**Carcinogenesis** vol.28 no.3 pp.760–766, 2007
doi:10.1093/carcin/bgl207
*Advance Access publication November 1, 2006*

# Expression profiling defines a recurrence signature in lung squamous cell carcinoma

**Jill Everland Larsen[1,3,*], Sandra Jane Pavey[3,4], Linda Hazel Passmore[1], Rayleen Bowman[3], Belinda Edith Clarke[2], Nicholas Kim Hayward[3,4] and Kwun Meng Fong[1,3]**

[1]Department of Thoracic Medicine, [2]Department of Pathology, The Prince Charles Hospital, Brisbane, 4032, Australia, [3]School of Medicine, University of Queensland, Herston, 4006, Australia and [4]Human Genetics Laboratory, Queensland Institute of Medical Research, Herston, 4006, Australia

*To whom correspondence should be addressed. Tel: +61 7 3139 4110; Email: Jill_E_Larsen@health.qld.gov.au

**Lung cancer remains the leading cause of cancer death worldwide. Overall 5-year survival is ~10–15% and despite curative intent surgery, treatment failure is primarily due to recurrent disease. Conventional prognostic markers are unable to determine which patients with completely resected disease within each stage group are likely to relapse. To identify a gene signature associated with recurrent squamous cell carcinoma (SCC) of lung, we analyzed primary tumor gene expression for a total of 51 SCCs (Stages I–III) on 22 323 element microarrays, comparing expression profiles for individuals who remained disease-free for a minimum of 36 months with those from individuals whose disease recurred within 18 months of complete resection. Cox proportional hazards modeling with leave-one-out cross-validation identified a 71-gene signature capable of predicting the likelihood of tumor recurrence and a 79-gene signature predictive for cancer-related death. These two signatures were pooled to generate a 111-gene signature which achieved an overall predictive accuracy for disease recurrence of 72% (77% sensitivity, 67% specificity) in an independent set of 58 (Stages I–III SCCs). This signature also predicted differences in survival [log-rank $P = 0.0008$; hazard ratio (HR), 3.8; 95% confidence interval (CI), 1.6–8.7], and was superior to conventional prognostic markers such as TNM stage or N stage in predicting patient outcome. Genome-wide profiling has revealed a distinct gene-expression profile for recurrent lung SCC which may be clinically useful as a prognostic tool.**

## Introduction

Lung cancer remains the leading cause of cancer death in Western countries with an overall 5-year survival of 10–15% (1). Complete surgical resection remains the most effective treatment, but despite clinical and technical advancements, the outcome of lung cancer has not improved significantly during the last 20 years. Treatment failure is frequently attributable to the presence of undetectable and unpredictable micrometastases (2). Generally, early-stage tumors have better clinical outcome and tumor staging aids treatment planning, yet there are instances where patients unexpectedly develop recurrent disease, illustrating the limitations of current clinical staging techniques in accurately predicting tumor recurrence.

Recurrent disease is disease that develops after initial treatment of the primary tumor with curative intent, and may occur in local (e.g. bronchial stump), regional (e.g. mediastinal lymph node) or distant (e.g. contralateral lung, brain, liver or bone) sites. Distant metastatic disease arises where malignant cells separate from the primary tumor and establish secondary tumor growth.

Non-small cell lung carcinomas (NSCLCs) constitute ~80% of all lung cancers, with small-cell carcinomas making up the remaining 20%. The NSCLC group can be further divided into histological subtypes with adenocarcinoma (AC), squamous cell carcinoma (SCC) and large cell carcinoma (LCC) being the most common, accounting for 40, 27 and 8% of all lung cancers, respectively (3). Gene-expression profiling has recently been used to characterize prognosis in lung cancer (4–14), mostly with a survival endpoint rather than tumor recurrence. Two previous studies, unstratified for histopathologic subtype, reported expression profiles associated with disease recurrence (9,13). As prognosis may differ between SCC and AC it is possible these signatures reflect, in part, the known gene-expression differences between histological types of lung cancer (4,5,15). The single study which has described a gene-expression signature of recurrence in SCC had limited sample size with 10 training samples and 5 test samples (11).

In this study, we sought to determine if a gene-expression profile of disease recurrence exists in SCC using a large well-defined cohort. Microarray analysis was used to profile gene expression in primary lung SCCs that were associated with clinically distant prognostic outcomes. We aimed to identify a clinically useful classification signature that could predict the likelihood of tumor recurrence and which may ultimately aid clinicians in making treatment decisions for this tumor type.

## Materials and methods

Additional details for the Materials and methods given here can be found in the Supplementary information.

### *Sample collection and selection*

Fresh-frozen primary lung tumor tissue specimens were obtained from The Prince Charles Hospital Tissue Bank. Tumor samples were collected from consecutive patients undergoing curative surgical resection, between 1990 and 2004, excluding patients who had undergone neo-adjuvant radiation or chemotherapy. The study protocol was approved by the Ethics Committee of The Prince Charles Hospital, and subjects gave informed written consent

**Abbreviations:** AC, adenocarcinoma; LCC, large cell carcinoma; NSCLC, non-small cell lung carcinoma; Rec(−), clinically disease-free for atleast 36 months following surgery; Rec(+), recurrent; SCC, squamous cell carcinoma.

for use of resected tissue. Study inclusion criteria were as follows: primary NSCLC of the squamous cell histological subtype (mixed histology excluded), tumor haematoxylin and eosin examination showed at least 50% tumor cells, surgical bronchial margins were free of disease and resection was considered complete, no adjuvant chemotherapy, and fitted to one of our two disease recurrence outcome criteria: non-recurrence [Rec(−)], clinically disease-free for at least 36 months following surgery; or recurrent disease [Rec(+)], unambiguous clinical, imaging, or histopathologic evidence of recurrence of the original primary lung cancer in a local or distant metastatic site occurring between 3 and 18 months post-resection in patients. The threshold of 36 months for Rec(−) cases was selected because the majority of patients develop disease recurrence within this period of time (16) and to allow for comparison with other similarly designed studies (14). A summary of individual sample characteristics of the training set is given in Supplementary Table I.

### Array hybridization

Microarray experiments conformed to MIAME guidelines (http://www.mged.org/Workgroups/MIAME/miame_checklist.html). Each tumor sample was compared with Universal Human Reference RNA (Stratagene, CA), which consists of RNA pooled from eleven cell lines of diverse tissue types. Twenty micrograms of total RNA from tumor samples (test) and universal reference RNA (reference) were subjected to first-strand cDNA synthesis with oligo-dT (0.1 μg) and random hexamer primers (0.1 μg) using a CyScribe Post-Labeling Kit (Amersham/GE Healthcare, PA) incorporating aminoallyl-modified dUTP. Modified nucleotides were then chemically coupled to Cy3 (reference) or Cy5 (test) fluors (Amersham/GE Healthcare, PA) and any uncoupled dyes removed using a CyScribe GFX™ Purification Kit (Amersham/GE Healthcare, PA). Each tumor sample (Cy5) was then pooled with a reference sample (Cy3) and hybridized to a commercially available 22 K Human Oligo Microarray (Operon Human Genome Oligo Set v2.1 (http://www.operon.com) containing 21 329 70mer probes representing ∼14 200 named transcripts) and printed by the British Columbia Gene Array Facility (http://prostatelab.org/arraycentre/index.html). Hybridization was carried out at 37°C for 16−18 h, and the slides were washed according to the manufacturer's protocol. Arrays were scanned by a GMS418 confocal scanner (Affymetrix, CA).

### Microarray analysis

Raw images were imported into Imagene V5.1 (BioDiscovery, CA) to extract pixel intensities and to flag spots with poor/absent signal. The raw background subtracted signal median for each probe was imported into GeneSpring GX V7.3 (Agilent Technologies, CA) for analysis. Data was normalized and probe signals filtered on pixel intensity and consistent spot morphology. For each probe, the logarithm to the base 2 of the ratio between the intensity in the tumor sample (red) channel and the reference (green) channel was used as the expression value for the probe. Of 21 329 probes present on the chip, 18 525 passed the filtering criteria. The data discussed in this publication have been made available in the NCBI Gene Expression Omnibus (GEO) public repository (http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE5123.

### Statistical analysis

To identify a signature that predicts the likelihood of recurrence, we use Cox proportional hazards modeling, leave-one-out cross-validation (LOOCV), and class prediction using 1-nearest neighbor (BRB ArrayTools Version 3.5; developed by Dr. Richard Simon and Amy Peng Lam, http://linus.nci.nih.gov/~brb/tool.htm). The major endpoint was time to recurrence, defined as the time from surgery to tumor recurrence (local, regional or distant). As recurrence usually leads to death, we also tested time to cancer-related death, defined as the time from surgery to death, where death was cancer-related. Cases that died of a non-cancer-related death were excluded from the latter analysis. To identify genes robustly correlated with time to recurrence and cancer-related death, we used Cox proportional hazards modeling with LOOCV. Briefly, 51 iterations of Cox modeling were performed so that each sample was left out once with the significance of each gene in relation to time to recurrence/cancer-related death calculated at each iteration. *P*-values for each gene were then averaged and ranked to identify genes that consistently, and robustly, correlated with outcome. We selected genes that met set criteria (average $P < 0.01$) to identify the genes that comprise our signature.

The signature was then tested for its predictive ability by 1-nearest neighbor class prediction in an independent set 130 primary lung SCCs (test set) (14) recently reported by Raponi *et al*. Hierarchical clustering was performed using Pearson correlation with bootstrapping of 1000 iterations. Principal component analysis was performed using Avadis V4.3 (Strand Genomics, India) scaled for equal variance. Kaplan–Meier survival plots and log-rank tests performed in SPSS Version 11.5 (SPSS, Chicago, Illinois, USA) were used to assess the differences in survival of the predicted good and poor outcome groups. Distributions of clinical and pathological parameters were analyzed using $\chi^2$, *t*-test, or log-rank tests where appropriate.

## Results

### Tumor samples

To identify a prognostic expression signature of tumor recurrence, a training set of 51 primary SCCs were studied, 29 with Stage I disease, 15 with Stage II and 7 with Stage III. The demographics of the patients and tumors in the training and test sets are outlined in Table I with detailed information on each patient in the training set given in Supplementary Table I. Univariate analysis showed that the maximum diameter of the primary tumor and pathological nodal stage (pN) were associated with tumor recurrence in these patients ($P = 0.006$, *t*-test and $P = 0.016$, $\chi^2$, respectively). Rec(−) and Rec(+) phenotypes were not associated with other clinical or pathological factors including age, gender, smoking history, TNM stage, differentiation, tumor invasion (lymphatic, vascular, pleural or perineural), and clinical investigation (Table I). Two patients had adjuvant radiation, R19: Rec(+) and N32: Rec(−). All patients had follow-up of a minimum 60 months or until death.

### Unsupervised analysis of SCCs

Unsupervised hierarchical clustering was initially performed on the 51 SCCs to identify any subgroups (clinical or otherwise) with distinct gene-expression profiles. A filtered gene list of 6748 probes was applied after excluding probes from the original filtered list of 18 525 probes with low log-ratio variation ($P > 0.01$). SCCs clustered into two distinct groups of 24 and 27 tumors (Figure 1), with a robustness index of 0.74 over 1000 permutations indicating high reproducibility. These two clusters differed in terms of time to cancer-related death with borderline statistical significance ($P = 0.05$, log-rank) (Figure 1). Additionally, Cluster 1 (with poorer survival), had a higher frequency of poorly differentiated tumors (67 versus 37%), although this was not statistically significant ($P = 0.07$, $\chi^2$). No significant association was identified between the clusters and recurrence phenotype, recurrence site (none, local, distant), TNM stage, tumor size (<3 cm/≥3 cm), N stage, invasion status (vascular, lymphatic, pleural or perineural), smoking status, gender, age or pack-years.

### Identification of a SCC signature to predict tumor recurrence

To identify a robust set of genes whose expression could identify primary tumors likely to develop recurrent disease we employed Cox proportional hazard modeling. For each probe in the filtered set of 6748 probes, we computed a statistical significance level for two endpoints—time to recurrence and the correlated, time to cancer-related death, based on univariate Cox proportional hazards models (17). *P*-values were used in a multivariate permutation test (10 000 permutations) in which the survival times and censoring indicators were randomly permutated among arrays. To test the robustness of the signature in the training set, probes selected by hazards modeling were subjected to LOOCV. This led to the identification of a recurrence signature of

**Table I.** Clinical, pathological and prognostic characteristics of cases in training and test sets

| Characteristics | Training set | | Test set All tumors |
|---|---|---|---|
| | All tumors | R(+)/R(−) | |
| Patient demographics | | | |
| Total number of samples (*N*) | 51 | 32/19 | 58 |
| Age, median (years) | 68 | 68/68 | 66 |
| Gender [*n* (%)] | | | |
| Male | 37 (73) | 23/14 | 35 (60) |
| Female | 14 (27) | 9/5 | 23 (40) |
| Smoking status [*n* (%)] | | | |
| Never | 2 (4) | 1/1 | NA |
| Former | 13 (25) | 8/5 | NA |
| Current | 36 (71) | 23/13 | NA |
| Pack-years, median | 50 | 50/55 | 60 |
| Primary tumor | | | |
| TNM Stage [*n* (%)] | | | |
| IA | 8 (16) | 6/2 | 8 (14) |
| IB | 21 (41) | 16/5 | 22 (38) |
| IIA | 1 (2) | 1/0 | 2 (3) |
| IIB | 14 (27) | 5/9 | 12 (21) |
| IIIA | 5 (10) | 2/3 | 10 (17) |
| IIIB | 2 (4) | 2/0 | 4 (7) |
| Max. tumor measurement (mm), median | 45 | 40/55[a] | NA |
| Differentiation [*n* (%)] | | | |
| P | 25 (49) | 14/11 | 19 (33) |
| M | 15 (29) | 10/5 | 33 (57) |
| W | 11 (22) | 8/3 | 6 (10) |
| Tumor invasion [*n* (%)] | | | |
| Lymphatic | 5 (10) | 3/2 | NA |
| Vascular | 15 (29) | 8/7 | NA |
| Perineural | 0 (0) | 0/0 | NA |
| Pleural | 13 (25) | 7/6 | NA |
| Clinical investigation [*n* (%)] | | | |
| Abdominal CT scan | 49 (96) | 30/19 | NA |
| Bone scan | 34 (67) | 19/15 | NA |
| Chest CT scan | 51 (100) | 32/19 | NA |
| Head CT scan | 21 (41) | 12/9 | NA |
| PET scan | 7 (14) | 4/3 | NA |
| Prognostic parameters | | | |
| Recurrence site [*n* (%)] | | | |
| NED[b] | 29 (57) | 29/0 | NA |
| Lung—primary site | 10 (20) | 2/8 | NA |
| Lymph node (loco-regional) | 2 (4) | 0/2 | NA |
| Lung—distant | 1 (2) | 0/1 | NA |
| Bone | 3 (6) | 0/3 | NA |
| Brain | 5 (10) | 1/4 | NA |
| Liver | 1 (2) | 0/1 | NA |
| Adrenal | 1 (2) | 0/1 | NA |
| Survival status [*n* (%)] | | | |
| Alive | 24 (47) | 24/0 | 24 (41) |
| Dead | 27 (53) | 8/19[c] | 34 (59) |

Univariate analyses were performed to identify significant associations between parameters and recurrence phenotype in training set ($\chi^2$, *t*-test, or log-rank). NA, not available.

[a] *P*-value = 0.006.

[b] NED, no evidence of recurrent disease.

[c] Death was cancer-related in 4 out of 8 [50%; Rec(−)] tumors and 18 out of 19 [94.7%; Rec(+)] tumors.

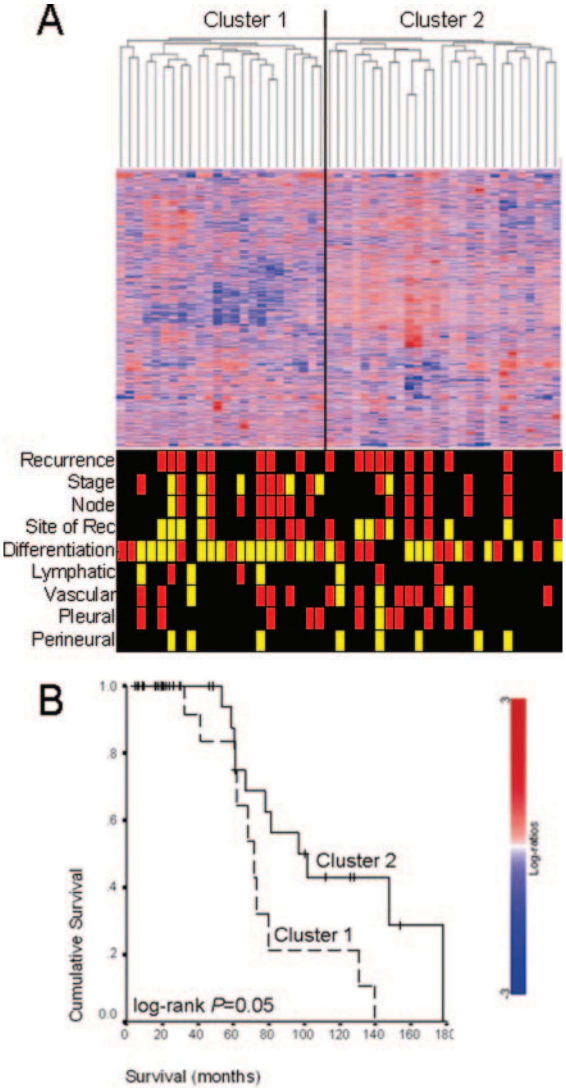

**Fig. 1.** Unsupervised analysis of the training set of 51 SCC samples identifies two clinically relevant subsets of SCC. (**A**) Unsupervised hierarchical clustering of the 51 SCCs (using Pearson correlation with 1000 bootstrap iterations using a filtered list of 6748 probes) identified two distinct clusters of 24 and 27 samples. Each column is a sample and each row a gene. Heatmap indicates level of gene-expression, red, high expression, blue, low expression. Prognostic parameters color-coded beneath heatmap: Recurrence (black, no recurrence; red, recurrence); TNM stage (black, Stage I; red, Stage II; yellow, Stage III); Nodal stage (black, N0; red, N1; yellow, N2); Site of recurrence (black, none; red, distant; yellow, local); Differentiation (black, well; red, moderate; yellow, poor); Tumor invasion (lymphatic, vascular, pleural or perineural) (black, no; red, yes; yellow, unknown). (**B**) Cluster 1 had poorer survival in Kaplan–Meier analysis (log-rank test) as well as a higher frequency of poorly differentiated tumors. Tick marks indicate patients whose data were censored at last follow-up.

71 probes [whose expression significantly correlated to time to tumor recurrence ($P < 0.01$)] and a survival signature of 79 probes [whose expression significantly correlated to time to cancer-related death ($P < 0.01$)]. Then, we combined the two outcome signatures to form one final overarching signature of 111 probes due to the close correlation between time to recurrence and time to cancer-specific death and as 39 probes were common to both signatures (Supplementary Table II). Principal component analysis (PCA) was used to assess the degree of separation in our set of 51 tumors according to disease recurrence status (Figure 2A). To ensure a signature will have predictive capability in independent cohorts, it is important not to over-fit the signature to the training set. PCA analysis illustrated that within the training set of 51 SCCs, our 111-gene signature could separate most, but not all, tumors by recurrence phenotype.

Hierarchical clustering was performed on the 111 genes and 51 samples to visualize the expression levels and identify genes with similar expression patterns (Figure 2B).

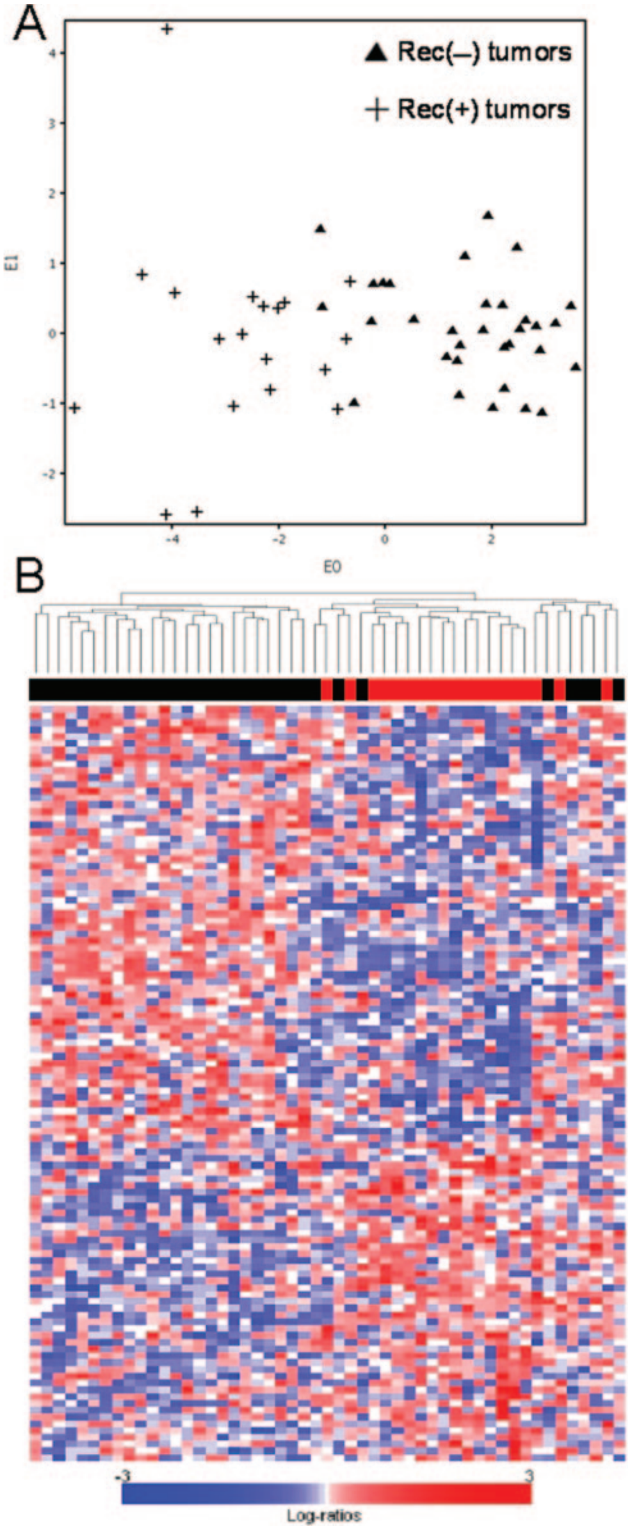

**Fig. 2.** Separation of the training set of 51 SCC samples by the 111-gene signature. (**A**) Principal component analysis (PCA) allowed evaluation of the level of separation between recurrence phenotypes [Rec(+) and Rec(−)]. (**B**) Hierarchical clustering using a Spearman's correlation was performed in both the sample and gene dimensions. Two distinct patterns of gene expression were observed in the 111-gene signature, genes whose expression correlates with better prognosis (upper half of heatmap) and genes whose expression correlates with poorer outcome (lower half of heatmap). Heatmap is colored by level of gene-expression, red, up-regulated genes, blue, down-regulated genes. Samples are color-coded in relation to recurrence phenotype, black, no recurrent disease at 36 months, red, recurrent disease within 18 months.

Two distinct groups of genes were identified, genes whose expression directly correlated with a better outcome, and genes whose expression inversely correlated with a better outcome in our training set.

### *Validation of the SCC recurrence signature in an independent test set*

To assess the predictive potential of a prognostic signature or classifier, validation is necessary in an independent cohort of samples (test set; Table I). We used a high quality publicly available dataset of 130 SCCs described previously (14) to test whether the 111-gene signature performed as well at predicting outcome in these tumors as in the training set of 51 SCCs. Signal intensity data scaled to an average intensity of 600 U was downloaded from http://www.ncbi.nlm.nih.gov/geo/ (GEO Series accession no. GSE4573) and imported into BRB ArrayTools. Analysis of variance (ANOVA) was performed to normalize experimental differences between our dataset and the test dataset. Since disease recurrence data was not available for this 130 sample test set, we used the published overall survival data as a surrogate marker, polarizing for recurrence. This was done by defining 'poor outcome' as patients who died within 1.5 years after surgery and 'good outcome' as patients having >5 years of clinical disease-free follow-up, leaving a test set of 58 SCCs (Stages I–III).

We used Chip Comparer (http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl) and independent mapping based on Unigene and Genbank accessions to determine that 78 (70%) of the 111 genes in our recurrence/survival signature were mapped to the Affymetrix U133A GeneChip microarray which was used for the test set samples. For probes that could not be mapped by these methods, a BLAST search was performed on the target sequence of the Operon HumanV2 probe. Supplementary Table II indicates which genes in the signature were present on the U133A array.

We applied the reduced recurrence/survival signature of 78 genes to predict the classification of the 58 SCCs as likely to have either good or poor outcome using a 1-nearest neighbor predictor (BRB ArrayTools). This determines which expression profile in the training set (51 SCC samples) is most similar to the expression profile of each sample in the test set. The expression profile is a vector of log-ratios or log-intensities for the signature. Euclidean distance is used as the distance metric for the Nearest Neighbor Predictor. The class of the nearest neighbor in the training set for each test sample is taken as the predicted class. In the test set of 58 SCC samples the signature had an accuracy of 72% in predicting good outcome (alive for a minimum of 5 years after surgery) or poor outcome (died within 1.5 years following surgery). The signature was slightly better at predicting tumors with poor outcome (77% sensitivity, 67% specificity) than tumors with good outcome (67% sensitivity, 77% specificity). To verify that this level of performance did not occur by chance, 1000 random permutations of class labels were performed. The proportion of permutations with accuracy similar to that attained by the true class labels was 0.001. Kaplan–Meier log-rank analysis confirmed that the predicted good and poor outcome groups had significantly different survival [log-rank $P = 0.0008$; hazard ratio (HR), 3.77; 95% confidence interval (95% CI), 1.63–8.70; Figure 3A].

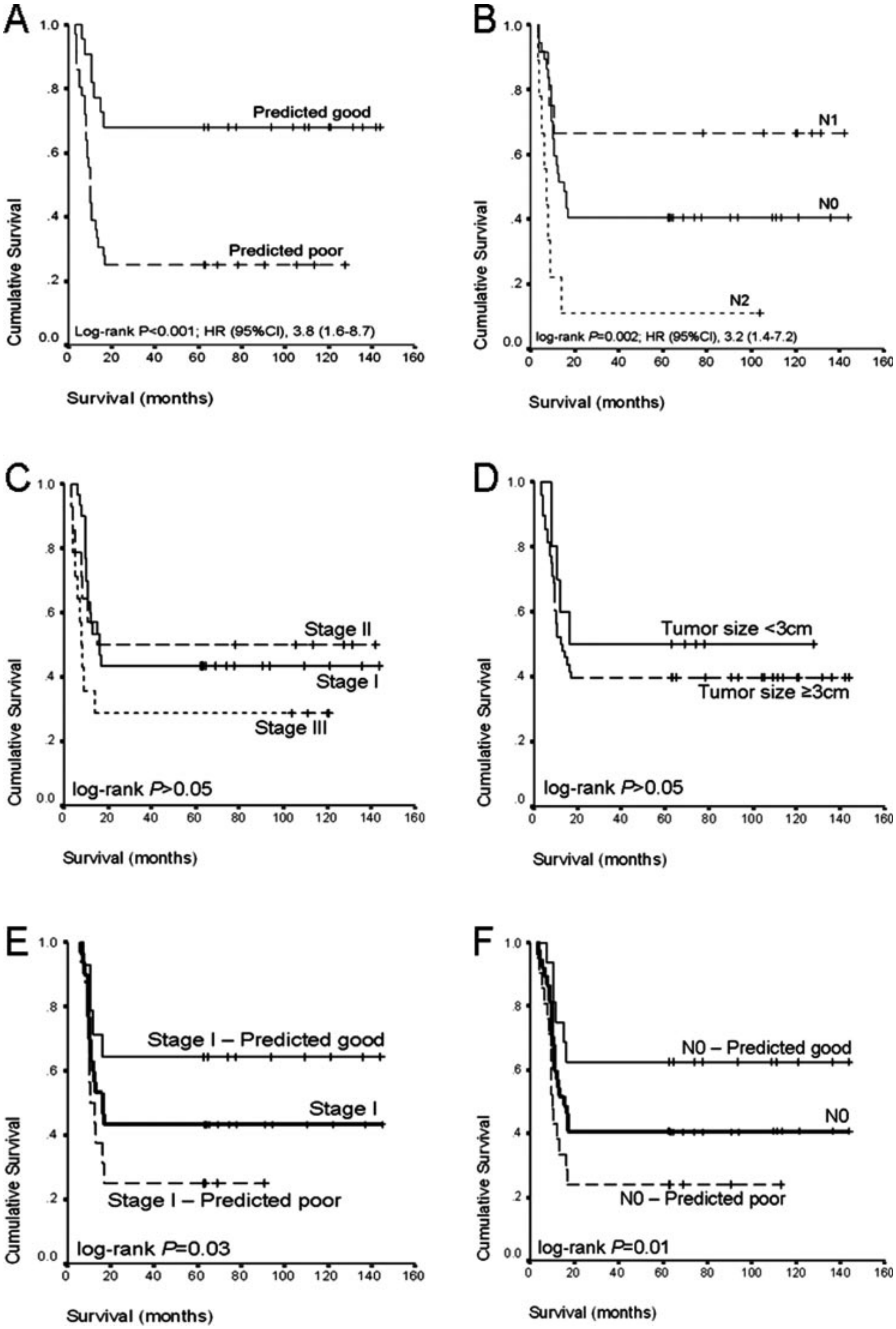

**Fig. 3.** Validation of the 111-gene signature in an independent test set of 58 SCCs. Comparison of Kaplan–Meier survival estimates (log-rank test) for predicted patient groups using the 111-gene signature with clinical patient groups. HRs and 95% CI obtained by Cox regression. Tick marks indicate patients whose data were censored at last follow-up. (**A**), Predicted good ($n = 22$; full line) and poor ($n = 36$; dashed line) outcome groups using 111-gene signature in 58 SCCs; (**B**) survival estimates in relation to nodal stage (N0: full line, N1: dashed line, N2: dotted line); (**C**) survival estimates in relation to TNM stage (Stage I: full line, Stage II: dashed line, Stage III: dotted line); (**D**), survival estimates in relation to tumor size (<3 cm: full line, ≥3 cm: dashed line); (**E**), in Stage I tumors ($n = 30$): overall survival (bold line), and signature predicted good ($n = 14$; full line) and poor ($n = 16$; dashed line) outcome groups; (**F**), in N0 tumors ($n = 37$): overall survival (bold line), and signature predicted good ($n = 16$; full line) and poor ($n = 21$; dashed line) outcome groups.

### *Comparison of recurrence/survival signature against conventional prognostic markers*

We used Kaplan–Meier log-rank tests to determine if the signature was a better predictor of outcome than conventional prognostic markers such as TNM stage, tumor size and N stage in the independent test set of 58 SCC samples. Univariately, N stage could stratify patients with respect to survival (log-rank $P = 0.002$; HR, 3.17; 95% CI, 1.39–7.22; Figure 3B) but inclusion of nodal stage as a confounder in the Cox regression model revealed that the signature could predict outcome independently of N stage (Wald $P = 0.002$; HR, 3.77; 95% CI, 1.63–8.70). TNM stage or tumor size

(<3 cm or $\geq$3 cm) were not significantly associated with survival ($P > 0.05$) (Figure 3C and D). Finally, we examined the prognostic discriminatory ability of the signature in very early stage SCCs which are generally expected to have a good outcome. The signature predicted differences in survival within both TNM Stage I samples (log-rank $P = 0.0310$; HR, 2.99; 95% CI, 1.05–8.52; $n = 30$) and N0 stage samples (log-rank $P = 0.0102$; HR, 3.22; 95% CI, 1.25–8.27; $n = 37$) (Figure 3E and F).

## Discussion

We investigated lung SCCs to determine if a gene-expression profile predictive of recurrence is present in the primary tumor. The samples used in this study were collected from patients treated by surgical resection with curative intent where a subset of patients (37%) developed tumor recurrence within 18 months of resection. Unsupervised hierarchical clustering on 51 SCCs identified clusters which differed in time to cancer-related death, but were not significantly different in other pathological or clinical characteristics suggesting the potential of primary tumor gene expression to be independently prognostic. Similar findings have also been reported for SCC (12,14).

To determine whether a signature was present within the primary tumor that could predict the likelihood of recurrence, we initially identified two groups of prognosis-related genes by using two defined clinical outcomes: one associated with time to disease recurrence and one associated with time to cancer-related death in a training set of 51 SCCs. These are often closely related outcomes; however, it is not uncommon for patients with low volume disease recurrence to have a relatively long survival. We considered it possible that prognostic signatures predictive of survival identified in previous studies (6,14) may not necessarily be able to predict recurrence. To develop our prognostic signature, we therefore included genes significantly correlated with either time to recurrence or time to cancer-related death. As expected, the gene ontology composition of the 111-gene signature included genes with biological relevance to disease recurrence, such as cell growth and movement, cell communication and cell signaling. A 111-gene signature was identified which was found to predict outcome in an independent test set of 58 SCCs with a 72% accuracy. It is likely that absence of 33 of the 111 genes from the array platform used for the test set will have limited the fidelity of the test set validation. Furthermore, the extrapolation from overall survival time (the surrogate outcome variable available for the test set) to dichotomize groups of cases likely to have developed early recurrence and those likely to have been recurrence-free required the setting of artificial censorship points. The accompanying introduced potential error could mean that the true predictive accuracy of the signature could be either higher or lower than the estimated 72%. Nevertheless, the ability of the signature to identify patients of significantly different overall survival in an independent set of SCC samples indicates its potential for application in the clinical setting. This is especially the case given that the predictive ability of the gene-expression signature was independent of and superior to conventional prognostic markers such as TNM stage, tumor size and N stage.

Several studies have used expression profiling to characterize prognosis in lung cancer (4–14), yet only one specifically explored an expression profile of disease recurrence in SCCs (11) using 10 samples. The authors described a 27-gene signature that could classify five unknown samples, however the training set comprised only 10 tumors—5 ‘aggressive’ (patients died of recurrence within 24 months) and 5 ‘less-aggressive’ (patients survived >54 months after surgery) (11). A recent, large study used a combined cohort of AC and SCC samples to define a lung metagene model capable of predicting the likelihood of recurrence in two independent test sets of 25 NSCLCs (11 AC and 14 SCC) and 84 ACs (13). However, as prognosis may differ between SCC and AC and there is substantial evidence for gene-expression differences between histological types of lung cancer (4,5,15), we used a homogeneous cohort of SCCs to determine expression differences due solely to tumor recurrence, without potentially confounding histological influences. Two studies have investigated survival in SCCs (11,12,14), one reporting distinct subgroups of SCCs with significant differences in the likelihood of survival to 6 years (12), the other identifying an optimized set of 50 prognostic genes in 129 lung SCCs with a predictive accuracy of 68% in classifying patients with good or poor overall survival in an independent set of 36 SCC samples (14).

A signature identified in gene-expression studies is not a set of genes but rather a function that can transform the expression levels for that set of genes to a risk score or predicted class (18). The lack of commonality of genes identified between the published prognostic signatures for lung cancer to date simply reflects the fact that numerous gene-expression signatures may be capable of predicting outcome in NSCLCs, a premise recently demonstrated in breast cancer (19–21). Cross-validation steps such as LOOCV can, if executed properly, give an unbiased estimate of how well the signature will perform in an independent set of samples but a major potential flaw in developing a predictive signature is over-fitting to the training dataset. An over-fitted signature will reflect characteristics of the individual samples represented in the training set and will not accurately predict outcome in independent test sets. Therefore the critical test of prognostic signatures is validation in independent datasets.

We envisaged that prediction of lung tumor recurrence will likely require a set of genes with moderate changes in expression as observed in breast cancer (19–21), rather than single genes. In this study, we have demonstrated the value of a genomic approach in identifying patients likely to develop tumor recurrence by validating our signature in an independent set of SCC samples. As a result, subjects at high risk of recurrent disease may require adjuvant treatment in addition to surgical resection. Genes and gene ontology pathways differentially expressed between recurrence phenotypes could also represent potential novel therapeutic targets.

## Supplementary data

Supplementary data are available at *Carcinogenesis* online.

## Acknowledgements

The authors would like the acknowledge Dr Edwina Duhig and Dr David Godbolt and staff in Anatomical Pathology at The Prince Charles Hospital for aiding in collection and pathological assessment of samples, staff in the Thoracic Research Laboratory at The Prince Charles Hospital for aid in specimen collection, Dr Maree L.Colosimo for clinical data, and Dr Ian A.Yang for general discussions on study design and analysis. J.E.L. is the recipient of an Edwin Tooth PhD Scholarship of the University of Queensland. K.M.F. is the recipient of a Queensland Smart State Fellowship. This work was supported in part by the National Health and Medical Research Council (338200) and The Prince Charles Hospital Research Foundation.

*Conflict of Interest Statement:* None declared.

## References

1. Mathers,C., Vos,T., Stevenson,C. and Australian Institute of Health and Welfare. (1999) *The burden of Disease and Injury in Australia: Summary Report*. Australian Institute of Health and Welfare, Canberra.
2. Fidler,I.J. (2003) The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer*, **3**, 453–458.
3. Feinstein,M.B. and Bach,P.B. (2000) Epidemiology of lung cancer. *Chest Surg. Clin. N. Am*, **10**, 653–661.
4. Bhattacharjee,A., Richards,W.G., Staunton,J. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
5. Garber,M.E., Troyanskaya,O.G., Schluens,K. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.
6. Beer,D.G., Kardia,S.L., Huang,C.C. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.
7. Miura,K., Bowman,E.D., Simon,R. *et al.* (2002) Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. *Cancer Res.*, **62**, 3244–3250.
8. Moran,C.J., Arenberg,D.A., Huang,C.C. *et al.* (2002) RANTES expression is a predictor of survival in stage I lung adenocarcinoma. *Clin. Cancer Res.*, **8**, 3803–3812.
9. Wigle,D.A., Jurisica,I., Radulovich,N. *et al.* (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.*, **62**, 3005–3008.
10. Gordon,G.J., Richards,W.G., Sugarbaker,D.J., Jaklitsch,M.T. and Bueno,R. (2003) A prognostic test for adenocarcinoma of the lung from gene expression profiling data. *Cancer Epidemiol. Biomarkers Prev.*, **12**, 905–910.
11. Sun,Z., Yang,P., Aubry,M.C., Kosari,F., Endo,C., Molina,J. and Vasmatzis,G. (2004) Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung? *Mol. Cancer*, **3**, 35.
12. Inamura,K., Fujiwara,T., Hoshida,Y. *et al.* (2005) Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*, **24**, 7105–7113.
13. Potti,A., Mukherjee,S., Petersen,R. *et al.* (2006) A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.*, **355**, 570–580.
14. Raponi,M., Zhang,Y., Yu,J. *et al.* (2006) Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.*, **66**, 7466–7472.
15. Ramaswamy,S., Tamayo,P., Rifkin,R. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.
16. Hawson,G., Zimmerman,P.V., Ford,C.A., Johnston,N.G. and Firouz-Abadi,A. (1990) Primary lung cancer: characterization and survival of 1024 patients treated in a single institution. *Med. J. Aust.*, **152**, 230–234.
17. Cox,D.R. (1972) Regression models and life-tables (with discussion). *J. Roy. Stat. Soc. B*, **34**, 187–220.
18. Simon,R. (2006) Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *J. Natl Cancer Inst.*, **98**, 1169–1171.
19. Glinsky,G.V., Higashiyama,T. and Glinskii,A.B. (2004) Classification of human breast cancer using gene expression profiling as a component of the survival predictor algorithm. *Clin. Cancer. Res.*, **10**, 2272–2283.
20. Paik,S., Shak,S., Tang,G. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.
21. Wang,Y., Klijn,J.G., Zhang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

*Received July 19, 2006; revised October 15, 2006;*
*accepted October 20, 2006*