard Hughes Medical Institute. A.S.-C. was supported in part by grants from the Robert Wood Johnson Foundation (Harold Amos Medical Faculty Development Program) and by a mentored clinical scientist grant from the National Cancer Institute. S.M. received partial support from an Alfred P. Sloan Foundation/U.S. Department of Energy Fellowship in Computational Molecular Biology.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 2 August; accepted 22 November 2004

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

- van 't Veer, L.J. *et al.* Expression profiling predicts outcome in breast cancer. *Breast Cancer Res* **5**, 57–58 (2002).
- 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).
- 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).
- Armstrong, S.A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30**, 41–47 (2002).
- Ferrando, A.A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75–87 (2002).
- Ross, M.E. *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102**, 2951–2959 (2003).
- 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).
- Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* **34**, 226–230 (2003).
- Tuveson, D.A. *et al.* Endogenous oncogenic K-ras(G12D) stimulates proliferation and widespread neoplastic and developmental defects. *Cancer Cell* **5**, 375–387 (2004).
- Repasky, G.A., Chenette, E.J. & Der, C.J. Renewing the conspiracy theory debate: does Raf function alone to mediate Ras oncogenesis? *Trends Cell Biol.* **14**, 639–647 (2004).
- Zuber, J. *et al.* A genome-wide survey of RAS transformation targets. *Nat. Genet.* **24**, 144–152 (2000).
- Tchernitsa, O.I. *et al.* Transcriptional basis of KRAS oncogene-mediated cellular transformation in ovarian epithelial cells. *Oncogene* **23**, 4536–4555 (2004).
- Johnson, L. *et al.* Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. *Nature* **410**, 1111–1116 (2001).
- Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
- Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
- Vapnik, V.N. *Statistical Learning Theory* (John Wiley and Sons, New York, 1998).
- Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **98**, 15149–15154 (2001).
- 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).
- Mootha, V.K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
- Bonner, A.E., Lemon, W.J., Devereux, T.R., Lubet, R.A. & You, M. Molecular profiling of mouse lung tumors: association with tumor progression, lung development, and human lung adenocarcinomas. *Oncogene* **23**, 1166–1176 (2004).
- Lo, R.S., Wotton, D. & Massague, J. Epidermal growth factor signaling via Ras controls the Smad transcriptional co-repressor TGIF. *EMBO J.* **20**, 128–136 (2001).
- Beer, D.G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8**, 816–824 (2002).
- Naoki, K., Chen, T.H., Richards, W.G., Sugarbaker, D.J. & Meyerson, M. Missense mutations of the BRAF gene in human lung adenocarcinoma. *Cancer Res.* **62**, 7001–7003 (2002).
- Paez, J.G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
- Iacobuzio-Donahue, C.A. *et al.* Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol.* **162**, 1151–1162 (2003).
- Klimstra, D.S. & Longnecker, D.S. K-ras mutations in pancreatic ductal proliferative lesions. *Am. J. Pathol.* **145**, 1547–1550 (1994).
- Peeper, D.S. *et al.* Ras signalling linked to the cell-cycle machinery by the retinoblastoma protein. *Nature* **386**, 177–181 (1997).
- Camps, M. *et al.* Induction of the mitogen-activated protein kinase phosphatase MKP3 by nerve growth factor in differentiating PC12. *FEBS Lett.* **425**, 271–276 (1998).
- Muda, M. *et al.* The dual specificity phosphatases M3/6 and MKP-3 are highly selective for inactivation of distinct mitogen-activated protein kinases. *J. Biol. Chem.* **271**, 27205–27208 (1996).
- Kawakami, Y. *et al.* MKP3 mediates the cellular response to FGF8 signalling in the vertebrate limb. *Nat. Cell Biol.* **5**, 513–519 (2003).
- Serrano, M., Lin, A.W., McCurrach, M.E., Beach, D. & Lowe, S.W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
- Toyoshima, Y. *et al.* TDAG51 mediates the effects of insulin-like growth factor I (IGF-I) on cell survival. *J. Biol. Chem.* **279**, 25898–25904 (2004).
- Ellwood-Yen, K. *et al.* Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* **4**, 223–238 (2003).
- Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Rubinson, D.A. *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nat. Genet.* **33**, 401–406 (2003).
- Mantel, N. The detection of disease clustering and a generalized regression approach **27**, 209–220 (1967).