Gene Expression Profiling Allows Distinction between Primary and Metastatic Squamous Cell Carcinomas in the Lung

Simon G. Talbot,[1] Cherry Estilo,[3] Ellie Maghami,[1] Inderpal S. Sarkaria,[1] Duy Khanh Pham,[1] Pornchai O-charoenrat,[7] Nicholas D. Socci,[5] Ivan Ngai,[1] Diane Carlson,[4] Ronald Ghossein,[4] Agnes Viale,[6] Bernard J. Park,[2] Valerie W. Rusch,[2] and Bhuvanesh Singh[1]

[1]Laboratory of Epithelial Cancer Biology, Head and Neck Service, [2]Thoracic Oncology Service, [3]Dental Service, Departments of Surgery and [4]Pathology, [5]Computational Biology Center, [6]Genomics Core Laboratory, Memorial Sloan-Kettering Cancer Center, New York, New York; and [7]Department of Head and Neck Surgery, Siriraj Hospital, Bangkok, Thailand

Abstract

Lung neoplasms commonly develop in patients previously treated for head and neck carcinomas. The derivation of these tumors, either as new primary lung cancers or as metastatic head and neck cancers, is difficult to establish based on clinical or histopathologic criteria since both are squamous cell carcinomas and have identical features under light microscopy. However, this distinction has significant treatment and prognostic implications. Gene expression profiling was performed on a panel of 52 sequentially collected patients with either primary lung (n = 21) or primary head and neck (n = 31) carcinomas using the Affymetrix HG_U95Av2 high-density oligonucleotide microarray. Unsupervised hierarchical clustering with Ward linkage and the Pearson correlation metric was performed. To assess robustness, bootstrap resampling was performed with 1,000 iterations. A t test of the normalized values for each gene was used to determine the genes responsible for segregating head and neck from lung primary carcinomas, and those with the most differential expression were used for later analyses. In the absence of a large "test" set of tumors, we used a supervised leave-one-out cross-validation to test how well we could predict the tumor origin. Once a gene expression profile was established, 12 lung lesions taken from patients with previously treated head and neck cancers were similarly analyzed by gene expression profiling to determine their sites of origin. Unsupervised clustering analysis separated the study cohort into two distinct groups which reliably remained segregated with bootstrap resampling. Group 1 consisted of 30 tongue carcinomas. Group 2 consisted of 21 lung cancers and 1 tongue carcinoma. The clustering was not changed even when normal lung or tongue profiles were subtracted from the corresponding carcinomatous lesions, and a leave-one-out cross-validation showed a 98% correct prediction (see Supplementary Data 1). A minimum set of 500 genes required to distinguish these groups was established. Given the ability to segregate these lesions using molecular profiling, we analyzed the lung tumors of undetermined origin. All cases clearly clustered with either lung or tongue tumor subsets, strongly supporting our hypothesis that this technique could elucidate the tissue of origin of metastatic lesions. Although histologically similar, squamous cell carcinomas have distinct gene expression profiles based on their anatomic sites of origin. Accordingly, the application of gene expression profiling may be useful in identifying the derivation of lung nodules and consequently enhances treatment planning. (Cancer Res 2005; 65(8): 3063-71)

Introduction

Squamous cell carcinoma of both the lung and the head and neck contributes significantly to morbidity and mortality worldwide. Lung cancer ranks as the leading cause of cancer death in the United States, predicted to be responsible for 173,770 cases and 160,440 deaths during 2004. Non–small cell lung cancer comprises the majority of these cases, and squamous cell carcinoma is a large subset of this group. Head and neck squamous cell carcinoma is ranked within the top 10 most prevalent cancers in males with oral, pharyngeal, and laryngeal cancers in the United States, predicted to cause 38,530 cases and 11,060 deaths during 2004 (1).

Numerous risk factors are shared by lung and head and neck cancers. The most important known risk factor for both is tobacco exposure, which is known to contain numerous mutagens and carcinogens. Tobacco use may increase the risk for development of head and neck squamous cell carcinomas up to 17 times (2) and lung cancers at least 2.6 times that of nonusers (3). Given the similar risk factors and etiologies, it is not surprising that these diseases frequently occur together in the same patient. Two broad explanations may account for the conjoint occurrence of these tumors. First, the concept of "field cancerization" is well-known in the literature (4, 5), and can perhaps be more broadly applied to the coexistence of these diseases throughout the body. Squamous mucosa of both the head and neck and lungs is exposed to mutagens that may cause genetic aberrations responsible for the initiation of the carcinogenic process at multiple sites. Alternatively, carcinogens such as tobacco may cause both diseases by different mechanisms in each site, still ultimately causing concurrent cancers. In patients with a previous lung cancer, the incidence ratio of a second primary lung cancer is 1.7 and that of an oral cavity or pharyngeal cancer is 2.7 (6). Conversely, patients with a previous laryngeal cancer have a 4.5-fold increased incidence ratio of a lung cancer when followed for >5 years (7). This perhaps attests to similar genetic mechanisms and biologies accounting for squamous cell carcinomas in general.

Confounding the diagnosis of these patients is the fact that the lungs are the most common site for metastasis of head and neck squamous cell carcinomas. Since both lung primary tumors and

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Presented at the American Association for Cancer Research Annual Meeting, 2003. Washington, DC, Poster Session, Abstract #6014.

Requests for reprints: Bhuvanesh Singh, Head and Neck Service, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021. Phone: 212-639-2024; Fax: 212-717-3302; E-mail: singhb@mskcc.org.

metastatic head and neck tumors arise from squamous cells, these carcinomas are often indistinguishable based on traditional histopathologic analysis; and thus, the site of origin of the tumor is indeterminate. This creates a complex conundrum in patients with lung lesions and a history of prior head and neck cancers: is this a second primary lung squamous cell carcinoma or is it a metastasis from the previously treated head and neck squamous cell carcinoma? Oligonucleotide and cDNA microarrays have been hailed as "the newest types of microscopes" (8) since they allow the classification of tumors on the assumption that the genotype directly leads to a phenotype, and can be fingerprinted in a way far more precise than our ability to detect subtle changes under the light microscope.

This concept is eloquently stated by Caldas et al. (9), who explains that the first step in rationally treating a disease is to correctly classify the disease in order to predict response. This process is dependent on the quality of the classification method, and molecular techniques seem to be refining this method. Others have shown molecular classification to have ~78% accuracy, compared with random classification's 9% accuracy (10). Whereas this still shows significant error, the inclusion of more genes, novel techniques, and applications shows considerable promise, with early studies generating significant interest (11). Moreover, whereas our biological techniques expand, it is often the ingenuity of the *in silico* modeling which gives the data clinical utility, as we show in this study.

Knowledge of whether a lung lesion constitutes a second lung primary versus a head and neck metastasis contributes significantly to treatment planning and prognosis. First, patients with an isolated primary lung lesion may be effectively treated by resection with an ~82% 5-year survival and 74% 10-year survival for T_1 non–small cell lung cancer (12). Even if the lesion is in fact a metachronous lung lesion, survival may be similar to that for resection of the initial primary (13). In contrast, those patients with metastatic head and neck cancers to the lung have clinically and biologically systemic disease, in which the cancer cells have acquired the ability to migrate, travel, and seed new metastases. Consequently, surgical therapy offers only a 34% 5-year survival for squamous cell carcinomas metastatic to the lung even in well selected patients (14). Second, the extent of lung resection (if the patient is a surgical candidate) is potentially more tissue-sparing if it is for a metastasis versus a primary lung tumor—consisting of a wedge or segmental resection rather than a lobe or more extensive resection. Third, if we know that a patient has a metastasis, given an otherwise poor survival with resection alone, we may be inclined to give multimodality therapy including chemotherapy and surgery. Thus, the distinction between whether a lung lesion constitutes a primary or metastatic lesion is not trivial and may be important in the clinical care of the patient (15). Whereas other groups have used similar methodologies to molecularly classify tumors into known subsets (16), our data may provide an answer to this troubling clinical question of the location of the primary disease.

To date, the vast majority of oligonucleotide and cDNA microarray studies have focused on delineating the genetic mechanisms of carcinogenesis (8, 17–19), isolating novel therapeutic and diagnostic targets (20–22), classifying tumors molecularly (23, 24), and prognosticating (11, 25–27). Here we show that validated oligonucleotide microarray technology can be used to clearly distinguish the tissue of origin of a squamous cell carcinoma, using an unsupervised hierarchical clustering analysis with supervised validation. This algorithm can then be applied to a "test set" of tumors of unknown origin, allowing us to determine where these originated, and whether they constitute new lung primary tumors or metastatic head and neck tumors.

Materials and Methods

Patient selection. Three separate tumor cohorts were used in this study. The first was a group of 21 sequentially collected non–small cell lung cancers from patients treated between February 1990 and July 1997. The second was a group comprising 31 sequentially collected oral tongue squamous cell carcinomas from patients treated between January 1998 and January 2002. Patients in each of these groups were previously untreated and had no known distant metastases at the time of presentation. Details of patient characteristics are shown in Supplementary Data 2. The third cohort was comprised of patients with a prior history of squamous cell carcinoma of the head and neck having squamous cell carcinoma tissue resected from the lungs between 1995 and 2002.

Tissue collection. Under informed consent and Institutional Review Board–approved protocols, tissue from patients undergoing operative procedures was collected directly following resection in the operating rooms at Memorial Sloan-Kettering Cancer Center. Tumor samples were collected from the advancing tumor edge along with adjacent normal tissue and snap-frozen in liquid nitrogen at the time of operation. These tumors were then banked at −80°C for storage until later use.

Confirmation of pathology. All tumors were examined by routine pathology using H&E staining to determine pathology and tumor grade. To confirm the initial pathology report and grade, we further reviewed every case with a pathologist (R.A. Ghossein and D. Carlson). We also confirmed a minimum of 70% tumor involvement of the final section in order to minimize contamination of gene expression profiles from surrounding normal stromal tissue, in keeping with reports from other investigators (10). Whereas tumor cells may be enriched using laser capture microdissection, this introduces the inherent problems of cDNA amplification and is impractical for clinical use.

Tissue processing. For RNA extraction, 100 mg of frozen tissue was homogenized in 1 mL of TRIzol Reagent (Life Technologies, Gaithersburg, MD) and RNA extracted according to the manufacturer's protocol. The obtained RNA was then further purified using the RNeasy (Qiagen, Valencia, CA) system and protocol. Samples were quantified using standard spectrophotometry, and considered acceptable if the A260/280 reading was >1.7.

Affymetrix oligonucleotide microarray. RNA quality was confirmed using an Agilent 2100 Bioanalyzer. Total RNA (25-50 ng) was run on a RNA 6000 Nano Assay (Agilent, Palo Alto, CA). Samples were accepted for further analysis if clear 28S and 18S RNA bands were present. Total RNA (5-10 μg) was then reverse-transcribed using an oligo dT-T7 primer and cDNA synthesis kit (Life Technologies). cDNA was isolated by phenol/chloroform extraction and resuspended in 12 μL of diethyl pyrocarbonate-treated water. Ten microliters of the resultant solution was then used in an *in vitro* transcription-amplification reaction in the presence of biotinylated nucleotides (Enzo Diagnostics, Farmingdale, NY). Fifteen micrograms of labeled cRNA was fragmented by incubation at 95°C for 35 minutes in fragmentation buffer [40 mmol/L Tris-acetate (pH 8.1), 100 mmol/L KOAc, 30 mmol/L MgOAc] and hybridized onto a Test 3 array and subsequently onto an Affymetrix HG_U95Av2 oligonucleotide chip for 16 hours at 45°C (Affymetrix, Santa Clara, CA). Posthybridization staining and washing were processed according to the manufacturer's instructions (Affymetrix). Scanning was done using a Hewlett-Packard argon-ion laser confocal scanner and analyzed using Microarray Suite 5.0 (Affymetrix). Images were quantified using MAS 5.1 (MicroArray Suite, Affymetrix) with the default parameters for the statistical algorithm and all probe set scaling with a target intensity of 500.

Validation of microarray data. Results from the Affymetrix oligonucleotide microarrays were validated by real-time reverse transcription-PCR. For the lung squamous cell carcinoma cohort, six up-regulated genes and three down-regulated genes were chosen to confirm the microarray

findings. For the oral tongue squamous cell carcinoma cohort, three genes were examined by real-time reverse transcription-PCR. Real-time reverse transcription-PCR methodology has been published previously (28).

Statistical analyses. In order to filter noise from the analysis, genes that were present in <25% of the samples were excluded from initial analyses. Hierarchical clustering on this set was done as follows: gene values were normalized by subtracting the means of the signal intensities for each gene; the distance metric used was $d = (1 - \rho) / 2$, where ρ is the Pearson correlation function between samples. The Ward linkage method was used. Bootstrap nonparametric resampling was done using 1,000 iterations. A majority rule consensus tree was constructed from the 1,000 bootstrap trees. Next, genes were selected to maximally discriminate the tongue from lung tumors using a standard t test on the log of the absolute signal intensity, and the data was reclustered using these genes. The number of genes selected was arbitrary. However, we found that approximately 100 genes perfectly discriminated the two groups; whereas choosing more genes improved the robustness of the clustering. Five hundred genes seemed to be an acceptable compromise, allowing a manageable number of genes yet maintaining stable clusters. To confirm and quantify how well tongue versus lung tumors could be discriminated, a supervised learning analysis was done using support vector machines and leave-one-out cross-validation on the original tumors. Thirteen non–head and neck squamous cell carcinoma samples of unknown origin (primary or metastatic) were then added to the analysis after the gene selection.

Results and Discussion

Gene expression profiling of primary tumors. The lung cancer microarray data set was internally validated using real-time reverse transcription-PCR for six genes found to be highly up-regulated on the array (HSN, GCS, BTAK, TTK, cyclin E2, and NLK with mean fold changes of 1,333.5-, 123.6-, 2.93-, 121.6-, 4.1-, and 2.83-fold, respectively) and three found to be highly down-regulated (ANG, PTPTM, and C1-Inh with mean fold changes of 0.037-, 0.0001-, 0.0055-, and 0.05-fold respectively). The overall trend and correlation confirmed the validity of the microarray data (P = 0.0002-0.02, Spearman r = 0.43-0.67). We have previously shown the lung cancer cohort to reliably segregate into three distinct subgroups by unsupervised hierarchical clustering (29).

In the tongue cancer patient cohort 79 genes were found to be significantly associated with oral tongue cancer when compared with case-matched, histologically normal mucosa. Three genes (GLUT3, HSAL2, and PACE 4) were selected for further analysis and validated as prognostic markers in a larger cohort of 49 patients by quantitative real-time reverse transcription-PCR. Using a criterion of 2-fold up-regulation as a cutoff, 30.6%, 24.5%, and 26.5% of patients expressed high levels of GLUT3, HSAL2, and PACE4, respectively; again confirming the validity of the microarray data (30).

Other groups have used Northern blot and/or immunohistochemistry validation (25). However, the specificity and accurate quantification afforded by real-time reverse transcription-PCR probably make it an ideal method for internal validation of array data.

Although it is clearly impractical to examine all genes responsible for the segregation of the various tumor subsets, several provide confirmatory validation of the segregation and also raise a number of interesting potential causal links. For example, the protein kinase C substrate is found significantly more often in the primary lung tumor subset. Given that this is a regulatory molecule that mediates mucin granule release by bronchial epithelial cells, it is not surprising to find it more commonly expressed in the lung tumor subset (31). The polycomb 2 homologue is a gene also highly expressed in lung tumor tissue. This gene is thought to be a repressor of proto-oncogene function and that interference with it may lead to cellular transformation, consistent with our finding that this gene is highly expressed in this malignant tissue (32). Conversely, EIF-2α is found more commonly in our tongue tumor subset. Notably, this gene may be involved in the hypoxic stress response (33). Hence, it is likely to be more important to cells in the relatively more hypoxic environment of the oral mucosa than in the lung.

Hierarchical clustering of primary tumors. All lung and tongue tumors were combined into an unsupervised hierarchical clustering model. From the 12,625 probe sets on the array, those expressed in <25% of samples were excluded (as is usual practice in microarray analysis) to filter out noise, leaving 6,716 probe sets for analysis. The two groups clearly segregated based on the tissue of origin of the primary tumor (Fig. 1). Unsupervised clustering was done without prior knowledge of how these tumors might cluster, and to determine if groups would form without biasing the data, as suggested by numerous authors (34). Thirty of the 31 tongue tumors grouped in cluster 1. All 21 lung tumors grouped in cluster 2, along with a single tongue tumor. Notably, the three lung adenocarcinoma samples formed a separate subcluster within the lung group, as our group and others have found previously (24, 35, 36). These two clusters and single subcluster were the only truly robust clusters, with bootstrap resampling of >95%. Together, these data show the uniqueness of tumors from each tissue, and indicate the importance of genes intrinsic to the tissue of origin in determining the expression profile of a tumor.

In order to ensure that effects of the surrounding normal tongue or lung stromal tissue were not interfering with expression profiling of tumor tissue, we subtracted the profiles of either normal lung or tongue tissue for each tumor data set, respectively. Normal tissue from each patient was mixed and run on three separate arrays which were then combined mathematically. This combined profile was then subtracted from the profiles generated for each patient's tumor tissue. Even taking this into account, the segregation into two separate groups was strongly maintained (see Supplementary Data 1).

We next aimed to determine both the approximate number and identities of the genes causing this powerful tissue-based segregation. On one hand, ideally a single gene would determine the origin of a tissue and/or its carcinogenic potential. However, we find that no less than 50 genes can be used to segregate these tissues. Consequently, some authors propose the use of the maximum number of probe sets (37); however, this can both mask the differences in genotype due to noise and decrease the predictive value when arrays are validated using an unfiltered gene set. Moreover, others have suggested that increasing gene numbers from small to moderate improves the error rate in classification algorithms, but as the number of genes used becomes large, the generalization performance worsens (38). Thus, to determine an optimally small gene set to use, we arbitrarily selected the 1,000, 500, and 100 genes which were most different between the lung and tongue data sets by t test. Table 1 lists the 50 most significantly different genes between the two groups. Even using as few as 100 genes, results were sufficient to completely distinguish the two data sets. This suggests that <100 genes were responsible for the distinction between the tissues of origin. However, as the number of genes decreased, the robustness of the clustering reduced. Thus, for later analyses we chose to use

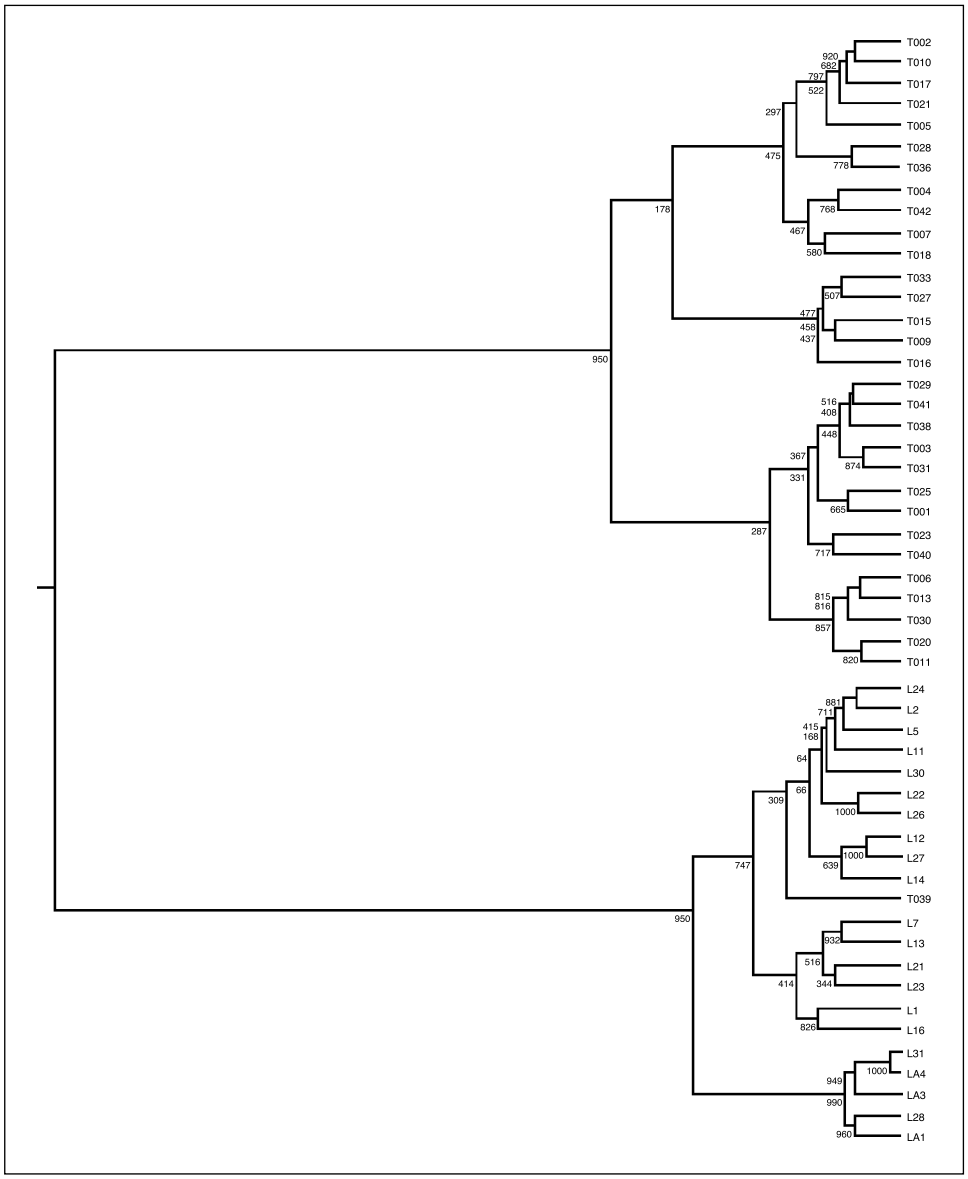

Figure 1. Consensus tree of unsupervised hierarchical clustering of lung tumors (L#) and tongue tumors (T#) using all probe sets. Horizontal distance represents the correlation of tumor samples to one another; numbers at each node represent frequency of the shown pattern using 1,000 iteration bootstrap resampling.

500 genes, as this maintained both the segregation and robustness of the clustering. Whereas 500 genes may seem to be a large number of genes to consider, it is a relatively small number in comparison with the total number of genes likely to determine a tissue type or the number of mutations thought to contribute to most carcinogenic processes.

Importantly, although we ultimately used 500 probe sets for our analyses, these were chosen statistically from *all* the probe sets on the array to be those which best segregated our clusters. They were not arbitrarily picked nor preselected by any criteria. This is in sharp contrast to many other similar studies in which focused and/or preselected genes are spotted onto cDNA microarrays (17, 20, 21). In these studies, similar numbers of genes may be used in the analyses, but due to a nonempirical selection of genes, many genes of potential importance may have been missed.

Clustering using just the subset of "unique genes" resulted in groupings similar to the initial unsupervised clustering (Fig. 2). However, this time *every* lung and tongue tumor segregated into their respective groups, with no outliers. Of note, within the lung tumor subgroup, we also included three lung adenocarcinomas (in addition to the 28 lung squamous cell carcinomas), as before, to delineate the significance of tissue of origin versus disease biology as the key segregating factor. These primary lung neoplasms of markedly different histology tended to segregate into a separate subcluster of the dendrogram within the lung subset, but more closely correlated than previously, confirming that although, as mentioned by others, disease biology contributed to the dendrogram clusters (36), the key feature in this analysis of 500 genes was tissue of origin.

We are not the first group to show that a relatively small set of genes may be required for tumor classification. Shedden et al. (38) used a tree-based framework to pathologically classify tumors using a model which ranked the importance of genes in the classification system as either coarse or fine in their importance within the tree structure. They used just 45 genes to accurately classify 157 of 190 malignant tumors according to pathologic type and were also able to determine the primary site in several metastases.

Another group, Leong et al. (39), have previously used microsatellite analyses of loss on chromosomal arms 3p and 9p to examine the origins of lung squamous cell carcinomas. Our work builds on this in that array-based analyses examines many thousands of genes (in comparison to just a small number of chromosomal arms). We are also able to retrospectively and algorithmically choose genes which seem to be important rather than having to predetermine and "best-guess" which areas of the

Table 1. The 50 most differentially expressed genes (by *t* test) allowing distinction of tongue from lung patient samples

T score (+ if lung > tongue)	Accession number	Description
15.74	L37033	FK-506 binding protein homologue (FKBP38)
14.74	U39840	hepatocyte nuclear factor-3α (HNF-3α)
13.07	AB014591	KIAA0691 protein
12.37	L21990	spliceosomal protein (SAP 62)
12.85	L25444	TAFII70-α
12.74	AF013956	polycomb 2 homologue (hPc2)
12.19	J04046	Calmodulin
11.20	J03075	80K-H protein (kinase C substrate)
10.64	L35013	spliceosomal protein (SAP 49)
10.54	D31840	DRPLA
10.45	X87852	*SEX* gene
10.34	U66617	SWI/SNF complex 60 kDa subunit (BAF60a)
10.33	X79780	YPT3
10.28	D85131	Myc-associated zinc-finger protein of islet
10.29	U35113	metastasis-associated gene (*mta1*)
10.24	AC004410	chromosome 19, fosmid 39554
10.09	W68046	zd42a12.s1 IMAGE
10.67	AF034373	ataxin-2-like protein A2LP (A2LG)
10.34	N58318	yv68h03.s1 IMAGE
−9.71	D42087	*KIAA0118* gene
9.87	X55885	presumptive KDEL receptor
−9.60	AI540958	PEC1.2_15_H01.r *Homo sapiens* cDNA
−10.76	U26032	translation initiation factor eIF-2α
10.06	M93036	carcinoma-associated antigen GA733-2 (GA733-2)
9.79	U56417	lysophosphatidic acid acyltransferase-α
9.85	AA628946	af28f05.s1 *H. sapiens* cDNA
9.95	X66362	HSSTHPKG *H. sapiens* mRNA PCTAIRE-3 for serine/threonine protein kinase
−9.43	AW003733	ws16b04.x1 IMAGE
9.24	AC005306	*H. sapiens* chromosome 19, cosmid R27216
−8.92	AB014566	KIAA0666 protein
−8.94	AL050224	DKFZp586L2123 (from clone DKFZp586L2123)
8.83	AL050118	DKFZp586C201 (from clone DKFZp586C201)
10.14	D87437	*KIAA0250* gene
9.24	U05681	HSBCL3S2 proto-oncogene (*BCL3*) gene
8.86	AB007929	KIAA0460 protein
8.64	AF086947	dynactin 1 (DCTN1)
−8.52	X05232	stromelysin
8.52	AD000092	chromosome 19p13.2 cosmids R31240, R30272 and R28549 (*EKLF*, *GCDH*, *CRTC*, and *RAD23A* genes)
8.85	X64037	RNA polymerase II associated protein RAP74
−8.69	AF097935	desmoglein 1 (DSG1)
8.50	U54778	HSU54778 14-3-3 epsilon
8.48	U52111	plexin-related protein
9.27	U60644	HU-K4
8.45	M62762	vacuolar H+ ATPase proton channel subunit
8.92	AB000520	APS
−9.00	AF070569	clone 24659
−9.50	X05451	myosin light chain 3 (MLC-3f)
8.85	Y00503	keratin 19
8.30	AF034102	NBMPR-insensitive nucleoside transporter ei (ENT2)
8.38	AB002559	hunc18b2

NOTE: Samples are in order of significance, with positive *t* scores representing the fold overexpression in lung, and negative *t* scores representing the fold overexpression in tongue samples.

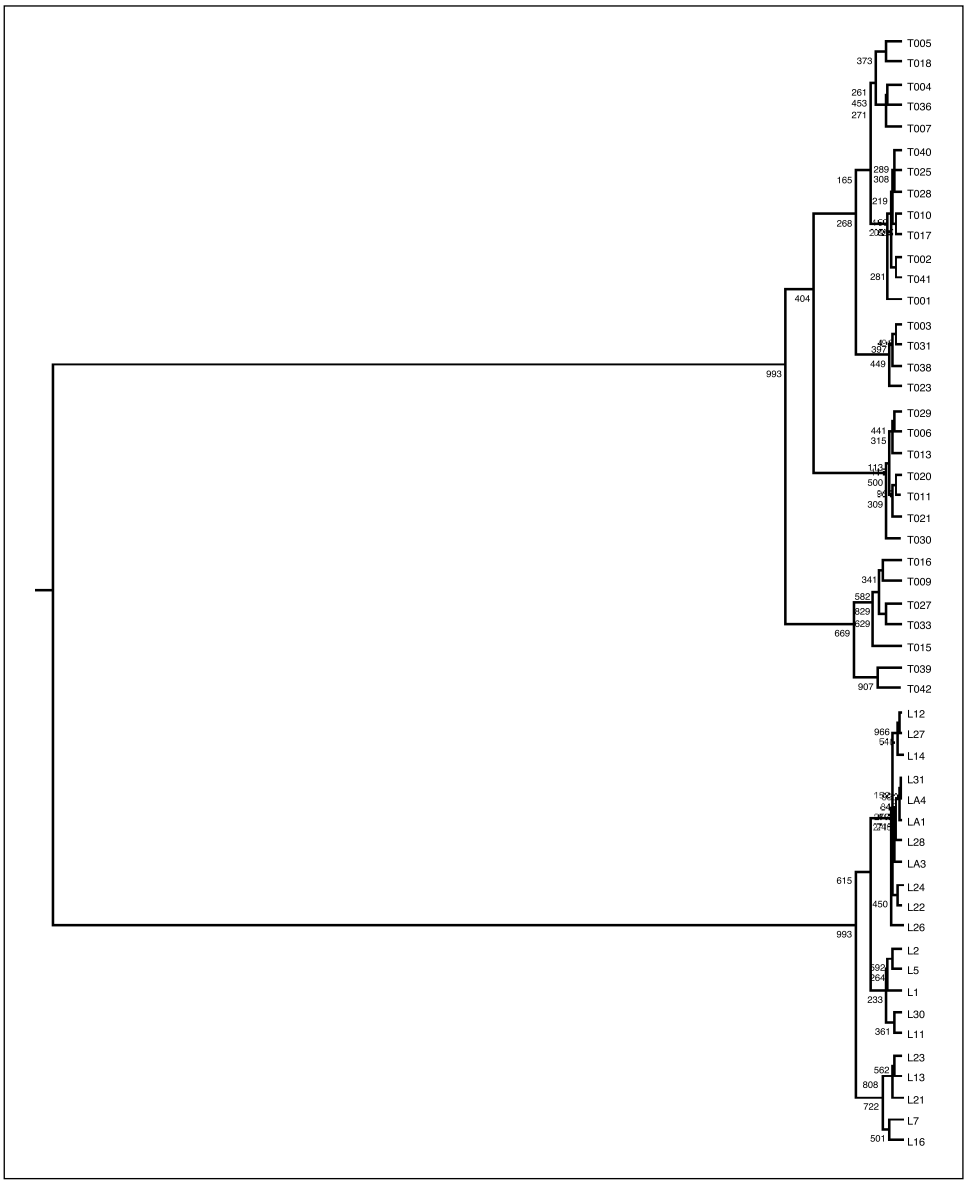

Figure 2. Consensus tree of hierarchical clustering of lung tumors (L#) and tongue tumors (T#) using 500 of the most differentially expressed genes selected by *t* test. Horizontal distance represents the correlation of tumor samples to one another; numbers at each node represent frequency of the shown pattern using 1,000 iteration bootstrap resampling.

genome are best to examine. Furthermore, computer pattern recognition allows us to compare a single biopsy sample from the lung against a generic normal tissue template of any tissue, alleviating the need for tissue from both the new tumor and previous primaries as are required for microsatellite analyses.

Our tumor cohorts for each group were not large enough to perform a complete validation by dividing the data into "training" and "test" sets, as would be ideal (16). In these circumstances, one of the optimal techniques is a supervised cross-validation rather than to re-test class assignment by repeated unsupervised clustering (34). Furthermore, supervised methods ensure that criteria relevant to the classification are what determine the groupings (40). We therefore elected to use a leave-one-out cross-validation using support vector machines to learn the classification (11, 25, 41). This analysis removes a sample, re-tests the categorization of remaining samples, and then returns the removed sample. This is repeated for every patient sample. The percentage of samples correctly predicted was 98%, further confirming that our prior clustering very accurately predicted class assignment of tumors to the head and neck or lung subsets.

Gene expression profiling of lung tumors of undetermined origin. Having identified a subset of 500 genes which reliably distinguished lung from tongue squamous cell carcinomas, we then chose to test this clustering algorithm on a clinical set of 12 lung squamous cell carcinomas resected from patients who were previously treated for head and neck squamous cell carcinomas (Table 2). As expected, histologic or clinical criteria could not precisely define the tissue of origin of these cases. However, the study group was biased as the majority of these tumors were thought to be of lung origin by virtue of the fact that they underwent resection. This reflects a clinical bias against resecting head and neck metastases to the lung given the poor patient outcome in this population. Nonetheless, many of the patient records documented the lack of certainty in deciding on the tissue of origin from the histologic biopsy or surgical samples. The tumors from the test set were processed and analyzed exactly as the training set and hybridized to Affymetrix HG_U95Av2 arrays.

The cluster algorithm was rerun using the 500 genes selected from the training set (Fig. 3). The addition of the samples from the test cohort did not destabilize the clusters. All of the cases from the test cohort reliably segregated into one of the two cluster groups. This offers strong molecular evidence to link the unknown cases either as lung or head and neck primary origin tumors, although no "gold standard" test exists to confirm this. As expected, the majority (11 samples, labeled U1-11) segregated with the lung tumors, supporting the clinical suspicion that these were lung primary tumors. One sample (labeled U12) robustly clustered with the

Table 2. Clinical details of unknown patient samples

Patient ID	Head and neck tumor	Lung tumor	Notes
U1	SCC hard palate 1994	SCC R hilum 2000	
	Treatment—excision	Treatment—resection	
U2	SCC R maxillary antrum 1993	SCC LUL 1998	Pathology favors metastasis from head and neck lesion, but complicated by presence of *in situ* carcinoma in lung specimen
	Treatment—radical maxilectomy with orbital exenteration and radiotherapy	Treatment—LUL wedge resection	
U3	SCC larynx 1986	SCC RLL 2000/2001	
	Treatment—laryngectomy and radiotherapy	Treatment—R lower lobotomy and middle lode wedge resection	
U4	SCC base of tongue 2002	SCC RUL 2002/2003	Clinically favors second primary based on nodal status of lung lesion
	Treatment—hemiglossectomy and RND and radiotherapy	Treatment—RU lobectomy	
U5	SCC R tonsil 1997	SCC upper mediastinum 2000	Cannot rule out met on pathology or clinical
	Treatment—radiotherapy	Treatment—resection but no completion lobectomy	
U6	SCC base of tongue and cervical LN met 1988/1989	SCC R bronchus intermedius/main stem 1992	
	Treatment—radiotherapy and brachytherapy plus L MRND	Treatment—radiotherapy (endobronchial)	
	SCC endotracheal 1990	SCC RUL 1994/1995	
	Treatment—radiotherapy	Treatment—thoracotomy and resection	
	SCC nasal cavity and palate 1995	SCC recurrence 1995	
	Treatment—resection plus radiotherapy/brachytherapy	Treatment—biopsy then chemotherapy and chemoprevention	
U7	SCC supraglottic larynx with R neck mets 1997	SCC LLL 1998	
	Treatment—radiotherapy and R RND	Treatment—L lower lobe resection and L upper lobe wedge	
U8	SCC false vocal cord (stage I) 1990	SCC RUL and LUL 1998	Pathology favors second primaries based on appearance
	Treatment—radiation	Treatment—bilateral wedge resections	
		Recurrences 1999	
		Treatment—R ant-basal segmentectomy, L upper lobectomy	
U9	SCC lower lip 1992	SCC RLL 2000	
		Treatment—R lower lobectomy	
U10	SCC R true vocal cord 1999	SCC RLL 2000	
	Treatment—R anterolateral vertical partial laryngectomy	Treatment—wedge resection	
U11	SCC larynx	SCC LUL1997	Pathology favors second primary based on *in situ* carcinoma
	Treatment—L hemilarygectomy	Treatment—LUL resection	
U12/U13	SCC larynx 1989	SCC RUL 1997	SCC pancreas 1998
	Treatment—? plus L RND	Treatment—R upper lobectomy	Treatment—distal partial pancreatectomy (labeled sample U13)
		Chest wall recurrence (? needle tract) 1998	
		Treatment—resected plus radiotherapy	

NOTE: SCC, squamous cell carcinoma; mets, metastasis; RND, radical neck dissection; LUL, left upper lobe; RUL, right upper lobe.

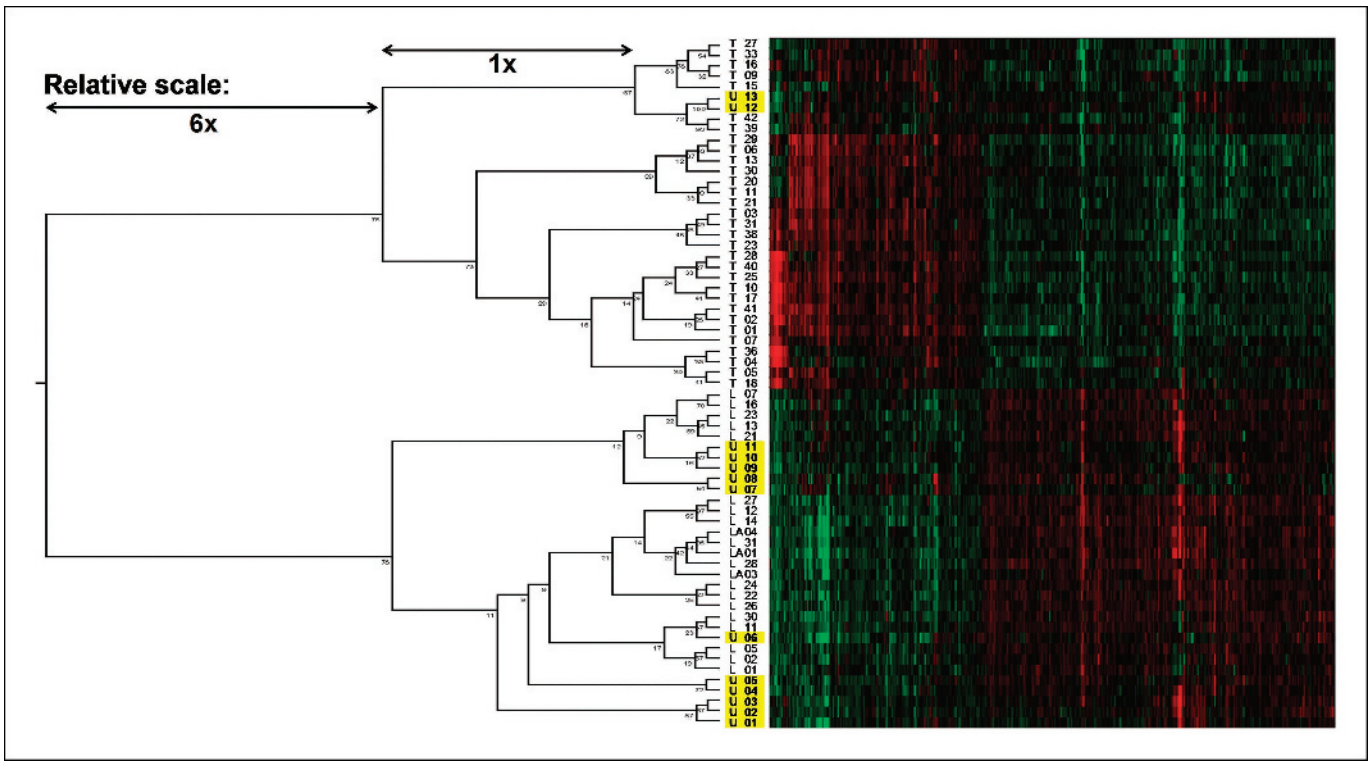

Figure 3. Consensus tree of hierarchical clustering of lung tumors (L#), tongue tumors (T#), and unknown metastases (U#), using 500 of the most differentially expressed genes between the initial groups. Horizontal distance represents the correlation of tumor samples to one another; numbers at each node represent frequency of the shown pattern using 100 iteration bootstrap resampling. Note the highly uncorrelated and robust grouping of lung tumors distinct from tongue tumors, with unknown samples interspersed. Note that samples U12 and U13 represent different metastases from the same patient.

tongue tumor subset, suggesting that it was likely a metastasis from a prior head and neck cancer. This suspicion was supported clinically by the development of a metachronous metastasis to the pancreas in this patient. To further validate the concept that metastatic tissues retain the genetic profile of the tissue of origin, we tested this pancreatic metastasis (labeled U13). This sample not only reliably clustered with the tongue subset, but despite being of pancreatic focus, clustered most closely with the same patient's lung lesion. This suggests a very strong maintenance of the tumor's gene expression profile regardless of the site to which it metastasizes.

This finding is supported by work of Bhattacharjee et al. (23) who were able to distinguish the site of origin of adenocarcinomas in the lung based on the expression profile. Another group has studied metastatic adenocarcinomas of unknown origin using computer modeling based on serial analysis of gene expression public databases confirming that metastases of adenocarcinoma cluster according to their sites of origin (42).

Oligonucleotide microarrays have shown considerable promise in determining genetic mechanisms, therapeutic targets, and prognostication. In this study, we use microarray data *clinically* to allow us to determine the tissue of origin of a tumor of unknown origin. We show that gene expression profiling can reliably inform us of a tumor's tissue of origin when compared with other like tissues of indistinguishable phenotypes. Furthermore, this profiling can be applied to metastases which seem to retain the expression profile of their primary tissue of origin, regardless of their location in the body. In this way, microarray technology may be used to improve clinical decision-making and patient care.

Acknowledgments

Received 6/14/2004; revised 1/25/2005; accepted 2/7/2005.

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.

We thank Nancy Bennett for her outstanding editorial assistance.

References

1. Jemal A, Tiwari RC, Murray T, et al. Cancer statistics. CA Cancer J Clin 2004;54:8–29.
2. Day TA, Davis BK, Gillespie MB, et al. Oral cancer treatment. Curr Treat Options Oncol 2003;4:27–41.
3. Miller DP, De Vivo I, Neuberg D, et al. Association between self-reported environmental tobacco smoke exposure and lung cancer: Modification by GSTP1 polymorphism. Int J Cancer 2003;104:758–63.
4. Braakhuis BJ, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications. Cancer Res 2003;63:1727–30.
5. Tabor MP, Brakenhoff RH, Van Houten VM, et al. Persistence of genetically altered fields in head and neck cancer patients: biological and clinical implications. Clin Cancer Res 2001;7:1523–32.
6. Levi F, Randimbison L, Te VC, La Vecchia C. Second primary cancers in patients with lung carcinoma. Cancer 1999;86:186–90.
7. Levi F, Randimbison L, Te VC, La Vecchia C. Second primary cancers in laryngeal cancer patients. Eur J Cancer 2003;39:265–7.

8. Webb T. Microarray studies challenge theories of metastasis. J Natl Cancer Inst 2003;95:350–1.

9. Caldas C, Aparicio SA. The molecular outlook. Nature 2002;415:484–5.

10. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001;98:15149–54.

11. Van de Vijver MJ, He YD, Van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.

12. Martini N, Bains MS, Burt ME, et al. Incidence of local recurrence and second primary tumors in resected stage I lung cancer. J Thorac Cardiovasc Surg 1995;109:120–9.

13. Battafarano RJ, Force SD, Meyers BF, et al. Benefits of resection for metachronous lung cancer. J Thorac Cardiovasc Surg 2004;127:836–42.

14. Liu D, Labow DM, Dang N, et al. Pulmonary metastasectomy for head and neck cancers. Ann Surg Oncol 1999;6:572–8.

15. Van de Wouw AJ, Jansen RL, Speel EJ, Hillen HF. The unknown biology of the unknown primary tumour: a literature review. Ann Oncol 2003;14:191–6.

16. Yamagata N, Shyr Y, Yanagisawa K, et al. A training-testing approach to the molecular classification of resected non-small cell lung cancer. Clin Cancer Res 2003;9:4695–704.

17. Dong G, Loukinova E, Chen Z, et al. Molecular profiling of transformed and metastatic murine squamous carcinoma cells by differential display and cDNA microarray reveals altered expression of multiple genes related to growth, apoptosis, angiogenesis, and the NF-κB signal pathway. Cancer Res 2001;61:4797–808.

18. Chen JJ, Peck K, Hong TM, et al. Global analysis of gene expression in invasion by a lung cancer model. Cancer Res 2001;61:5223–30.

19. Vigneswaran N, Wu J, Zacharias W. Upregulation of cystatin M during the progression of oropharyngeal squamous cell carcinoma from primary tumor to metastasis. Oral Oncol 2003;39:559–68.

20. Leethanakul C, Knezevic V, Patel V, et al. Gene discovery in oral squamous cell carcinoma through the head and neck cancer genome anatomy project: Confirmation by microarray analysis. Oral Oncol 2003;39:248–58.

21. Nakamura H, Saji H, Ogata A, et al. cDNA microarray analysis of gene expression in pathologic stage IA nonsmall cell lung carcinomas. Cancer 2003;97:2798–805.

22. Hippo Y, Taniguchi H, Tsutsumi S, et al. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res 2002;62:233–40.

23. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.

24. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.

25. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.

26. Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361:1590–6.

27. Van't Veer LJ, Dai H, Van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.

28. Talbot SG, O-charoenrat P, Sarkaria IS, et al. Squamous cell carcinoma related oncogene regulates angiogenesis through vascular endothelial growth factor-A. Ann Surg Oncol 2004;11:530–4.

29. O-charoenrat P, Rusch V, Talbot SG, et al. CSNK2A1 and C1-Inh are independent predictors of outcome in patients with squamous cell carcinoma of the lung. Clin Cancer Res 2004;10:5792–803.

30. Estilo CL, O-charoenrat P, Socci ND, et al. Identification of genetic prognosticators in oral tongue cancer using gene expression profiling and real-time PCR analysis [abstract]. Proc Annu Meet Am Assoc Cancer Res; 2003.

31. Li Y, Martin LD, Spizz G, Adler KB. MARCKS protein is a key molecule regulating mucin secretion by human airway epithelial cells *in vitro*. J Biol Chem 2001;276:40982–90.

32. Satijn DP, Olson DJ, van der Vlag J, et al. Interference with the expression of a novel human polycomb protein, hPc2, results in cellular transformation and apoptosis. Mol Cell Biol 1997;17:6076–86.

33. Blais JD, Filipenko V, Bi M, et al. Activating transcription factor 4 is translationally regulated by hypoxic stress. Mol Cell Biol 2004;24:7469–82.

34. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. In Brownstein MJ, Khodursky AB, editors. Functional genomics: methods and protocols. Totowa (NJ): Humana Press; 2002.

35. Ouyang X, Gulliford T, Doherty A, Huang GC, Epstein RJ. Detection of ErbB2 oversignalling in a majority of breast cancers with phosphorylation-state-specific antibodies. Lancet 1999;353:1591–2.

36. Kikuchi T, Daigo Y, Katagiri T, et al. Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. Oncogene 2003;22:2192–205.

37. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–44.

38. Shedden KA, Taylor JM, Giordano TJ, et al. Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. Am J Pathol 2003;163:1985–95.

39. Leong PP, Rezai B, Koch WM, et al. Distinguishing second primary tumors from lung metastases in patients with head and neck squamous cell carcinoma. J Natl Cancer Inst 1998;90:972–7.

40. Xu Y, Selaru FM, Yin J, et al. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res 2002;62:3493–7.

41. Ramaswamy S, Perou CM. DNA microarrays in breast cancer: the promise of personalised medicine. Lancet 2003;361:1576–7.

42. Dennis JL, Vass JK, Wit EC, Keith WN, Oien KA. Identification from public data of molecular markers of adenocarcinoma characteristic of the site of origin. Cancer Res 2002;62:5999–6005.

Gene Expression Profiling Allows Distinction between Primary and Metastatic Squamous Cell Carcinomas in the Lung

Simon G. Talbot, Cherry Estilo, Ellie Maghami, et al.

Cancer Res 2005;65:3063-3071.