[CANCER RESEARCH 62, 3005–3008, June 1, 2002]

Advances in Brief

Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival[1]

Dennis A. Wigle, Igor Jurisica, Niki Radulovich, Melania Pintilie, Janet Rossant, Ni Liu, Chao Lu, James Woodgett, Isolde Seiden, Michael Johnston, Shaf Keshavjee, Gail Darling, Timothy Winton, Bobby-Joe Breitkreutz, Paul Jorgenson, Mike Tyers, Frances A. Shepherd, and Ming Sound Tsao[2]

Thoracic Oncology Site Group, Princess Margaret Hospital [D. A. W., M. J., S. K., G. D., T. W., F. A. S., M. S. T.], Ontario Cancer Institute [I. J., N. R., M. P., N. L., C. L., J. W., I. S., M. S. T.], Samuel Lunenfeld Research Institute of Mount Sinai Hospital [D. A. W., J. R., B-J. B., P. J., M. T.], and University of Toronto, Toronto, Ontario, Canada M5G 2M9

Abstract

Recent studies have suggested that information from gene expression profiles could be used to develop molecular classifications of cancer. We hypothesized that expression levels of specific genes in operative specimens could be correlated to recurrence risk in non-small cell lung cancer (NSCLC). We performed expression profiling using 19.2 K cDNA microarrays on tumor specimens from a total of 39 NSCLC patients with known clinical follow-up information. Statistical analysis and clustering approaches were used to determine patterns of gene expression segregating with clinical outcome. The results provide evidence that molecular subtyping of NSCLC can identify distinct profiles of gene expression correlating with disease-free survival.

Introduction

Lung cancer continues to be the most common cause of cancer-related mortality in both men and women in North America (1). It accounts for ~30% of all cancer deaths, a total greater than that from the next three cancers (breast, colon, and prostate) combined. The current staging system for lung cancer has remained largely unchanged for >30 years, and continues to be based on histopathology and extent of disease at presentation (2). However, these classification systems alone have reached their limit in providing critical information that may influence management strategy. The treatments of lung cancer are primarily based on the broad classification of tumors into small cell and non-small cell types, and histological subtyping of the latter does not play a significant role either in prognosis or therapeutic options. The heterogeneity of lung cancer patients at each disease stage with respect to outcome and treatment response suggests that additional subclassification and substaging remains possible. Molecular heterogeneity within individual lung cancer diagnostic categories is evident in the variable presence of specific mutations, deletions of tumor suppressor genes, and numerous chromosomal abnormalities found to date (3). Reports that some of these genetic aberrations are prognostic factors for NSCLC[3] patients provide evidence that additional information on risk of relapse or death from cancer may be defined at molecular levels (4). Thus, correlations of molecular profiles from individual tumor samples to clinical outcome data hold the promise of better classification of lung cancer, and subsequently improved diagnostic and prognostic information for patient management (5–10). In this study, we provide evidence that information regarding DFS or relapse risk for NSCLC can be obtained in gene expression data from cDNA microarrays.

Materials and Methods

Patient Samples. NSCLC samples were obtained from lobectomy or pneumonectomy specimens within 30 min of resection, and samples were snap-frozen and stored in liquid nitrogen until use. Samples were collected using informed consent according to institutional guidelines and the study protocol approved by our institutional Human Tissue Committee and Research Ethics Board. Details of clinical specimens are described in Table 1. Frozen sections were performed on all of the specimens to verify the quality of tumor samples, as well as estimating the percentage of tumor cells, extent of inflammatory infiltrate, and the differentiation grade of tumor cells.[4]

cDNA Microarray Experiments. All of the samples were referenced to an RNA pool derived from the immortalized human bronchial epithelial cell line HBE135-E6E7 (11). Microarrays were printed on glass slides from a human 19.2 K clone set. Total RNA (10–20 μg) was added to 8 μl of 5× first strand buffer, 0.75 μl of 200-μM AncT primer (5′ T20VN 3′), and 4 μl of 0.1 M DTT. The reaction mixture was heated to 65°C for 5 min and then placed at 42°C for 5 min. One μl of 20 mM dGTP, dCTP, and dATP mix, 1.3 μl of 5-mM dTTP, 1.35 μl of 10-mM amino-allyl dUTP (Sigma, Oakville, Ontario, Canada), and 2 μl of 200 units/μl Superscript II reverse transcriptase (Life Technologies, Inc., Burlington, Ontario, Canada) were added to each RNA sample to a total volume of 40 μl. Reverse transcription was performed at 42°C for 2 h. RNA was hydrolyzed by adding 4 μl of 50 mM EDTA and 2 μl of 10 N NaOH, and incubation at 65°C for 20 min. Samples were then neutralized with 4 μl of 5 M acetic acid, and reactions purified using 30-μm columns (Millipore, Bedford, MA). Samples were dried using a Speedvac concentrator and resuspended in 9 μl of 0.1 M sodium bicarbonate (pH 9.0) for dye labeling. Cy3 and Cy5 dyes (Pharmacia, Piscataway, NJ) were resuspended in 18 μl DMSO and 0.625 μl of dye added to each 4.5 μl aliquot of cDNA. Coupling was performed at room temperature in darkness for 45 min. Reactions were quenched by addition of 2.25 μl of 4 M hydroxylamine and incubated for 15 min at room temperature in the dark. Cy3 and Cy5 reactions were then combined for each set of reference and test samples, and unincorporated nucleotides removed using a High Pure PCR purification kit (Roche, Laval, Quebec, Canada). Probes were dried using a Speedvac concentrator before resuspension. Arrays were prehybridized for 1 h at 37°C in 100 μl of DIGeasy hybridization buffer (Roche) mixed with 5 μl of 10 mg/ml yeast tRNA, 5 μl of 10 mg/ml salmon sperm DNA, and 5 μl of 10% BSA in humid hybridization chambers. Slides were washed with double deionized water and then dried with isopropanol. Probes were resuspended in 100 μl of

Received 12/28/01; accepted 4/15/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by grants from the Canadian Cancer Society and the National Cancer Institute of Canada (#012150), the National Science and Engineering Research Council of Canada (#203833-98), the Physicians' Services Incorporated Foundation (R00-10), and the Princess Margaret Hospital Microarray Clinical Research Program. D. A. W. is a CIHR postdoctoral fellow. I. J. is supported by an IBM Shared University Research Grant and an IBM Faculty Partnership Award. J. R. is a CIHR Distinguished Investigator.

[2] To whom requests for reprints should be addressed, at Ontario Cancer Institute and Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario, Canada M5G 2M9. Phone: (416) 946-4426; Fax: (416) 946-6579; E-mail: Ming.Tsao@uhn.on.ca.

[3] The abbreviations used are: NSCLC, non-small cell lung cancer; DFS, disease-free survival.

[4] Internet address: http://www.cs.utoronto.ca/~juris/publicationsData.html.

Table 1 *The clinical and pathological characteristics of patients and their tumors*

	Early recurrence	No recurrence
Total number of patients	24	15
Mean age (range)	65 (43–73)	64 (42–78)
Sex		
Male	13	9
Female	11	6
Mean follow-up (months)	26	24
Stage		
I (1A and 1B)	8	10
II (2A and 2B)	13	2
III (3A)	3	3
Histological type		
ADC	11	8
SQCC	10	4
Others[a]	3	3

[a] Includes adenosquamous carcinoma (ADSQ), large cell undifferentiated carcinoma (LCUC), and carcinoid tumor (CND).

DIGeasy hybridization buffer (Roche) combined with 5 μl of 10 mg/ml yeast tRNA, 5 μl of 10 mg/ml salmon sperm DNA, denatured for 5 min at 65°C, then hybridized to glass arrays overnight at 37°C in humid hybridization chambers. Slides were washed the next day six times for 5 min at room temperature in 0.1× SSC, 0.1% SDS, and then 2 times with 0.1× SSC. Slides were dried by centrifugation before scanning using a Genepix 4000 scanner (Axon Instruments Inc., Union City, CA) and Tiff files analyzed using Quantarray image analysis software (GSI Lumonics, Ottawa, Ontario, CA). Cy3 and Cy5 channels were normalized by subarray intensity, and ratios were calculated as test sample divided by reference sample on each array.

Data Analysis. A subset of 2899 genes that contained datapoints in at least 80% of the samples and of which the transcripts had at least two or more samples with an absolute value of two in $\log_2$ space were included for additional analysis. Hierarchical clustering was performed on all 39 of the patients according to Eisen *et al.* (12). For statistical testing, the main outcome was DFS defined as the time between surgery and the first failure (either recurrence or death). When an event did not occur the time was calculated between surgery and last follow-up date, and was considered censored. The median follow-up (calculated for the 16 without events) was 2 years. The exact time of 2 patients with relapse was not clear from the medical record, and they were excluded from this part of the analysis. There were 21 events in this set of 37 patients. A Cox proportional hazards model was used in testing which of the 2899 genes considered well measured were significant. To preserve an overall significance value of 0.05 an adjustment for multiplicity was done based on Dubey's approach

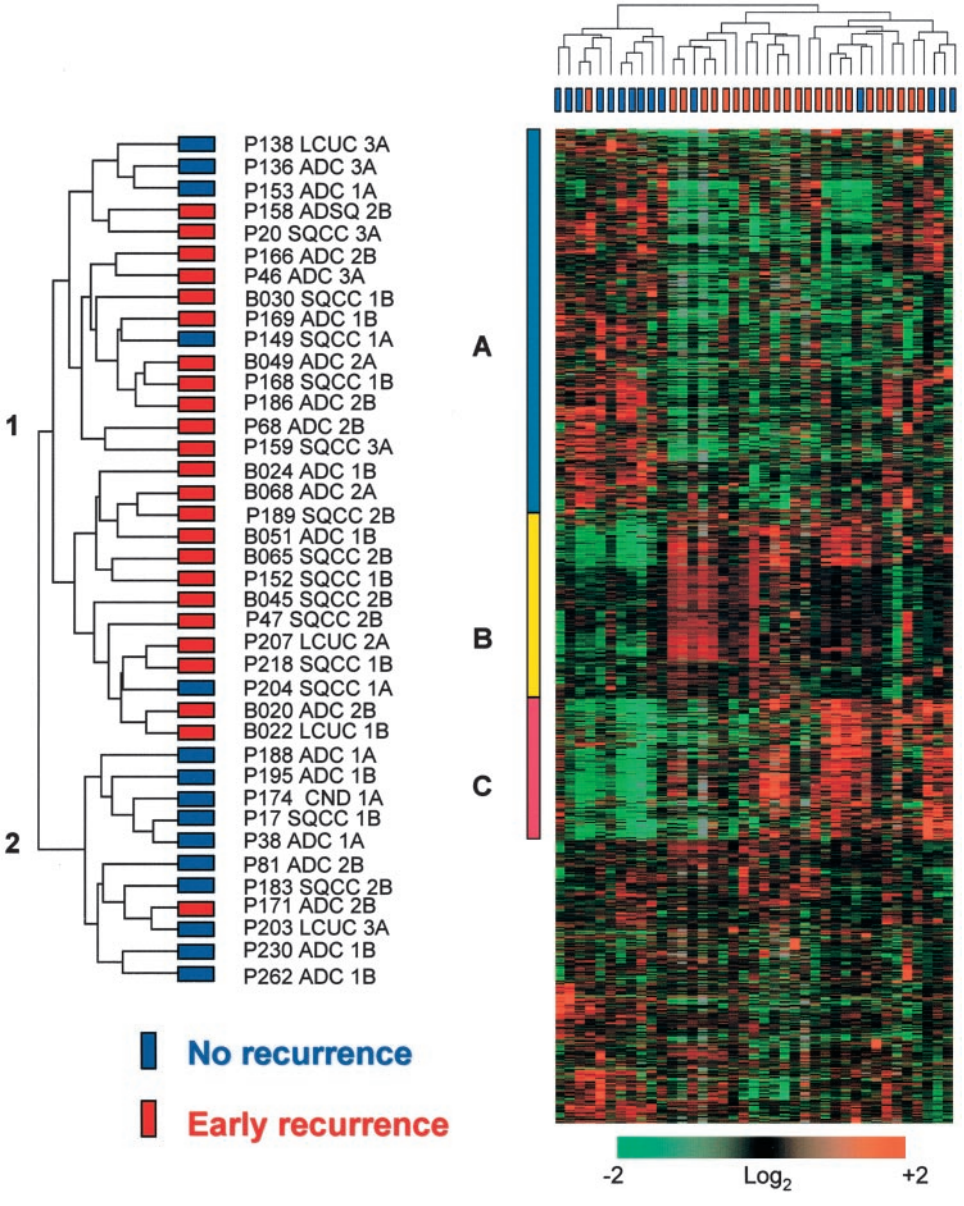

Fig. 1. Gene expression profile and early tumor recurrence. Thirty-nine patients with clinical outcome data were clustered hierarchically on the basis of 2899 genes. Cluster tree of individual patient samples and overall pattern of median-centered gene expression data are shown. The gene clusters that appear most obvious to be differentially expressed in patient group 1 with high recurrence rate as compared with patient group 2 with low relapse rate are labeled *A*, *B*, and *C*. Table 1 illustrates selected genes that may be of oncogenic interest for each of these clusters. *ADC*, adenocarcinoma; *SQCC*, squamous cell carcinoma; *ADSQ*, adenosquamous carcinoma; *LCUC*, large cell undifferentiated carcinoma; *CND*, carcinoid. *1A*, *1B*, *2A*, *2B*, and *3A* refer to pathological stage of the tumors.

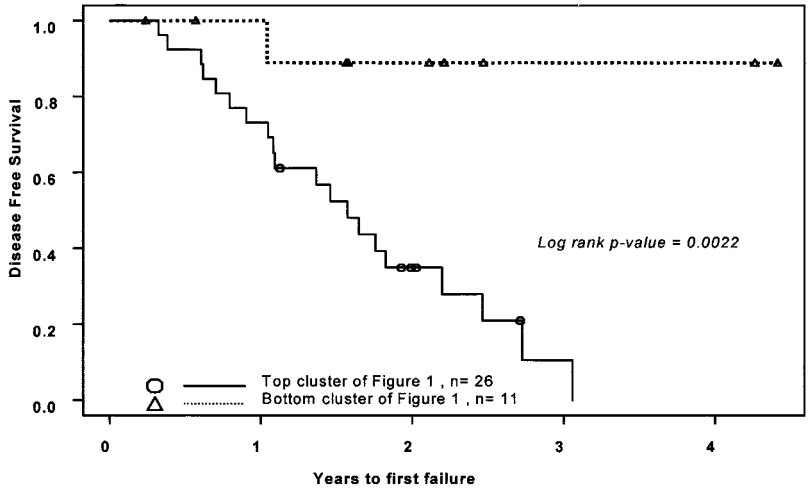

Fig. 2. Kaplan-Maier plot of DFS. DFS for the two primary patient clusters shown in Fig. 1.

(12). The correlation coefficient needed in this formula was the median of the correlation coefficients between the gene with the highest likelihood calculated in the Cox model and the remaining 2898 genes. A gene was considered significant if the P was ≤ 0.0023 (13).

Results and Discussion

Our analysis focused on tumor expression profiles from 39 NSCLC patients whom we had obtained clinical outcome data with a minimum of 1 year follow-up. Within this group, 24 patients had experienced relapse of their tumor either locally or as a distant metastasis. The remaining 15 patients are disease-free based on both clinical and radiological testing. As listed in Table 1, the two groups were broadly similar in distribution of age, sex, length of follow-up, and histological subtype. The only observable difference was a higher number of stage II cancers in the recurrence group; however, we did not find correlations of clinical stage with gene expression profile in a larger database of NSCLC expression profiles from a total of 84 patients nor within a preliminary analysis of raw data from published NSCLC microarray datasets (not shown; Refs. 14, 15).

We applied a number of different approaches in an attempt to define gene expression patterns segregating with early tumor recurrence within this group of 39 patients. Unsupervised hierarchical clustering (Fig. 1) shows the pattern differences between the groups on the basis of 2899 genes that contained datapoints in at least 80% of the samples and whose transcripts had at least two or more samples with an absolute value of 2 in $\log_2$ space. Remarkably, two groups emerged that appeared to separate patients in the dataset with relapse compared with those that remain disease-free. A number of oncologically relevant genes are observed within the most prominent clusters marked in Fig. 1. Table 2 lists these selected genes of interest, a number of which have received attention in the context of NSCLC biology. Ataxia telangiectasia mutated is a checkpoint kinase that helps to transduce genomic stress signals that halt cell cycle progression and promote DNA repair, and consequently may mark more aggressive tumor biology as indicated by our profiling result (16). Up-regulation of the flt1 vascular endothelial growth factor receptor is another notable finding, implying a role for increased angiogenic activity in more severe NSCLC lesions. The remaining genes listed in Table 2 contain a variety of transcripts, including a number with known roles in kinase-based signaling (PIK3R2, PPP2R3, RABIF, and MAK).

Table 2 *Selected genes of interest from indicated clusters in Figure 1*

	GenBank accession no.	Gene annotation
Cluster A	R56646	PNUTL1 peanut-like 1 (*Drosophila*)
	AA001333	FUS Fusion, derived from t(12;16) malignant liposarcoma
	H41817	Hippocalcin
	R80974	Wee1 + homologue (*Schizosaccharomyces pombe*)
	R55942	serine protease 25 (PRSS25)
	H58462	Nop10p Nucleolar protein family A, member 3
Cluster B	H45943	ATM (Ataxia telangiectasia mutated)
	H30017	PIK3R2 (Phosphoinositide-3-kinase regulatory subunit)
	W47003	HIF1A (hypoxia-inducible factor 1A)
	AA043244	MMP2 (matrix metalloproteinase 2)
	N92620	TIMP3 (tissue inhibitor of metalloproteinase 3)
Cluster C	T84132	PPP2R3 (Protein phosphatase 2 regulatory subunit)
	W67485	ZNF136 (Zinc finger protein 136)
	AA035437	FLT1 (Fms-related tyrosine kinase 1)
	H63198	RABIF (RAB interacting factor)
	AA059211	MAK (Male germ cell-associated kinase)

Kaplan-Maier plots of the two major arms of the cluster tree (Fig. 2) showed a statistically significant difference in DFS between these two groups ($P = 0.0022$). Using Cox proportional hazards model testing, we determined that the expression level of 22 of the 2899 gene subsets used for clustering analysis was statistically significant for DFS (Table 3; complete list available[4]). The majority of these genes are found within cluster A in Fig. 1. Of these 22 genes, 15 map to unigene clusters, 6 of which are predicted proteins found as expressed sequence tags in numerous libraries with no annotated function. Of the known genes, a number have potentially interesting relationships with putative mechanisms of malignancy. Hippocalcin has been shown recently to interact with the neuronal apoptosis inhibitory protein NAIP, and coexpression of the two proteins is protective against calcium-mediated cell death stimuli (17). The G_2 DNA damage and DNA replication checkpoints in many organisms act through the inhibitory phosphorylation of Cdc2, catalyzed by the Wee1/Mik1 family of kinases (18). The Rho GDP dissociation inhibitor is implicated in the effects of Rho on cytoskeletal function (19), a process for which there are numerous implications of altered behavior in malignancy. On the basis of our analysis using a series of artificial intelligence-based approaches to the full 19.2 K dataset,[5] many genes that were below the level of statistical significance based on a Cox proportional hazards model also appeared to be capable of class distinction between the groups of patients with early recurrence *versus* no recurrence. Notable examples include the altered expression of two

[5] D. A. Wigle, I. Jurisica, F. A. Shepherd, and M. S. Tsao, Molecular subtyping of lung cancer from an artificial intelligence-based analysis of gene expression profiles, manuscript in preparation.

Table 3 *Statistically significant genes for DFS from Cox proportional hazards model testing*

GenBank accession no.	Unigene ID	Gene annotation	*P* and expression change[a]
AB001105	Hs.272891	Hippocalcin-like protein 4 (HPCAL4)	0.0006 ↓
H29772	Hs.32501	predicted protein	0.0006 ↓
H06235	Hs.12294	predicted protein	0.0007 ↓
H14691	Hs.159161	Rho GDP dissociation inhibitor (GDI) α	0.0007 ↓
H17978	Hs.21814	Interleukin 20 receptor, α	0.0011 ↓
T97149	Hs.83575	KIAA1533 predicted protein	0.0012 ↓
T65437	Hs.268053	KIAA0029 protein	0.0012 ↓
H58462	Hs.14317	NOLA3 Nucleolar protein family A, member 3	0.0014 ↓
R40106	Hs.286145	RNA polymerase II holoenzyme component SRB7	0.0016 ↓
R80974	Hs.75188	Wee1+ (*S. pombe*) homolog protein kinase	0.0017 ↓
AF020760	Hs.115721	serine protease 25 (PRSS25)	0.0017 ↓
R80770	Hs.30819	C40 Hypothetical protein	0.0018 ↓
R73128	Hs.73931	HLA-DQB1 MHC	0.0019 ↓
AF201940	Hs.283740	DC6 predicted protein	0.0020 ↓
R20380	Hs.12758	predicted protein	0.0021 ↓

[a] Expression level change correlating with poorer DFS; ↓, decreased expression.

putative p53 pathway proteins; including TP53BP1, a p53-binding protein (GenBank accession no. AF078776) and TP53TG1, a p53 target gene (GenBank accession no. BC002709).

Recent studies involving gene expression profiling of clinical specimens have had a profound impact on cancer research. In many examples, correlations have been made between the expression levels of a gene or set of genes and clinically relevant subclassifications of specific tumor subtypes. These results have compounded expectations that true molecular classification and substaging of multiple tumor types may be possible, leading to measurable improvements in prognosis and patient management. Two recent publications have identified gene expression profiles correlating with histological subtypes of lung cancer, in addition to subtypes of adenocarcinoma with apparent differences in survival (14, 15). Although in our dataset we did not observe a primary correlation of expression profiles with tumor histological type, we could identify gene subsets capable of this distinction. It is worth noting that our study included roughly equal numbers of adenocarcinoma and squamous cell carcinoma (49% and 36%, respectively), whereas in contrast, the tumors analyzed in the prior two publications showed higher proportions of adenocarcinomas (61% and 68%) compared with squamous cell carcinoma (11% and 24%). Overall, these results suggest a number of implications for the classification and staging of NSCLC. Most prominent is that a true molecular staging system, built on either the current system or constructed anew, has the potential to additionally refine diagnosis, prognosis, and patient management for this lethal disease. Our eventual goal involves distilling the clone set down to the critical genes necessary for class distinction and using this to construct a lung cancer-specific gene chip for prospective testing in clinical trials. This forms part of the larger objective of defining and implementing a true molecular staging system for NSCLC.

References

1. Greenlee, R. T., Hill-Harmon, M. B., Murray, T., and Thun, M. Cancer Statistics, 2001. CA Cancer J. Clin., *51:* 15–36, 2001.
2. Mountain, C. F. Revisions in the international system of staging for lung cancer. Chest, *111:* 1710–1717, 1997.
3. Sekido, Y., Fong, K. W., and Minna, J. D. Progress in understanding the molecular pathogenesis of human lung cancer. Biochim. Biophys. Acta, *1378:* F21–F29, 1998.
4. Gandara, D. R., Lara, P. M., Lau, D. H., Mack, P., and Gumerlock, P. H. Molecular-clinical correlative studies in non-small cell lung cancer: application of a three-tiered approach. Lung Cancer, *34:* 75–80, 2001.
5. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.
6. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature (Lond.), *406:* 536–540, 2000.
7. Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A., and Hampton, G. M. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA, *98*: 1176–1181, 2001.
8. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.
9. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA, *98:* 10869–10874, 2001.
10. Takahashi, M., Rhodes, D. R., Furge, K. A., Kanayama, H., Kagawa, S., Haab, B. B., and Teh, B. T. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proc. Natl. Acad. Sci. USA, *98:* 9754–9759, 2001.
11. Tsao, M. S., Zhu, H., and Viallet, J. Autocrine growth loop of the epidermal growth factor receptor in normal and immortalized human bronchial epithelial cells. Exp. Cell Res., *223:* 268–273, 1996.
12. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95*: 14863–14868, 1998.
13. Dubey, S. D. Adjustment of p-values for multiplicities of intercorrelating symptoms. C. R. Buncher and J. Y. Tsay (eds.), Statistics in Pharmaceutical Industry (2nd Ed.), pp. 517–518. New York: Marcel Dekker, 1994.
14. Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. USA, *98*: 13784–13789, 2001.
15. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA, *98*: 13790–13795, 2001.
16. Cortez, D., Guntuku, S., Qin, J., and Elledge, S. J. ATR and ATRIP: partners in checkpoint signaling. Science (Wash. DC), *294:* 1713–1716, 2001.
17. Mercer, E. A., Korhonen, L., Skoglosa, Y., Olsson, P. A., Kukkonen, J. P., and Lindholm, D. NAIP interacts with hippocalcin and protects neurons against calcium-induced cell death through caspase-3-dependent and -independent pathways. EMBO J., *19:* 3597–3607, 2000.
18. Rhind, N., and Russell, P. Roles of the mitotic inhibitors Wee1 and Mik1 in the G(2) DNA damage and replication checkpoints. Mol. Cell. Biol., *21:* 1499–1508, 2001.
19. Sasaki, T., and Takai, Y. The Rho small G protein family-Rho GDI system as a temporal and spatial determinant for cytoskeletal control. Biochem. Biophys. Res. Commun., *245:* 641–645, 1998.

Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival

Dennis A. Wigle, Igor Jurisica, Niki Radulovich, et al.

Cancer Res 2002;62:3005-3008.